# Implicative Analysis for Multivariate Binary Data using an Imprecise Dirichlet Model

**Jean-Marc Bernard**
Laboratoire Cognition et Activités Finalisées, Université Paris 8 & CNRS ESA 7021, France
berj@univ-paris8.fr

## Abstract

Bayesian implicative analysis was proposed for summarizing the association in a $2\times2$ contingency table in terms possibly asymmetrical such as, *e.g.*, "presence of feature $a$ implies, in general, presence of feature $b$" ("$a$ quasi-implies $b$" in short). Here, we consider the multivariate version of this problem: having $n$ units which are classified according to $q$ binary questions, we want to summarize the association between questions in terms of quasi-implications between features. We will first show how at a descriptive level the notion of implication can be weakened into that of quasi-implication. The inductive step assumes that the $n$ units are a sample from a $2^q$-multinomial population. Uncertainty about the patterns' true frequencies is expressed by an imprecise Dirichlet model which yields upper and lower posterior probabilities for any quasi-implicative statement. This model is shown to have several advantages over the Bayesian models based on a single Dirichlet prior, especially whenever $2^q$ is large and many patterns are thus unobserved by design.

**Keywords.** Quasi-implication, logical model, measure of association, multivariate implicative index, Boolean methods, Bayesian inference, upper and lower probabilities.

## 1 Introduction

Our purpose in this paper is to propose a new method for analyzing binary questionnaires, *i.e.* data in the form of $n$ units providing responses to $q$ binary questions. Most existing methods for analyzing a binary questionnaire take a symmetrical view of association between questions. Our purpose is, on the contrary, to possibly reach asymmetrical conclusions of the type "response $a$ to question $A$ implies, in general, response $b$ to question $B$" which we abbreviate to "$a$ quasi-implies $b$". The present work is thus directly related to "Scaling methods" ([13] and [18]) and, more generally, to methods that put forward Boolean/ordinal structures on the modalities, such as "Galois lattice theory" ([1] and [23]) and "Boolean analysis of questionnaires" ([7] and [10]).

Previous research in that field focussed on the qualitative analysis of exhaustive datasets merely based on the presence/absence of patterns. This approach leads to characterizing the structure of associations in the data by a set of implicative relations between modalities (see [8] and [12]). But, when $q$ is large, the implicative structure may be quite complex and calls for some kind of simplification or approximation. Introducing quantitative elements has been envisaged by filtering out the rarest patterns ([10] and [9]) but, as pointed out by [15] and [5], rarity of patterns on its own does not constitute a satisfying basis for defining quasi-implications.

Hildebrand *et al.* [15] investigated bivariate questionnaires for which they proposed the "*Del*" index as a measure of the agreement between the observed data and some given logical model. Inferences about this index were envisaged using frequentist methods, which presented serious difficulties for small samples and extreme data (see [15, ch. 6]). Bernard & Charron [5] proposed an alternative Bayesian approach free from these difficulties. Hildebrand *et al.*'s proposals for generalizing the *Del* index to multivariate binary data do not satisfy some minimal requirements of invariance by logical equivalence (see [15, ch. 7]) and the related inferential technique they propose suffers from the same limitations as in the bivariate case.

The aim of this paper is to propose an alternative approach for multivariate binary data. At the descriptive level, it provides a descriptive summary expressed in terms of quasi-implications whose definition is based on a new descriptive index $d$, called the *multivariate implicative index*. The quasi-logic that we define generalizes standard logic and is shown to satisfy several of its properties (*e.g.* transitivity).

For the inductive step, we assume that the data constitute a random sample from a $2^q$-multinomial population characterized by some unknown patterns' true frequencies (parameters). This step is envisaged within the Bayesian framework of inference: some prior state of uncertainty about parameters is updated to a posterior state of uncertainty by means of Bayes' theorem. One major advantage of the Bayesian approach is that it provides a straightforward means to make inferences about any derived parameter of interest and, more generally, about any property that parameters may satisfy. This point is particularly important for our present purpose — searching for an approximate implicative summary of the data —, since this requires making inferences about complex indices and properties.

More specifically, inference is envisaged from a "data analysis viewpoint" [4], *i.e.*, with the aim of bringing out what the data have to say about the parameters regardless of any prior information. In the Bayesian framework, this viewpoint amounts to choosing a prior state of prior ignorance about parameters. The usual Bayesian solutions to this problem, which we call "objective Bayesian models", all involve choosing a unique Dirichlet prior distribution for the parameters. We say "solutions" because several of them have actually been proposed as representing a prior state of ignorance. None of them simultaneously satisfies all desirable principles for formalizing prior ignorance, but, in problems involving a small number of categories, they generally lead to close numerical results. In our present context, however, the categories are the possible patterns and their cardinal, $2^q$, may be quite large.

Another solution is to resort to an imprecise probability model, that is, to a model where prior uncertainty about parameters is described by a set of prior distributions rather than by a single one. As a result any prior or posterior statement about parameters is associated with a probability interval instead of a single probability value. A thorough presentation of statistical models based on imprecise probabilities may be found in [20].

Walley [21] recently proposed the imprecise Dirichlet model (IDM) for analyzing multinomial data. This model satisfies several general principles of inference (*e.g.*, the likelihood principle, coherence). Two other important principles it satisfies are the "embedding principle" and the "representation invariance principle"; these last two principles ensure that two identical parameters (possibly defined for different numbers of categories) will receive the same inferential treatment (either prior or posterior); most objective Bayesian models do not satisfy them. For these reasons, the IDM appears to be a better model for formalizing prior ignorance than any objective Bayesian model (for more details, see [21] and [22]). Another important feature is that, through the degree of imprecision of probabilities, the IDM distinguishes between a relative lack of information (high imprecision) and a more substantial state of knowledge (low imprecision). This is particularly striking in our present context when the number of questions $q$ is large and, thus, many of the $2^q$ patterns are unobserved by design. In such a case, some objective Bayesian models lead to the surprising conclusion that any unobserved pattern is absent from the whole population. In contrast, the IDM distinguishes between those which truly point to an incompatibility of modalities and those which might just result from a small sample size and the independent conjunction of rare modalities.

