# Estimating the size of a population from three incomplete lists: a Bayesian approach

Jérôme Dupuis

# Estimating the size of a population from three incomplete lists: a Bayesian approach

**Jérôme Dupuis**

IMT, Université Paul Sabatier, Toulouse, France

**Abstract**

We consider the problem of estimating the size $N$ of a closed population from three incomplete lists. Estimation of $N$ is based on capture-recapture type models. Our approach uses graphical models to deal with some possible dependences between the lists. Our parametrization involves marginal and conditional capture probabilities rather than clique probabilities, which facilitates the incorporation of the prior information available on capture. We link both parametrizations and we show that assuming an hyper-Dirichlet distribution for the clique parameter of any sub-model boils down to assume that some capture probabilities follow independently beta distributions (and conversely). Prior information on capture is incorporated via a descending procedure which guarantees that the prior distributions are, in a certain sense, compatible across the different models. As far as $N$ is concerned, an improper prior is usually adopted for $N$; as a result, the posterior distribution of $N$ may not exist. We provide a necessary and sufficient condition for it exists, when inference is based on a Bayesian model averaging.

**Key Words**. Bayesian estimation; Capture-recapture; Clique; Graphical models; Hyper-Dirichlet distribution; Incomplete lists; Population size.

# 1 Introduction

Estimating the size $N$ of a closed population is an important issue in several scientific fields: such as medecine, ecology, computer science (eg Pollock, 1990). As far as human populations are concerned, estimating $N$ is typically based on $J \geq 2$ incomplete lists and the resulting data are thus capture-recapture type data: an individual which appears on a given list being, in a way, 'captured' by this list (eg Hook and Regal, 1995). Owing to this analogy, the satistical analysis uses capture-recapture type models. The estimation of $N$ can be based on two lists; but, in such a case, it is not possible to take into account in the model a possible dependence between the two lists, and then there is a risk of overestimating (or underestimating) $N$ if such a dependence exists. In fact, at least three lists are necessary to model dependences between lists (eg Chao, 2015).

The precision of the estimate of $N$ based on capture recapture data may be not satisfactory, in that the confidence interval may be considered as being too large by the user. This typically happens when some capture probabilities are small (Seber, 1982). It is thus important to be able to incorporate some prior information on capture when it exits. The first (and main) objective of this paper is to propose a framework within which it is possible to incorporate some prior on the capture probabilities (marginal or conditional).

Two approaches have been proposed for modelling the possible dependences between the lists: the one of Madigan & York (1997), and the one of King & Brooks (2001) ; both implementing a Bayesian model averaging procedure for estimating $N$. The one of King & Brooks uses a log-linear ap-

proach and is particularly well suitable when prior information on capture is related to the existence and/or sign of correlation between the lists: now, it is not this kind of prior information on capture which is considered in this paper. The one of Madigan & York uses graphical models and the paper takes place in this framework. Madigan & York placed the prior distributions on the clique probabilities and limited themselves to a non informative situation (as far as capture is concerned). In fact, several authors have stressed the difficulty to incorporate -in an effective way- some prior information on capture in the Madigan & York's approach (King & Brooks, 2001), and more generally in discrete graphical models (Consonni & Leucari 2006). Let us briefly indicate how Madigan & York (1997) proceeds to point where lies the difficulty. These authors place the priors on the clique probabilities from which they derive an hyper-Dirichlet distribution (introduced by Dawid and Lauritzen, 1993) for each sub-model of the saturated model. Then, they require that priors are compatible across models, and, to achieve this objective, they derive the prior on any clique probability from the one adopted for the saturated model, by marginalisation. Therefore, in practice, the prior has to be elicitated on each of the $2^J$ joint capture probabilities, each probability being associated with a particular combination of presence/absence on the $J$ lists: for example, to appear on lists $\mathrm{L}_1$ and $\mathrm{L}_2$, but not on list $\mathrm{L}_3$ (if $J = 3$). Now, in practice, it is hard for an Expert to give some information on capture under this form. (We observe that Madigan & York (1997) do not provide any indication to determine the parameters of the Dirichlet distribution placed on the saturated model parameter if the available prior information is on the clique probabilities.) Contrary to these authors,

we do not place the priors on the cliques probabilities, but on the marginal and conditional capture probabilities; arguing that this greatly facilitates the elicitation of prior beliefs, inso far as these probabilities are quantities which have a concrete meaning and that praticians are used to manipulate (eg Dupuis, 1995). We show that there is a one-to-one and onto correspondence between our parametrization and the Madigan & York's one. We also show that assuming an hyper-Dirichlet for a given sub-model comes down to assume that the marginal and conditional capture probabilities present in this sub-model follow independently beta distributions, and conversely. This key result allows to link both approaches. As Madigan & York (1997) we require that priors are compatible across models, and we adopt a similar strategy to do so. In the framework of our approach, elicitating the prior comes down to determine the $2^J$ parameters of the Dirichlet distribution placed on the parameter of the saturated model, while the available prior information is on the marginal and conditional capture probabilities; a specific *descending* procedure is developed in this aim. This procedure easily implements when $J = 3$, but it becomes tedious to implement beyond this value; that is why we will afterwards limit ourself to this important particular value.