The definition of quasi-implications will first be adressed in Sections 2 and 3 at a descriptive level, *i.e.* considering that the data constitute the entire population. We will turn to the inferential aspect of our method in Sections 4 (objective Bayesian inference) and 5 (IDM). An application to the analysis of a sociological survey is then described in Section 6. Finally, Section 7 provides some directions for future research.

## 2 Binary questionnaire and associated implications

### 2.1 Binary questionnaire

Let $U = \{u_1, \cdots, u_n\}$ be a set of $n$ *units* which may possess or not some of the $q$ *features* amongst the set of features $\{a, b, c, \ldots\}$. Each feature $a$ is associated with a *binary question* $A = \{a, a'\}$ whose *modalities* denote the presence ($a$) or the absence ($a'$) of the feature. The data can thus be described as a questionnaire about $n$ units, each providing responses to $q$ binary questions, $A = \{a, a'\}$, $B = \{b, b'\}$, $C = \{c, c'\}$, *etc.*. Any unit is described by some *pattern* $p \in P$, where $P = A \times B \times C \cdots$; for example, a unit which possesses features $b$ and $c$, but not feature $a$, will be described by a pattern of the type $p = a'bc\ldots$.

In a statistical context, the set $U$ of units typically constitutes a sample from a larger population, the elements of which need not be distinguished. At the same time, the interest will typically lie in studying the association between features, so that features or questions always need to be distinguished. With this asymmetrical view, the data can be described as a *weighted protocol* composed of weighted patterns $p \in P$, with associated weight $n_p$, where $n_p$ represents the number of units having pattern $p$.

As a simple example, we consider a fictitious questionnaire with $q = 3$ questions, $A$, $B$ and $C$ and $n = 100$ units given in the form of a weighted protocol, $(p, n_p)_{p \in P}$ (see Table 1).

| $n_p$ | $b$ | | $b'$ | |
|---|---|---|---|---|
| | $c$ | $c'$ | $c$ | $c'$ |
| $a$ | 13 | 0 | 0 | 7 |
| $a'$ | 1 | 1 | 61 | 17 |

Table 1: Questionnaire with $q = 3$ binary questions, $A$, $B$ and $C$, and $n = 100$ units (fictitious data).

Basic patterns are defined with reference to some basic set of $q$ questions $Q = \{A, B, C, \ldots\}$. A weighted protocol on $Q$ can be *projected* on any non-trivial subset $Q' \subset Q$. Basic patterns are then projected into *partial patterns* whose weights are obtained by adding up those of the constituent basic patterns. For instance, in the example of Table 1, projection onto $Q' = \{B, C\}$ leads to four partial patterns $bc$, $bc'$, $b'c$ and $b'c'$ with respective weights 14, 1, 61 and 24.

### 2.2 Logical expressions

In what follows, features $a, b, c, \ldots$ will be assimilated to elementary propositions which may be true or false. We use concatenation (*e.g.* in $ab$) for logical "and" between modalities, the prime symbol (*e.g.* in $a'$) for negation, and the symbol "$\Longrightarrow$" for logical implication. We shall also use symbols "$\wedge$" and "$\vee$" for logical "and" and "or" respectively, and the symbol "$\emptyset$" for designating the false proposition.

For any three propositions $r$, $s$ and $t$, it will easily be checked that,

$$r \Longrightarrow \emptyset \quad \equiv \quad r', \tag{1}$$
$$(r \wedge s) \Longrightarrow t \quad \equiv \quad r \Longrightarrow (t \vee s'), \tag{2}$$
$$(rs \Longrightarrow \emptyset) \wedge (rs' \Longrightarrow \emptyset) \quad \equiv \quad r \Longrightarrow \emptyset, \tag{3}$$

where symbol "$\equiv$" indicates logical equivalence between two propositions.

### 2.3 Implications between modalities

If interest is focussed on *strict* dependencies between modalities of the questions, the only aspect in the data that matters is whether each pattern $p$ is present ($n_p \geq 1$) or absent ($n_p = 0$). At this qualitative level, the observed protocol can be characterized by a list of implications between modalities (see [10] and [12]).

**Elementary implications.** The absence of one basic pattern $p \in P$ in the protocol is simply expressed by the logical statement $p'$ ($p$ is false), that is, using identity (1), by the *elementary implication*

$$p \Longrightarrow \emptyset,$$

where the adjective "elementary" indicates that the implication bears on a *single basic pattern*.

In the example of Table 1, pattern $ab'c$ is absent, *i.e.* $ab'c \Longrightarrow \emptyset$. By identity (2), this can also be expressed by either $ab' \Longrightarrow c'$, $b'c \Longrightarrow a'$ or $ac \Longrightarrow b$. The last expression, $ac \Longrightarrow b$ ("the conjunction of $a$ and $c$ implies $b$"), is the most simple for interpretative purposes. However, we shall prefer to use the form $r \Longrightarrow \emptyset$ which identifies absent patterns (basic or partial) since those play a central role in our method.

**Implicative structure of the protocol.** The simultaneous absence of several basic patterns, $p_1$, $p_2$, $p_3$, …, is expressed by the conjunction

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

or equivalently, using (1), by a conjunction of elementary implications

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

In the example of Table 1, both patterns $abc'$ et $ab'c$ are absent, that is $abc' \Longrightarrow \emptyset$ and $ab'c \Longrightarrow \emptyset$. Using (2), the implicative structure of the protocol can thus be expressed by

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

**Binary implications.** The recursive application of identity (3), when it is applicable (which is not the case in our example), enables one to condense the list of elementary implications into a more synthetic list of implications, each of which expresses the simultaneous absence of several basic patterns. Eventually, the most compact form will contain implications in which only *two* questions appear, such as $ab' \Longrightarrow \emptyset$ or $a \Longrightarrow b$, and which are of particular interest for the interpretation of the data. We call them *binary implications*.