It is possible that, for a sub-model $m$, the posterior distribution does not exist when a non informative prior is adopted for $N$. We note that this possibility is not considered by Madigan and York. The second objective of this paper is thus to give conditions ensuring the existence of the posterior distribution for each sub-model. We also provide a necessary and sufficient condition so that the model-averaged posterior distribution of $N$ exists. Interestingly, this condition applies to the approach of Madigan and York.

## 2 Data description

For any individual $i$ of the population of interest, we denote by $\mathbf{x}_i$ the vector $(x_{ij}; j = 1, \ldots, 3)$ where $x_{ij} = 1$ if individual $i$ appears on list $j$ and zero otherwise; $\mathbf{x}_i$ is called the history of individual $i$. There are 8 possible histories, namely: (0 0 0), (0 0 1), (0 1 0), (0 1 1), (1 0 0), (1 0 1), (1 1 0), (1 1 1) which are, for convenience, numbered in this order from 0 to 7. For example, the history (0 0 0) (or equivalently the history 0) is the one of a individual which appears on no list, and (1 1 1) (or equivalently the history 3) is the one of a individual which appears on the three lists. We denote by $n_h$ the number of individuals whose history is $h$. Note that the count $n_0$ is not observable and that $d = \sum_{h \neq 0} n_h$ represents the number of individuals appearing in at least one list. Data consist in the seven counts $\{n_h; h = 1, \ldots, 7\}$.

## 3 Assumptions, models and parameters

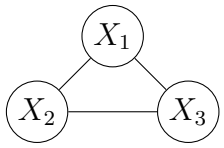
*Assumption* A1. We assume that $\mathbf{X}_1, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_N$ are independent and identically distributed.

It remains to specify the probabilistic assumptions made on the three random variables $X_{i1}$, $X_{i2}$, and $X_{i3}$ (index $i$ being fixed). These assumptions consist in conditional independance assumptions and are represented by a graph. Because the assumptions made on $X_{i1}$, $X_{i2}$, and $X_{i3}$ (index $i$ being fixed) do not depend on $i$ (cf Assumption A1), it is convenient to omit in what it follows index $i$ in these three random variables. As in the Madigan and York approach, eight models are considered: the saturated model and seven sub-models obtained by removing one or several arrows in the graph

of the saturated model: see below.

• The saturated model is denoted by [123].

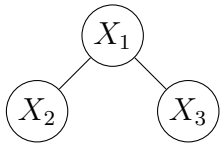
It is characterized by the fact that no independence assumption concerning $X_1$, $X_2$, and $X_3$ is made. The graph of the saturated model [123] is given below:




The saturated model is parametrized by the $\theta_h$'s where $\theta_h$ denotes the probability that an individual has $h$ as history; so, $\theta_{111}$ represents the probability that an individual appear on the three lists; note that $\sum_h \theta_h = 1$.

• The models $[12, 13]$, $[21, 23]$, and $[31, 32]$ (called models of type I).

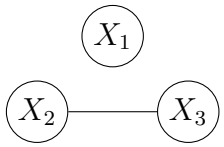
For example, the model $[12, 13]$ assumes that $X_2 \perp X_3 | X_1$. In other terms, we assume that $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$. The graph of the model $[12, 13]$ corresponding to thus assumption is:




since an absence of edge between two vertices (2 and 3 in our case) means that the corresponding random variables (namely $X_1$ and $X_2$) are independent conditionaly on the remaining ransom variable(s) (namely $X_1$ in our case); see eg Lauritzen (1996).

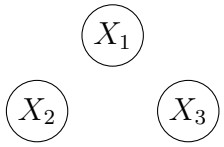
• The models $[1, 23]$, $[2, 13]$ and $[3, 21]$ (called models of type II).

For example, the model $[1, 23]$ assumes that $X_1 \perp (X_2, X_3)$. In other terms, we assume that $p(x_1, x_2, x_3) = p(x_1)p(x_2, x_3)$. The graph of this model is given below:




• The independant model, denoted by $[1, 2, 3]$.

It assumes that the three random variables $X_1$, $X_2$, and $X_3$ are independant. Thus, one has $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$. The graph of the independant model is given below.




The parametrization of the sub-models of type I and II involve marginal and conditional capture probabilities, while the independent model involve only marginal probabilities (considering the decompostion of $p(x_1, x_2, x_3)$ in these models). To clarify this point we introduce the following notation. Marginal capture probabilities are denoted by $q_r$ which represents the probability that an individual appears on list $r \in \{1, 2, 3\}$. Conditional capture probabilities are denoted as follows: $\mu_{s|r}$ represents that an individual appears on list $s$ given that it appears on list $r \neq s$, and $\lambda_{s|r}$ represents that an individual appear on list $s$ given that it does not appear on list $r \neq s$. We thus define twelve conditional probabilities and three marginal probabilities.

With this notation, the independent model includes three parameters, namely $q_1$, $q_2$, $q_3$. Models of type I includes five parameters. For example,

the parameters of the model [12, 13] are $q_1$, $\lambda_{2|1}$, $\mu_{2|1}$, $\lambda_{3|1}$ and $\mu_{3|1}$. As far as models of type II are concerned, two parametrizations are possible. For example, if one considers the model [1, 23], one can decompose $p(x_1, x_2, x_3)$ as $p(x_1)p(x_2)p(x_3|x_2)$ or as $p(x_1)p(x_3)p(x_2|x_3)$. In the first case, the resulting parametrization includes parameters $q_1$, $q_2$, $\lambda_{3|2}$, $\mu_{3|2}$, and parameters $q_1$, $q_3$, $\lambda_{2|3}$ and $\mu_{2|3}$ in the second case.

# 4 Prior on capture

## 4.1 Properties of the beta and Dirichlet distributions.

Let $\alpha \in [0, 1]$ be a parameter of a statistical model. In a Bayesian context, it is usual to adopt a beta distribution for $\alpha$. It is well known that it exists three distinct parametrisations for a beta distribution (each being usefull). The standard parametrisation corresponds to the density $\pi(\alpha) \propto \alpha^{b_1-1}(1-\alpha)^{b_2-1}$ if $\alpha \sim$ beta $(b_1, b_2)$. The second parametrisation involves $(M, S)$ where $M = b_1/(b_1 + b_2)$ represents the prior mean of $\alpha$ and $S = b_1 + b_2$, and the third parametrisation involves $(M, V)$ where $V$ denotes the prior variance of $\alpha$. The sum $S$ interprets as the precision (or the strength) of the prior available on $\alpha$ since $S$ and $V$ are linked by the relation $1 + S = M(1 - M)/V$; in particular, $M$ being fixed, small values of $S$ are associated with non informative priors: for example, $S$ is equal to 2 if $\alpha$ follows a uniform distribution, and to 1 if $\alpha \sim$ beta$(1/2, 1/2)$ (the Jeffreys prior). The precision of the prior attached to a parameter $\alpha \in [0, 1]$ will be denoted by $\mathcal{P}(\alpha)$.

Let $\alpha = (\alpha_1, \ldots, \alpha_k)$ be now a vector of probabilities which sum to 1, and assume that $\alpha$ represents the parameter of a statistical model. In a Bayseian context, it is usual to adopt a Dirichlet distribution for $\alpha$. If $\alpha \sim$

Dirichlet$(b_1, \ldots, b_k)$ the sum $S$ of these coefficients similarly interprets as the precision of the prior; small values of $S$ being associated with non informative priors.

We now recall a usefull property of the Dirichlet distribution, called the *agregation property*; it states as follows. If one assumes that

$$\alpha = (\alpha_1, \ldots, \alpha_k) \sim \text{Dirichlet}(b_1, \ldots, b_k)$$

with $k \geq 3$, and that $\{1, \ldots, k\}$ is partitioned into two parts $\mathcal{P}_1$ and $\mathcal{P}_2$, then

$$(\sum_{i \in \mathcal{P}_1} \alpha_i, \sum_{i \in \mathcal{P}_2} \alpha_i) \sim \text{ beta}(\sum_{i \in \mathcal{P}_1} b_i, \sum_{i \in \mathcal{P}_2} b_i).$$

Hence, for $1 \leq j \leq k$, one has $\alpha_j \sim$ Beta $(b_j, b - b_j)$ where $b = \sum_{i=1}^{k} b_i$.

## 4.2 Prior distributions on the capture probabilities

For each sub-model $m$, we assume that the marginal and conditional capture probabilities present in sub-model $m$ are a priori independent. For example, if $m = [12, 13]$, we assume that $q_1$, $\mu_{3|2}, \lambda_{2|1}, \mu_{3|1}, \lambda_{3|1})$ are a priori independant.