## 3 Descriptive analysis: quasi-implications

When some patterns, though present, are rarely observed, one would like to weaken the qualitative distinction "present/absent" into the quantitative one "well represented/quasi-absent", and thus to weaken the notion of implication into that of quasi-implication. This will lead us to summarize the data by an approximate implicative structure.

We shall first define the notion of a quasi-implication (*q-implication* in short) on a descriptive basis, that is by only considering the relative frequencies of patterns, $\boldsymbol{f} = (f_p)_{p \in P}$, with $f_p = \frac{n_p}{n}$.

### 3.1 Two binary questions ($q = 2$)

Bernard & Charron [5] based the implicative analysis of a $2 \times 2$ contingency table ($q = 2$) on the "*Del*" index, proposed in this context by [15] and defined, for any $i \in \{a, a'\}$ and $j \in \{b, b'\}$, by

$$d_{ij \Longrightarrow \emptyset} \quad = \quad 1 - \frac{f_{ij}}{f_i f_j}, \tag{4}$$

where $f_i$ and $f_j$ are the marginal frequencies of $i$ and $j$ respectively.[1] This index measures the degree of agreement of the data with the logical model $ij \Longrightarrow \emptyset$: It takes the value 1 whenever $ij \Longrightarrow \emptyset$ is verified (*i.e.* cell $ij$ is empty) and the value 0 when the two questions are independent (in which case, $f_{ij} = f_i f_j$ for any $i, j$). Intermediate values can thus be considered as representing various degrees of $q$-implication from $i$ to $j$. For some fixed non-negative reference value $d_{quasi} \leq 1$, a *$q$-implication at degree $d_{quasi}$*, denoted by "$ij \longrightarrow \emptyset$", was defined in [5] by

$$ij \longrightarrow \emptyset \quad \textit{iff} \quad d_{ij \Longrightarrow \emptyset} \geq d_{quasi}. \tag{5}$$

As an example, consider the projection of the protocol in Table 1 onto $Q' = \{B, C\}$. For each partial pattern $p$ in the set $\{bc, bc', b'c, b'c'\}$, the index $d_{p \Longrightarrow \emptyset}$ takes the values $-0.24$, $0.73$, $0.04$ and $-0.13$ respectively. If we take $d_{quasi} = 0.50$, we thus find a single $q$-implication, $bc' \longrightarrow \emptyset$ *i.e.* $b \longrightarrow c$.

This example illustrates that this type of analysis provides the means to specify the precise direction of the association between variables: The present analysis, not only says that there is a positive association between $B$ and $C$, but also states that "$b$ $q$-implies $c$" (and not the reverse).

### 3.2 Generalization to several binary questions ($q > 2$)

A first direction for generalizing $q$-implications to more than two questions is to focus on binary implications, *i.e.* to only consider pairs of questions from the overall questionnaire. Unfortunately, this first approach may lead to a non-transitive summary of the protocol. Other attempts have been made by trying to express elementary implications in such a way as to bring them back to the case of two questions, using either some *conditioning* or some *asymmetrical compounding* (see [15, ch. 7]). However, as these last authors acknowledge, these attempts lead to inconsistent results since they produce statements that are not invariant by logical equivalence (see also [6]).

[1] This descriptive index is also known in the literature as "Loevinger's homogeneity index" [18].

#### 3.2.1 Multivariate implicative index

Instead, we propose a new index which generalizes (4), called the *multivariate implicative index*. For any pattern $p = ijk\ldots$, with $i \in \{a, a'\}$, $j \in \{b, b'\}$, $k \in \{c, c'\}$, *etc.*, this index is defined as

$$d_{p \Longrightarrow \emptyset} \quad = \quad 1 - \frac{f_p}{f_i f_j f_k \cdots}, \tag{6}$$

where $f_i$, $f_j$, $f_k$, *etc.* denote the marginal relative frequencies of $i$, $j$, $k$, *etc.* respectively.

The multivariate implicative index $d_{p \Longrightarrow \emptyset}$ constitutes a local measure of the departure of the protocol from *complete independence*, which is defined as: $f_p = f_i f_j f_k \ldots$, for all $p = ijk\ldots$ (see [17, p. 600]). This measure is local because it measures a departure in some *specific direction*, the one of the logical model $p \Longrightarrow \emptyset$. The index $d_{p \Longrightarrow \emptyset}$ varies within $]-\infty, 1]$; it takes the value 0 in case of local independence ($f_p = f_i f_j f_k \ldots$), positive values whenever pattern $p$ is under-represented ($f_p < f_i f_j f_k \ldots$), and equals 1 whenever pattern $p$ is absent ($f_p = 0$).

#### 3.2.2 Quasi-implications

Given some non-negative reference value $d_{quasi}$, we generalize the notion of an elementary implication "$p \Longrightarrow \emptyset$" into that of an *elementary $q$-implication*, denoted by "$p \longrightarrow \emptyset$", by defining

$$p \longrightarrow \emptyset \quad \textit{iff} \quad d_{p \Longrightarrow \emptyset} \geq d_{quasi}, \tag{7}$$

which reads "pattern $p$ quasi-implies $\emptyset$ (at degree $d_{quasi}$)" and which we shall also express as "pattern $p$ is *quasi-absent*" (*$q$-absent* in short). When needed, we shall also write "$p \overset{ABC\ldots}{\longrightarrow} \emptyset$" with an explicit mention of the level of projection at which elementary $q$-implications are defined.

The lower the reference value $d_{quasi}$, the higher is the number of observed patterns treated as $q$-absent. In the sequel of this paper, we shall often use the value $d_{quasi} = 0.50$ which corresponds to patterns that are under-represented by 50% relative to the case of complete independence.

Non-elementary $q$-implications are defined by the recursive use of the same combination rule (identity (3)) as for strict logic. Formally, we thus define a *$q$-implication* $r \longrightarrow s$ for any propositions $r$ and $s$ by

$$\begin{aligned} & r \overset{Q}{\longrightarrow} s \ \textit{iff} \\ & \quad \forall\, p \in P, \quad \textit{if}\ (r \Longrightarrow s) \Longrightarrow (p \Longrightarrow \emptyset), \\ & \qquad \textit{then}\ p \overset{Q}{\longrightarrow} \emptyset, \end{aligned} \tag{8}$$

where all elementary $q$-implications are assessed with the same reference value $d_{quasi}$.