We note that certain marginal and conditional capture probabilities are linked; one has indeed:

$$q_r = \mu_{r|s} q_s + \lambda_{r|s}(1 - q_s), \qquad (1 - \mu_{s|r}) q_r = \lambda_{r|s}(1 - q_s) \tag{4.1}$$

and

$$q_r \mu_{s|r} = q_s \mu_{r|s}, \qquad (1 - \lambda_{r|s})(1 - q_s) = (1 - \lambda_{s|r})(1 - q_r) \tag{4.2}.$$

Surprisingly, no a priori independence assumption (made above) comes into conflict with the identities appearing in (4.1) and (4.2). Indeed, each of them

involve four distinct parameter which are never simultaneously present in any given sub-model. As far as models of type I are concerned this point is easily checked by observing that each identity involves two marginal probabilites while a model of type I includes only one marginal probability. As far models of type II are concerned, it is sufficient to observe that any model of type II includes four parameters ($q_r$, $q_s$, $\mu_{s'|s}$, $\lambda_{s'|s}$ where $r$, $s$ and $s'$ are all different) which are never simultaneously present in the above identities. Moreover, we adopt a beta distribution for each elementary parameter; we mean by elementary parameter, any one of the the marginal or conditional capture probability. An elementary parameter will be afterwards denoted by $\theta_e$.

We denote by $\theta_{\text{Sat}}$ the global parameter of the saturated model. As in the paper of Madigan & York (1997), we adopt for $\theta_{\text{Sat}}$ a Dirichlet distribution which is the standard prior distribution for a parameter lying in a simplex. More precisely we assume that:

$$\theta_{\text{Sat}} \sim \text{Dirichlet}\,(a_{000}, a_{001}, a_{010}, a_{011}, a_{100}, a_{101}, a_{110}, a_{111}).$$

We thus assume that:

$$\pi(\theta_{\text{Sat}}) \propto \prod_{h \in \mathcal{H}} \theta_h^{a_h - 1}$$

where $\mathcal{H}$ denotes the set of the eight possible histories (see Section 2). The precision of the Dirichlet is denoted by $a$, thus on has $a = \sum_h a_h$.

For $j, k$ in $\{0, 1\}$, it is convenient to introduce the following notations:

$$\theta_{jk+} = \sum_{l=0}^{1} \theta_{jkl} \quad \text{and} \quad \theta_{j++} = \sum_{k=0}^{1} \theta_{jk+},$$

where $\theta_{jkl}$ denotes the probability that a case has $jkl$ as history. Notations $\theta_{j+l}$, $\theta_{+kl}$, $\theta_{+k+}$ and $\theta_{++l}$ are defined similarly.

As far as a model averaging procedure is implemented for estimating $N$, it is strongly desirable that the priors are -as much as possible- compatible across the different models (eg Dawid and Lauritzen, 1993). In particular, it is rather natural to require that the priors put on elementary parameters are not chosen independently of the one put on $\theta_{\text{Sat}}$, because all expresses in function of the $\theta_h$'s; for example, one has:

$$q_1 = \theta_{1++} \quad \text{and} \quad \mu_{2|1} = \frac{\theta_{110} + \theta_{111}}{\theta_{1++}}.$$

A simple way to meet this requirement, is to deduce (by marginalisation) the prior on any elementary parameter from the prior placed on $\theta_{\text{Sat}}$ (see the Proposition 3, below). An analoguous strategy has been adopted by Dawid & Lauritzen (1993) and by Madigan and York (1997) which deduced (by marginalisation) the prior distribution on a clique probability of any graph (different from the complete one) from the one adopted for the complete graph (saturated model).

The following Proposition allows to derive the prior distribution on any marginal and conditional capture probabilities from the one placed on $\theta_{\text{Sat}}$.

**Proposition 1**. For any $j, k \in \{0, 1\}$, we have: $\theta_{j++} \sim \text{beta}(a_{j++}, a - a_{j++})$ and $\theta_{jk+}/\theta_{j++} \sim \text{beta}(a_{jk+}, a_{j++} - a_{jk+})$ with obvious notations for $a_{jk+}$ and $a_{j++}$.

*Proof.* See Appendix A.

Similar results of course hold for $\theta_{+k+}$, $\theta_{++l}$, and for all the possible ratios similar as the one appearing in Proposition 1. Fixing the hyperparameters $a_h$ induces a prior distribution for all the marginal and conditional capture probabilities, since $q_1 = \theta_{1++}$, $q_2 = \theta_{+1+}$, $q_3 = \theta_{++1}$, and that, for example,

$\mu_{2|1} = \theta_{11+}/\theta_{1++}$ and $\lambda_{2|1} = \theta_{01+}/\theta_{0++}$.