### 3.2.3 Properties of this quasi-logic

**Generalization of standard logic.** Standard logic is obtained for $d_{quasi} = 1$: $d_{p \Longrightarrow \emptyset} \geq 1$ is equivalent to $d_{p \Longrightarrow \emptyset} = 1$, and hence to $p \Longrightarrow \emptyset$.

**Invariance.** The definition (8) of $q$-implications in terms of elementary $q$-implications, at a *unique level of projection*, guarantees invariance of $q$-implications by logical equivalence:

$$\textit{If } (r \Longrightarrow s) \Longleftrightarrow (u \Longrightarrow v), \quad \textit{then} \quad \forall\, d_{quasi} \geq 0,\ (r \xrightarrow{Q} s) \Longleftrightarrow (u \xrightarrow{Q} v). \quad (9)$$

**Transitivity.** Invariance by logical equivalence also ensures that $q$-implications satisfy transitivity. From definition (8), any valid $q$-implicative statement is equivalent to a conjunction of elementary $q$-implications and several such conjunctions cannot generate any contradiction.

**Coherence by projection.** Another important property is the *coherence by projection* of $q$-implications. Assume that "$abc\ldots \longrightarrow \emptyset$" and "$a'bc\ldots \longrightarrow \emptyset$" both hold; this can be summarized by "$bc\ldots \overset{ABC\ldots}{\longrightarrow} \emptyset$", where "$ABC\ldots$" indicates the level of projection. It can be shown that "$d_{bc\ldots \Longrightarrow \emptyset}$" is a weighted average of "$d_{abc\ldots \Longrightarrow \emptyset}$" and "$d_{a'bc\ldots \Longrightarrow \emptyset}$", so that having "$bc\ldots \overset{ABC\ldots}{\longrightarrow} \emptyset$" implies "$bc \overset{BC\ldots}{\longrightarrow} \emptyset$" (see [6]).

The recursive application of this result entails that any $q$-implication satisfied at some level $Q$ of projection is necessarily satisfied at any less refined level, $Q' \subset Q$. Contrarily to what happens in standard logic, the reverse is not true. This is the major difference between the quasi-logic presented here and standard logic: in standard logic, the level of projection is not relevant, whereas in our quasi-logic, $q$-implications are preserved by projection but not necessarily by refinement of the questionnaire.

### 3.3 Descriptive implicative summary of the protocol

The implicative summary of the protocol at the descriptive level consists in the set of patterns $p \in P$ such that $p \longrightarrow \emptyset$ holds. When data allow it, elementary $q$-implications may be combined into more synthetic $q$-implications and, eventually, into binary ones.

When $d_{quasi}$ is varied, several nested implicative summaries are obtained, from the qualitative vision of the protocol where only absent patterns are identified ($d_{quasi} = 1$) to the most drastic summary ($d_{quasi} = 0$) which considers that any under-represented pattern ($f_p < f_i f_j f_k \cdots$) is $q$-absent. For the example in Table 1, for any $d_{quasi} \in [0.90, 1.00]$, only the two patterns $abc'$ and $ab'c$ give rise to a $q$-implication; for $d_{quasi} \in [0.68, 0.89]$, an additional $q$-absent pattern, $a'bc$, is found; for $d_{quasi} \in [0.00, 0.66]$, there is another extra $q$-absent pattern, $a'b'c'$. For this last choice of $d_{quasi}$, the implicative summary is thus

$$(b \longleftrightarrow ac) \quad \wedge \quad (a' \longrightarrow c).$$

## 4 Objective Bayesian inference

### 4.1 From description to Bayesian inference

For the inductive step, we assume that the weighted protocol $\boldsymbol{n} = (n_p)_{p \in P}$ is a multinomial sample (with $K = |P| = 2^q$ categories) of size $n$ from an infinite population characterized by the parameters, or true relative frequencies, $\boldsymbol{\theta} = (\theta_p)_{p \in P}$: $\boldsymbol{n} \sim Mn(n, \boldsymbol{\theta})$.

In the usual version of the Bayesian framework, the state of uncertainty about parameters $\boldsymbol{\theta}$ is, at any moment (prior or posterior to the data), described by a *unique* probability distribution. For categorical data, prior uncertainty is usually expressed by a distribution from the conjugate Dirichlet family. From a Dirichlet prior on $\boldsymbol{\theta}$, $\boldsymbol{\theta} \sim Di(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_p)_{p \in P}$ a vector of non-negative reals, Bayes' theorem leads to an updated Dirichlet posterior on $\boldsymbol{\theta}$ conditionally on the observed data: $\boldsymbol{\theta} | \boldsymbol{n} \sim Di(\boldsymbol{n} + \boldsymbol{\alpha})$. The $K$ hyper-parameters composing vector $\boldsymbol{\alpha}$ can be thought of as *prior strengths* put on the various patterns; each prior strength, $\alpha_p$, is incremented by the observed frequency of the pattern, $n_p$, and thus updated into the posterior strength $n_p + \alpha_p$. We denote by $\nu$ the total prior strength: $\nu = \sum \alpha_p$.[2]

The posterior expectations of $\boldsymbol{\theta}$ are given by the relative posterior strengths, *i.e.*

$$E(\theta_p) \quad = \quad \frac{n_p + \alpha_p}{n + \nu}. \qquad (10)$$

Let $P(.)$ be some *property of interest* that a K-dimensional vector of relative frequencies might satisfy. Within the Bayesian approach, inference consists in deriving the posterior probability $Prob(P(\boldsymbol{\theta}))$ from the overall posterior on $\boldsymbol{\theta}$. If this probability is greater than some given guarantee $\gamma$, then the property can be assessed for the population with guarantee $\gamma$.

A typical property of interest here is that a single elementary $q$-implication $p \longrightarrow \emptyset$ (or that some more

[2] The standard definition of the Dirichlet distribution involves strictly positive $\alpha_p$'s. We extend this definition by allowing the possibility of null values for some (but not all) $\alpha_p$'s; a posterior strength $n_p + \alpha_p$ should be understood as meaning that $\theta_p = 0$.

complex model composed of several elementary $q$-implications) is satisfied. More precisely, for some pattern $p = ijk\ldots$, consider the derived parameter $\delta_{p\Longrightarrow\emptyset} = g(\boldsymbol{\theta})$ which is the population counterpart of the descriptive index $d_{p\Longrightarrow\emptyset} = g(\boldsymbol{f})$:

$$\delta_{p\Longrightarrow\emptyset} = 1 - \frac{\theta_p}{\theta_i\theta_j\theta_k\cdots}, \qquad (11)$$

where $\theta_i$, $\theta_j$, $\theta_k$, *etc.* denote the population marginal relative frequencies of $i$, $j$, $k$, *etc.* respectively. Here, the goal of inference will be to provide probabilistic statements about properties of the type $P(\boldsymbol{\theta})$ = "$\delta_{p\Longrightarrow\emptyset} \geq d_{quasi}$".

### 4.2 Usual reference priors and their difficulties

In the "data analysis approach to inference" [4], the $K$ prior strengths $\boldsymbol{\alpha}$ must be chosen so as to express a "prior state of ignorance", so that the posterior distribution essentially expresses the information on parameters brought by the data. Several reference priors have been proposed to achieve such a goal including Bayes-Laplace's ($\forall p$, $\alpha_p = 1$), Haldane's [14] ($\forall p$, $\alpha_p = 0$), Jeffreys' [16] ($\forall p$, $\alpha_p = 1/2$) and Perks' [19] ($\forall p$, $\alpha_p = 1/K$).

Bernard & Charron [5] used Perks' prior as a reference prior for the case $q = 2$. However, for the multivariate case, especially when $q$ is large, all of the four above priors appear quite unsatisfactory. The number of possible patterns $K = 2^q$ increases exponentially with $q$, so that when $q$ is large enough, $K$ will typically be much larger than $n$, $K >> n$. In such a case a large number of the possible patterns are unobserved *by construction*. Haldane's prior leads to the conclusion that these unobserved patterns are absent from the population. Perks' prior, in which the total prior strength $\nu$ also does not depend on $K$ ($\nu = 1$), allocates a very small prior strength to any pattern and thus tends to lead to the same undesirable conclusion. On the other hand, for the other two proposed priors, $\nu$ depends on $K$ ($\nu = K$ in Bayes-Laplace' prior, $\nu = K/2$ in Jeffreys'), so that the total prior strength will be much larger than $n$, the total strength provided by the data: the weight of evidence is overwhelmed by the prior. When $K << n$, the four Bayesian priors actually lead to very similar inferences, but when $K >> n$, large discrepancies appear between them, especially for rare patterns.

## 5 Imprecise Dirichlet model (IDM)

### 5.1 Presentation of the model

Instead of using a single prior, an alternative idea is that of using several priors within a restricted *ignorance zone* ([2] and [4]). This suggestion is also made by [21] under the name of an *imprecise Dirichlet model (IDM)*. The IDM consists in fixing the total prior strength $\nu$ and considering all possible Dirichlet priors satisfying the constraint:

$$0 \leq \alpha_p \quad \text{and} \quad \sum \alpha_p = \nu. \qquad (12)$$

Each Dirichlet prior in this set is then updated into a Dirichlet posterior using Bayes' theorem. Posterior uncertainty about $\boldsymbol{\theta}$ is thus described by the resulting set of Dirichlet posteriors. For any property of interest about $\boldsymbol{\theta}$, $P(\boldsymbol{\theta})$, the IDM yields a *lower* and an *upper probability* for statement $P(\boldsymbol{\theta})$, respectively denoted by $\underline{Prob}(P(\boldsymbol{\theta}))$ and $\overline{Prob}(P(\boldsymbol{\theta}))$.

The IDM as defined in (12) depends on the choice of $\nu$. The constant $\nu$ determines how fast the lower and upper probabilities converge one towards the other when $n$ increases. Walley [21] gives several arguments for chosing $\nu$ between 1 and 2 and notes that $\nu = 2$ might be overly cautious. For a one-sided test about a proportion in the case $K = 2$, the value $\nu = 1$ encompasses all above Bayesian solutions and several of their frequentist alternatives (see [2]). In the following, we shall use $\nu = 1$ for which the range of the resulting posterior imprecise probabilities always covers those obtained from both Haldane's and Perks' solutions.

### 5.2 Inductive implicative summary of the protocol

Under the IDM, the $q$-implication $p \longrightarrow \emptyset$ is said to be *inductively satisfied* (at degree $d_{quasi}$, with guarantee $\gamma$), if and only if

$$\underline{Prob}(\delta_{p\Longrightarrow\emptyset} \geq d_{quasi}) \quad \geq \quad \gamma. \qquad (13)$$

As we shall see, the degree of imprecision in the probabilities reflects prior ignorance. Basing the above inductive statement upon a lower probability $\underline{Prob}(.)$ amounts to producing a cautious statement: we may ensure that the property of interest, $\delta_{p\Longrightarrow\emptyset} \geq d_{quasi}$, has *at least* probability $\gamma$.

The conjunction of the elementary $q$-implications that are inductively satisfied constitutes a logical model which is an inductive implicative summary of the protocol (relative to $d_{quasi}$ and $\gamma$). (In a further step, one could also compute the lower probability of this logical model considered as a whole, which we shall not do here.)

### 5.3 Lower expectation for $\delta_{p\Longrightarrow\emptyset}$

Finding the lower probability required by (13) appears to be a hard task. Instead, we propose an indirect approximate procedure which, as intuition suggests, might actually provide the exact answer. The first level of approximation consists in trying to minimize the posterior expectation $E(\delta_{p \Longrightarrow \emptyset})$. In addition, since finding a closed form for this posterior expectation does not appear very easy either, we resort to a second level of approximation using a simple approximation for $E(\delta_{p \Longrightarrow \emptyset})$.