During the prior elicitation, it is usual to set $E[\theta_e] = \theta_e^*$ where $\theta_e^*$ denotes the estimation of $\theta_e$ furnished by the Expert (e.g. O'Hagan, 1998). In other respects, it would be also strongly desirable (for evident reasons of consistency) that the above identities (4.1) and (4.2) also hold when one replaces each parameter $\theta_e$ by its estimation $\theta_e^*$. Now, surprisingly, all these identities hold when the parameters are replaced by the corresponding prior means (see the second part of the Proposition 2 below). This result shows that deriving the priors distributions on the $\theta_e$'s from the Dirichlet placed on $\theta_{\text{Sat}}$ induces another kind of compatibility between sub-models which is particularly satisfactory.

The fact of deriving the pior distributions of the $\theta_e$'s from the Dirichlet distribution placed on $\theta_{\text{Sat}}$ induces some constraints on the resulting beta laws: they appear in the Proposition 3 below.

**Proposition 2**. For all $r, s, t$ in $\{1, 2, 3\}$, one has:

$$\mathcal{P}(q_s) = \mathcal{P}(q_t), \qquad \mathcal{P}(\mu_{s|r}) = \mathcal{P}(\mu_{t|r}), \qquad \mathcal{P}(\lambda_{s|r}) = \mathcal{P}(\lambda_{t|r})$$

and

$$\mathcal{P}(\mu_{s|r}) + \mathcal{P}(\lambda_{s|r}) = \mathcal{P}(q_r).$$

Moreover, for all $r$ and $s$ one also has:

$$E[q_r] = E[\mu_{r|s}]E[q_s] + E[\lambda_{r|s}]E[1 - q_s], \quad E[\mu_{s|r}]E[q_r] = E[\mu_{r|s}]E[q_s]$$

and

$$E[1-\mu_{s|r}]E[q_r] = E[\lambda_{r|s}]E[1-q_s], \quad E[1-\lambda_{r|s}]E[1-q_s] = E[1-\lambda_{s|r}]E[1-q_r].$$

*Proof.* It is easily deduced from the Proposition 1, so details are omitted.

## 4.3 The links with the Madigan-York parametrisation.

For each fixed sub-model $m$, there is a one-to-one and onto correspondence between the Madigan & York parametrisation and ours (details appear in Appendix B). Moreover, the Proposition 3 below clarifies the links between the hyper-Dirichlet distribution adopted by Madigan & York as prior, and the priors we place on the elementary parameters. In the statement of this Proposition, we mean by *compatible* hyper-Dirichlet that the prior distributions placed on the cliques probabilities of sub-model $m$ are derived from the Dirichlet distribution placed on $\theta_{\text{Sat}}$, via marginalisation; similarly, the marginal and conditional capture probabilities are said to be compatible if they are similarly derived from the Dirichlet placed on $\theta_{\text{Sat}}$.

**Proposition 3**. Sub-model $m$ being fixed, if one adopts a compatible hyper-Dirichlet distribution as prior for the clique parameter of sub-model $m$, thus the marginal and the conditional capture probabilites playing a part in sub-model $m$ follow independently beta distributions which are all compatible; moreover the converse holds.

*Proof.* As far as the models of type I and II are concerned, the proof appears in Appendix B. For the independant model, the proposition is trivial.

## 4.4 A descending procedure.

The objective of this Section is to propose a procedure to determine the parameters $a_h$ of the Dirichlet distribution placed on $\theta_{\text{Sat}}$ from the prior information available on the marginal and conditional capture probabilities (while satisfying the constraints appearing in Proposition 3). The $a_h$'s being the unknowns, the problem is in fact to find a collection $\mathcal{C}$ of such prior

quantities which leads to a unique solution; note that we have to solve a system with eight unknowns. Our procedure proceeds as follows.

• First, we begin by including in $\mathcal{C}$ the prior means of $q_1$, $q_2$, $q_3$ derived from the estimations $q_1^*$, $q_2^*$, $q_3^*$ furnished by the Expert.

• Second, for each of the three pairs $\{r, s\}$ where $r$ and $s$ belongs to $\{1, 2, 3\}$, we include in $\mathcal{C}$ the prior mean of any conditional capture probability chosen among $\{\mu_{s|r}, \mu_{r|s}, \lambda_{s|r}, \lambda_{s|r}\}$; each being derived from the corresponding estimation given by the Expert. It is easy to check that the prior means of the remaining conditional probabilities are in fact determined by this choice made upstream (owing to Proposition 2). Thus, at the end of this second step, $\mathcal{C}$ includes six prior means.

• Third, we include in $\mathcal{C}$ the global precision $a$ (see below for details). At the end of this step, it is easy to check that all the other constraints appearing in the first part of the Proposition 2 are automatically fulfiled; and the parameters of the beta distributions placed on the elementary parameters are all determined.

This procedure is called *descending procedure* because we begin by including the marginal capture probabilities for which the precision of the prior is maximal (namely $a$), and we pursue with the conditional capture parameters for which it is lesser. This descending procedure leads thus to a system with 7 equations (see Appendix D). An additional equation coming from some prior information put on a particular $\theta_h$ is thus necessary to have any hope that the system has a unique solution. In practice, it is expected that the Expert is able to furnish an estimation for the parameter $\chi = 1 - \theta_{000}$ which represents the probability that an individual of the target population appears

on at least on a list, and interprets as the global efficiency of the three lists. In such a situation, the following Proposition allows us to conclude.

**Proposition 4**. Assume that the prior consists of: $\mathrm{E}[q_r]$ for each $r = 1, 2, 3$; a value for the global precision of the prior (that is for $a$); a prior mean for one parameter among $\{\mu_{s|r}, \mu_{r|s}, \lambda_{s|r}, \lambda_{r|s}\}$, and that, for each pair $\{r, s\}$, where $r$ and $s$ are distinct; a prior mean for the parameter $\chi = 1 - \theta_{000}$. For such a prior, the determinant of the system is equal to 1, and the problem has thus a unique solution.

*Proof.* It appears in Appendix C.

We stress that there are no constraint between the different prior means appearing in Proposition 5; this is, from a practical point of view, an attractive characteristic of our procedure. From the Expert point of view, the constraints on the prior means are in fact transparent: he has simply to furnish seven estimations, namely: $q_1^*$, $q_2^*$, $q_3^*$, $\chi^*$ and, for example, $\mu_{2|1}^*$, $\mu_{3|2}^*$, $\mu_{1|3}^*$. A possible scenario for fixing the global precision $a$ is as follows. For $\chi$ and for each capture probability $q_r$, one collects from the Expert the prior mean and a 95% prior credible interval, and we derive for each parameter the beta distribution having these characteristics, as in Dupuis (1995). There is of course no reason that the resulting beta distributions will have the same precision. To deal with this constraint we simply advocate to retain for $a$ the smaller precision of the four beta distribution (even if that means a global loss of precision). Once the system (having the $a_h$'s as unknown) is solved, it is thus possible to derive (from the Proposition 1) the prior distributions of the parameters having play no part during the elicitation of the prior by the Expert.

The Proposition 4 shows that the descending procedure does not lead in fact to a unique collection $\mathcal{C}$. Moreover, it is clear that other procedures can be viewed to constitute a collection. For example, if the prior information consists in: $\mathrm{E}[\mu_{2|1}]$, $\mathrm{E}[\lambda_{2|1}]$, $\mathrm{Var}[\mu_{2|1}]$, $\mathrm{Var}[\lambda_{2|1}]$, $\mathrm{E}[\mu_{2|3}]$, $\mathrm{E}[\mu_{1|3}]$, $\mathrm{E}[q_3]$, and $\mathrm{E}[1-\theta_{000}]$, the resulting system has again a unique solution (for brievety the proof is omitted). We do not go futher in this direction.

# 5 Priors on the size $N$ of the population.

First, we assume that, for all models $m$, $N$ and $\theta_m$ are a priori independent, thus one has: $\pi(N, \theta_m) = \pi(N)\pi(\theta_m)$. In the absence of any prior information on $N$, one usually adopts either the Jeffreys prior $\pi(N) = 1/N$, or the uniform prior $\pi(N) = 1$ (eg Basu and Ebrahimi, 2001). note that both are improper. In an informative context, we could adopt a negative binomial distribution as in Dupuis and Goulard (2011).

# 6 Conditions of existence of $\pi(N|\mathbf{y}, m)$.

Recall that the estimation of $N$ resulting from a Bayesian model averaging procedure is based on the posterior distribution of $N$ (that is the distribution of $N|\mathbf{y}$). One has:

$$\pi(N|\mathbf{y}) = \sum_m \pi(N|\mathbf{y}, m)p(m|\mathbf{y}) \tag{5.1}$$

where $p(m|\mathbf{y})$ represents the posterior probability of model $m$, and $\pi(N|\mathbf{y}, m)$ the posterior density of $N$ under model $m$. Note that the distribution of $N|\mathbf{y}$ is also called the averaged-model posterior distribution of $N$, considering (5.1). For obtaining a condition of existence of $N|\mathbf{y}$ we need first to state a

condition of existence of $\pi(N|\mathbf{y}, m)$ for each model $m$; it is the objective of this Section.