Under a single Dirichlet prior, $Di(\boldsymbol{\alpha})$, a simple approximate value for the posterior expectation $E(\delta_{p \Longrightarrow \emptyset})$ is given by replacing each $\theta_p$ in (11) by $E(\theta_p)$ given in (10), that is,

$$E^\star(\delta_{p \Longrightarrow \emptyset}) = 1 - \frac{\frac{n_p+\alpha_p}{n+\nu}}{\left(\frac{n_i+\alpha_i}{n+\nu}\right)\left(\frac{n_j+\alpha_j}{n+\nu}\right)\left(\frac{n_k+\alpha_k}{n+\nu}\right)\cdots}, \quad (14)$$

where $\alpha_i$, $\alpha_j$, $\alpha_k$, ... are the marginal sums of the $\alpha_p$'s.

It can be shown that the minimum value of $E^\star(\delta_{p \Longrightarrow \emptyset})$ with respect to $\boldsymbol{\alpha}$ constrained by (12) and $\nu = 1$, is attained for

$$\begin{cases} \alpha_p = 1, & \alpha_{p''} = 0, \ p'' \neq p \quad \text{if } f_p < f_{p'} \\ \alpha_{p'} = 1, & \alpha_{p''} = 0, \ p'' \neq p' \quad \text{if } f_{p'} \leq f_p \end{cases}, \quad (15)$$

where $p' = i'j'k' \ldots$ is opposite to pattern $p = ijk \ldots$.

For each pattern $p$, we shall take as an approximate value for $\underline{Prob}(\delta_{p \Longrightarrow \emptyset} \geq d_{quasi})$ the probability $Prob(\delta_{p \Longrightarrow \emptyset} \geq d_{quasi})$ obtained from a single Dirichlet distribution with prior strengths given in (15).[3]

### 5.4 Computational issues

The proposed method involves heavy computations because each Dirichlet posterior distribution bears on $(K-1 = 2^q-1)$ parameters. A general computing algorithm consists in Monte-Carlo (MC) sampling from each Dirichlet posterior, as suggested by [11, pp. 76–77], and [4]; this method is easy to implement using independence properties of the Dirichlet distributions (see *e.g.* [3]).

In the general theory, the IDM model requires an additional level of computational complexity since MC sampling must be performed for every Dirichlet posterior constrained by $\sum \alpha_p = 1$. But, as we saw in the previous Section, if we are only interested in statements of the type "$\delta_{p \Longrightarrow \emptyset} \geq d_{quasi}$", a single vector $\boldsymbol{\alpha}$ (composed of one "1" and $2^q - 1$ "0"'s) needs to be considered for the approximate procedure we suggest. All numerical results given further are obtained in this way.

### 5.5 Properties of the IDM

Several general properties of the IDM are given in [21] and in [22] from a predictive viewpoint. We shall only stress here some properties that are particularly important for our present purpose.

**Prior ignorance.** An important property of the IDM is that it distinguishes the case of a relative lack of information for some statement — it then produces a wide posterior probability interval — from the case of a more substantial state of knowledge — the interval is then narrower. In particular, the *prior* IDM yields vacuous probability intervals for any $q$-implication; for any pattern $p \in P$ and any $d_{quasi} \in [0, 1]$, we have

$$\begin{aligned} \underline{Prob}(\delta_{p \Longrightarrow \emptyset} \geq d_{quasi}) &= 0, \quad \text{and} \\ \overline{Prob}(\delta_{p \Longrightarrow \emptyset} \geq d_{quasi}) &= 1. \end{aligned} \quad (16)$$

**Absent patterns and inference.** A related property is that patterns that are absent in the data ($p \Longrightarrow \emptyset$ descriptively) are treated quite differently according to whether the *product frequency* associated with $p = ijk\cdots$, $\widehat{f_{ijk\cdots}} = f_i f_j f_k \cdots$, is close to 0 or not.

Consider the example in Table 1 for which both patterns $abc'$ and $ab'c$ are absent. For this protocol, the marginal frequencies of the various features are $f_a = 0.20$, $f_b = 0.15$ and $f_c = 0.75$. Pattern $abc'$ thus appears to be composed of rare modalities only and, so, the associated product frequency, $\widehat{f_{abc'}} = f_a f_b f_{c'} = 0.0075$, is small. If there was actually complete independence between $A$, $B$ and $C$, out of a sample of $n = 100$ units one would expect $\widehat{n_{abc'}} = n\widehat{f_{abc'}} = 0.75$ unit for pattern $abc'$, and, even in such a case, the observation that $n_{abc'} = 0$ would not appear surprising. The absence of pattern $abc'$ in the data does not necessarily point towards some incompatibility between modalities $a$, $b$ and $c'$, but might only result from independence between questions together with a pattern composed of rare modalities and a relatively small sample size.

The IDM expresses this uncertainty: though the descriptive step leads to the most extreme statement, $d_{abc' \Longrightarrow \emptyset} = 1$, the probability interval for the inductive statement $\delta_{abc' \Longrightarrow \emptyset} \geq 0.50$ is very wide: $[0.31, 1]$. In other words, the data allow neither to conclude that pattern $abc'$ is $q$-absent (0.31 is too low), neither to conclude in the opposite direction (1 is too high).

[3] A similar approximate procedure can be used for the upper probability $\overline{Prob}(.)$ by finding the vector $\boldsymbol{\alpha}$ which maximizes (14). In brief, the solution consists in allocating the total prior strength to a single pattern (in general) or to several patterns (in some cases) among the patterns that are neighbours of $p$ (patterns which differ from $p$ by a single feature), depending on the marginal observed frequencies of the various features.

For the absent pattern $ab'c$, on the contrary, we find $\widehat{f_{ab'c}} = 0.1275$ and $\widehat{n_{ab'c}} = 12.75$, and the hypothesis of complete independence does not appear very compatible with the observation $n_{ab'c} = 0$. Here the probability interval for $\delta_{ab'c \Longrightarrow \emptyset} \geq 0.50$ is $[1.00, 1]$ and the inductive conclusion is that pattern $ab'c$ is $q$-absent (for $d_{quasi} = 0.50$) with a guarantee close to 1.