Using the Bayes formula, one has

$$\pi(N|\mathbf{y}, m) \propto p(\mathbf{y}|N, m)\pi(N),$$

with

$$p(\mathbf{y}|N, m) = \int_{\Theta_m} L(\theta_m, N; \mathbf{y})\pi(\theta_m)\,\mathrm{d}\,(\theta_m), \tag{5.2}$$

where $L(\theta_m, N; \mathbf{y})$ denotes the likelihood of data $\mathbf{y} = \{n_h; h = 1, \ldots, 7\}$ under model $m$, and where $\theta_m$ denotes the global parameter of model $m$.

For obtaining $L(\theta_m, N; \mathbf{y})$ we have to compute $\Pr(\mathbf{Y} = \mathbf{y}|\theta_m, N)$. From Assumption A1 we deduce that:

$$(N_1, \ldots, N_7)|N, \theta_m \sim \text{Multinomial}(N; \theta_1, \ldots, \theta_7)$$

where $\theta_h$ easily expresses in function of the capture parameters present in model $m$ (see below). It follows that:

$$L(\theta_m, N; \mathbf{y}) \propto \frac{N!}{(N-d)!}\left[1 - \sum_{h=1}^{7} \theta_h\right]^{N-d} \prod_{h=1}^{7} \theta_h^{n_h}. \tag{5.3}$$

In the above expression of the likelihood, the term $\prod_{h=1}^{7} n_h!$ has been omitted, keeping thus only the part useful for inference. The likelihood of data $\mathbf{y}$ under model $m$ is now easily derived from (5.2) and from the assumptions giving the dependence structure between $X_1$, $X_2$ and $X_3$ (cf Section 2). The expressions of the different likelihoods appear in Annex E.

For obtaining $p(\mathbf{y}|N, m)$ one has to integrate the likelihood over $\theta_m$. As in Madigan and York (1997), this integral, denoted by $I_m(N)$, can be write down in a closed form (for each model $m$). which allows us to to obtain an

explicit condition for the existence of the posterior distribution of $N$ under each model $m$. We provide in Annex D, the expressions of $I_m(N)$.

In Proposition 5 below, we provide a necessary and sufficient condition so that the posterior distribution of $N$ exists under model $m$. It is assumed that no prior information is available of $N$, and thus an improper prior has been adopted for $N$; in our statement, $t = 1$ is associated with the Jeffreys prior $\pi(N) = 1/N$, and $t = 0$ with the uniform prior $\pi(N) = 1$.

**Proposition 5**. The posterior distribution of $N$ exits:

- under the saturated model, if and only if, $t + a - a_{000} > 1$,
- under $[12, 13]$, if and only if, $t + n_{011} + (a - a_{000}) + a_{011} > 1$,
- under $[1, 23]$, if and only if, $t + d_1 - n_{100} + (a - a_{000} + a_{1++} - a_{100}) > 1$,
- under $[1, 2, 3]$, if and only if, $t + (d_1 + d_2 + d_3) - d + a_{1++} + a_{+1+} + a_{++1} > 1$.

*Proof.* It appears in Appendix E only for the models $[12, 13]$ and $[123]$; for the other models, the developments are very similar, and the proof has been omitted for brievety.

For obtaining the statement of Proposition 6, when the posterior mean of $N$ is of interest, replace 1 by 2 in the right member of inequalities.

Before commenting these results, let us recall the main non informative priors distributions for a parameter which lies in the simplex of $\mathbb{R}^{2^3}$; it is the case of $\theta_{\text{Sat}}$. One usually adopts the uniform distribution, that is a Dirichlet $(1, \ldots, 1)$; it is the prior adopted by Madigan and York (1997). In the litterature, one meets mainly three alternatives to the uniform prior distribution: the Jeffrey prior which corresponds to a Dirichlet $(1/2, \ldots, 1/2)$, the Perks prior which corresponds to a Dirichlet $(1/2^3, \ldots, 1/2^3)$, and the

improper Haldane prior, defined as follows:

$$\pi(\theta_{\text{Sat}}) \propto \prod_{h \in \mathcal{H}} \frac{1}{\theta_h}.$$

In this paper, we rather advocate to use a Dirichlet $(1/4, \ldots, 1/4)$, arguing that the resulting conditional and marginal capture probabilities follows non informative priors; it is indeed easy to check (by applying the Proposition 1) that any marginal capture probability follows a uniform distribution and any conditional capture probability follows a Jeffreys distribution. Note that the prior distribution adopted by Madigan and York (1997) does not share this appealing property since $q_j \sim \text{beta}(4, 4)$ which is far from to be non informative.

We now comment the results of existence stated in Proposition 5.

• When some prior information prior is available on $\theta_{\text{Sat}}$, these conditions will be typically all satisfied whatever the value of $t \in \{0, 1)$; consequently, the Proposition 5 has mainly an interest when a non informative prior has been considered for $\theta_{\text{Sat}}$.