To summarize, the IDM separates truly $q$-absent patterns in the population from unobserved patterns which were actually likely not to be observed in a small sample only because they are composed of rare modalities.

**Coherence by projection.** Another important property of the IDM in our present context is that it satisfies the *representation invariance principle*; this principle states that inferences about $\boldsymbol{\theta}$, and hence about any derived parameter $g(\boldsymbol{\theta})$, should not depend on the number of categories $K$ used for defining $\boldsymbol{\theta}$.

For example, inferences (lower and upper probabilities) relative to $\delta_{ab \Longrightarrow \emptyset}$ are identical whether, *(i)* one considers the protocol projected onto $\{A, B\}$ and uses the IDM on the $2^2$ partial patterns so defined, or, *(ii)* $\delta_{ab \Longrightarrow \emptyset}$ is decomposed into $\delta_{abc \Longrightarrow \emptyset}$ and $\delta_{abc' \Longrightarrow \emptyset}$ and the IDM is defined at the $\{A, B, C\}$ projection level, *i.e.* on the $2^3$ basic patterns. This result ensures that coherence by projection is satisfied by inferences about $q$-implications from the IDM.

## 6 Application: the "Religion data"

The following real example is taken from [7]. A sample of $n = 1524$ individuals were asked $q = 4$ binary questions about their religious opinions or behaviour. The weighted protocol on the $K = 2^q = 16$ patterns is shown in Table 2.

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

Table 2: Religion data. Weighted protocol for a binary questionnaire on $n = 1524$ individuals with $q = 4$ questions: $A$ (*Do you often pray?*), $B$ (*Do you go to church regularly?*), $C$ (*Do you believe in paradise?*), and $D$ (*Do you or will you give your children any religious education?*); "yes" answers are denoted $a$, $b$, $c$, $d$; "no" answers are denoted $a'$, $b'$, $c'$, $d'$. From a 1967 study of the "Institut Français d'Opinion Publique", presented in Degenne [7, pp. 37–39].

### 6.1 Descriptive analysis

Table 3 gives the value of the descriptive index $d_{p \Longrightarrow \emptyset}$ for each pattern $p$. If we take $d_{quasi} = 0$, 12 patterns out of 16 are $q$-absent (exceptions are $abcd$, $a'bcd$, $a'b'c'd$ and $a'b'c'd'$), so that the implicative structure in the protocol may be summarized by

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

This summary can be expressed as: "Praying often" $q$-implies both "Church attendance" and "Belief in paradise", which are $q$-equivalent and which both $q$-imply "Religious education for children". Very schematically, this list of $q$-implications seems to indicate that personal religious beliefs tend to imply religious social behaviour, but not the reverse. Of course, this first summary is rather brutal since it considers as $q$-absent any pattern which is under-represented relative to the independence case ($d_{quasi} = 0$). In fact, this summary discards more than 20% of the individuals (343 out of 1524).

| | $c$ | | $c'$ | |
|---|---|---|---|---|
| | $d$ | $d'$ | $d$ | $d'$ |
| $ab$ | −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 |

Table 3: Religion data. Descriptive index $d_{p \Longrightarrow \emptyset}$ for each pattern $p$.

Higher values for $d_{quasi}$ are likely to provide a more subtle — but also more complex — summary of the data. With the more selective choice of $d_{quasi} = 0.50$, 10 of the 12 under-represented patterns are found to be $q$-absent (the two excluded patterns are $a'b'cd$ and $abc'd$) and the implicative structure can now be descriptively summarized by

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

This second summary covers about 90% of the individuals (1367 out of 1524) and reads as follows: *(i)* "Praying often" $q$-implies "Church attendance" which in turn $q$-implies "Religious education for children"; *(ii)* "Belief in paradise" also $q$-implies "Religious education for children"; *(iii)* "Church attendance" $q$-implies either "Praying often" or "Belief in paradise".

### 6.2 Inductive analysis

Out of the 12 patterns that were found to be under-represented descriptively, all but one ($abcd'$) are also under-represented inductively, with guarantee 0.90. The implicative inductive summary of the data (for

$d_{quasi} = 0$ and $\gamma = 0.90$) may be written

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

This summary can be expressed as: *(i)* "Praying often" $q$-implies both "Church attendance" and "Belief in paradise", which are $q$-equivalent; *(ii)* "Church attendance" $q$-implies either "Praying often" or "Religious education for children".

For the reference value $d_{quasi} = 0.50$, amongst the 10 patterns that were found $q$-absent at the descriptive level, 7 are found inductively $q$-absent (for $\gamma = 0.90$) as shown in Table 4. The resulting inductive summary (for $d_{quasi} = 0.50$ and $\gamma = 0.90$) may be expressed by

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

This last summary is more difficult to express in a few simple words, because it only involves non-binary $q$-implications. Nevertheless, it contains substantial information as it indicates that 7 patterns out of the 16 possible ones are certified (with guarantee 0.90) to be highly under-represented (at least a 50% under-representation relative to the case of complete independence.)

| | $c$ | | $c'$ | |
|---|---|---|---|---|
| | $d$ | $d'$ | $d$ | $d'$ |
| $ab$ | 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 |

Table 4: Religion data. For each pattern $p$, lower probability that the $q$-implication $p \longrightarrow \emptyset$ is inductively satisfied (with $d_{quasi} = 0.50$), *i.e.* $\underline{Prob}(\delta_{p \Longrightarrow \emptyset} \geq 0.50)$. Quasi-absent patterns at guarantee $\gamma = 0.90$ appear in boldface.

### 6.3 Comments

We have given four different summaries of the same data and, clearly, others are possible by varying $d_{quasi}$ and $\gamma$. Which of them should be preferred? First of all, the descriptive summaries only answer the question: "If the data were the entire population, what would we be entailed to conclude? Varying $d_{quasi}$ produces a sequence of nested summaries, from $d_{quasi} = 1$ which yields only the absent patterns, *i.e.* the strict implications, to $d_{quasi} = 0$ which indicates which patterns are under-represented, however slightly. We suggest that considering this sequence of nested summaries is the better way to analyze the data.

When the question is how to generalize from the data to the underlying population, the answer is provided by inductive summaries. They can be thought of as filters of the corresponding descriptive summaries which filter out patterns that cannot be certified to be under-represented (at a given guarantee $\gamma$). Here, varying the value of $\gamma$ will also produce a sequence of nested summaries corresponding to a more or less strong filtering. Statistical conventions suggest using a high value for $\gamma$, such as 0.90, 0.95 or 0.99.

## 7 Conclusions

We would like to conclude by mentionning a few directions for future research:

*(1)* The method we have proposed here is suited when there is no prior information about the pattern's true frequencies. However, a questionnaire could incorporate structural constraints between the questions which indicate *a priori* incompatibilities between modalities. We think that our method could be easily adapted by specifying null prior strengths for patterns that are structurally unobservable and using the IDM on the remaining patterns. The incorporation of other types of *a priori* knowledge needs to be investigated as well.

*(2)* An inductive summary produced by the method is a joint statement where each elementary $q$-implication is obtained separately at a given guarantee. It may be of importance to investigate how to compute the lower probability of this joint statement considered as a whole. A related issue is to provide a global probability assessment for any quasi-logical model, and in particular one which would be constructed partly from the results of our method, and partly using less blind criteria such as, *e.g.*, domain knowledge or comprehensibility.

*(3)* Based as it is on a lower probability, the inductive summary is designed to be a cautious one: Certified $q$-absent patterns are the only ones it separates out. The other patterns could be separated into those that are ensured to be over-represented (or highly represented), and those for which lack of information dominates. A few of our experiments seem to indicate that this last category covers an increasing proportion of patterns as $q$ increases. Far from diminishing the interest of our method, the eventual confirmation of this fact would serve as a reminder to the analyst that having (at least) as many points as there are dimensions is a prerequisite for saying anything serious about a highly multidimensional space. This fundamental limitation would imply that our method can only produce useful results when $n >> 2^q$.

## Acknowledgements

We would like to thank Peter Walley and two anonymous referees for useful and encouraging comments that helped us to improve the paper.

## References

[1] M. Barbut, B. Monjardet. *Ordre et Classification: Algèbre et Combinatoire*, Paris: Hachette, 1970.

[2] J.-M. Bernard. Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process. *The American Statistician*, 50:7–13, 1996.

[3] J.-M. Bernard. Bayesian Analysis of Tree-Structured Categorized Data. *Revue Internationale de Systémique*, 11:11–29, 1997.

[4] J.-M. Bernard. Bayesian Inference for Categorized Data. In *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference*, by H. Rouanet *et al.*, European University Studies, Series 6: Psychology, Bern: Peter Lang, pp. 159–226, 1998.

[5] J.-M. Bernard, C. Charron. L'Analyse Implicative Bayésienne: une Méthode pour l'Etude des Dépendances Orientées. I: Données Binaires. *Mathématiques, Informatique et Sciences Humaines*, 134:5–38, 1996.

[6] J.-M. Bernard, S. Poitrenaud. L'Analyse Implicative Bayésienne Multivariée d'un Questionnaire Binaire: Quasi-implications et Treillis de Galois Simplifié. *Mathématiques, Informatique et Sciences Humaines*, in press, 1999.

[7] A. Degenne. *Techniques Ordinales en Analyse des Données*, Paris: Hachette, 1972.

[8] V. Duquenne. Contextual Implications Between Attributes and some Representation Properties for Finite Lattices. In *Beiträge zur Begriffsanalyse*, Ganter, Wille, Wolf (eds), Mannheim (Germany): Wissenschaftsverlag, pp. 213–239, 1987.

[9] V. Duquenne. On Lattice Approximations: Syntactic Aspects. *Social Networks*, 18:189–199, 1996.

[10] C. Flament. *L'Analyse Booléenne de Questionnaires*, Paris: Mouton, 1976.

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. *Bayesian Data Analysis*, London: Chapman & Hall, 1995.

[12] J.-L. Guigues, V. Duquenne. Familles Minimales d'Implications Informatives Résultant d'un Tableau de Données Binaires. *Mathématiques et Sciences Humaines*, 95:5–18, 1986.

[13] L. Guttman. A Basis for Scaling Qualitative Data. *American Sociological Review*, 9:139–150, 1944.

[14] J. B. S. Haldane. The Precision of Observed Values of Small Frequencies. *Biometrika*, 35:297–300, 1948.

[15] D. K. Hildebrand, J. D. Laing, H. Rosenthal. *Prediction Analysis of Cross Classifications*, New York: Wiley, 1977.

[16] H. Jeffreys. *Theory of Probability*, 3rd ed., Oxford: Clarendon Press, 1961.

[17] M. G. Kendall, A. Stuart. *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship*, 3rd ed., London: Griffin, 1973.

[18] J. Loevinger. The Technic of Homogeneous Tests Compared with some Aspects of Scale Analysis and Factor Analysis. *Psychological Bulletin*, 45:507–530, 1948.

[19] F. J. A. Perks. Some Observations on Inverse Probability Including a New Indifference Rule. *Journal of the Institute of Actuaries*, 73:285–334, 1947.

[20] P. Walley. *Statistical Reasoning with Imprecise Probabilities*, Monographs on Statistics and Applied Probability 42, London: Chapman & Hall, 1991.

[21] P. Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society*, Ser. B., 58:3–57, 1996.

[22] P. Walley, J.-M. Bernard. *Imprecise Probabilistic Prediction for Categorical Data.* Technical report CAF-9901, Laboratoire Cognition et Activités Finalisées, Université Paris 8, Saint-Denis (France), 1999.

[23] R. Wille. Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts. In *Symp. Ordered Sets* (I. Rival ed.), Dordrecht (Netherlands): D. Reidel, pp. 445–470, 1982.