• Proposition 5 shows that the posterior distribution (or the posterior mean) of $N$ may not exist when the data are scarce. For example, under model $[12, 13]$, if the count $n_{011}$ is null and if one uses the uniform prior for $N$, then the posterior distribution of N does not exit when one adopts the Perks non informative prior distribution for $\theta_{\text{S}}$, since $t+n_{011}+(a-a_{000})+a_{011} = 1$. If, one only changes the prior on $N$ by adopting now the Jeffreys prior, the posterior distribution of N exists, but the the posterior mean of N does not exits, since $t + n_{011} + (a - a_{000}) + a_{011} = 2$. If one uses the Jeffreys prior for $N$ and the prior we propose in Section 3 for $\theta_{\text{Sat}}$ the posterior mean of $N$ exits (and thus

also the posterior distribution) of $N$, since $t + n_{011} + (a - a_{000}) + a_{011} = 3$.

• When the improper Haldane prior is used for $\theta_{\text{Sat}}$ all the $a_h$ are null (as well as $a$). Consequently, whatever the improper prior adopted for $N$ (namely $\pi(N) = 1/N$ or $\pi(N) = 1$) and the improper t as well as the , .

• The result appearing in Proposition 5 also applies of course to the Madigan & York approach, because there is a one-to-one and onto correspondence between their parametrisation and ours; this particular point can be also checked by starting from the formula they give for $I_m(N)$ in their paper.

• As far the model $[1, 23]$ is concerned, the above result shows that the obtained condition does not depend on the factorization adopted for $p(x_1, x_2, x_3)$ since each of the terms $n_{011}$, $a_{000}$ and $a_{011}$ remain unchanged when one permutes the index $k$ (related to list 2) and $l$ (related to list 3).

• The conditions for models $[23, 21]$, $[31, 32]$, $[2, 31]$, and $[3, 12]$ are obtained by an appropriate permutation of the indices $j$, $k$, $l$ in $a_{jkl}$ and $n_{jkl}$ (details are omited).

# 7 Condition of existence of $\pi(N|\mathbf{y})$.

As in the previous Section, it is assumed that no information is available on $N$, and $t = 0$ is associated with the uniform prior, $t = 1$ with the Jeffeys prior.

**Proposition 6**. The averaged-model posterior distribution of $N$ exists if and only if $t + a - a_{000} > 1$.

*Proof.* It appears in Appendix F.

If the averaged-model posterior mean of $N$ if of interest, one obtains a

similar result by replacing the above condition by $t + a - a_{000} > 2$.

We now comment these two results. As previously, when some prior information prior is available on $\theta_{\text{Sat}}$, these two conditions will be typically satisfied, whatever the value of $t \in \{0, 1\}$; The comments below concern situations for which no prior information on capture is available.

We note that the data plays no part in the statement of Proposition 6. From Proposition 6, we deduce that, when one uses a uniform prior for $N$ or the non informative Perks prior for $\theta_{\text{Sat}}$, the posterior distribution of $N$ does not exist, while it exits when the one we propose in Section 4.5 is adopted. Moreover, if one uses the Jeffrey prior for $N$ and the Perks non informative prior distribution for $\theta_{\text{Sat}}$, the posterior mean of $N$ does not exist, while it exists with ours; it why we consider that ours is here preferable to the Perks's one.

In fact, if the precision $a$ of the Dirichlet distribution placed on $\theta_{\text{Sat}}$ is taken too large, the condition $t + a - a_{000} > 2$ will be automatically satisfied (whatever the value of $t \in \{0, 1\}$), but the resulting priors on the marginal capture probabilities may not be non-informative (it is what it occurs with the uniform prior where $a = 8$); in contrast, if $a$ is taken too small, the resulting priors on the marginal capture probabilities will be this time non-informative, but the condition $t + a - a_{000} > 2$ will be not satisfied (it is what it occurs with the Perks prior, when $\pi(N) = 1$). The precision $a = 2$ we have retained for the Dirichlet solves, in a way, this conflict.

## References


Basu and Ebrahimi (2001). Bayesian capture-recapture methods methods for error detection and estimation of population size: heterogeneity and dependence. *Biometrika* **88**, 269-279.

Chao, A. (2015). Capture-recapture for human populations. Wiley StatRef: Statistics Reference Online, 1-16.

Consonni and Leucari (2006). Reference prior for discrete graphical models. *Biometrika* **93**, 23-40.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable models. *Annals of Statistics*. **21**, 1272-1317.

Dupuis, J. A. (1995). Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika* **82**, 761-772.

Dupuis, J. A. and Goulard, M. (2011). Estimating species richness from quadrat sampling data: a general approach. *Biometrics* **67**, 1489-1497

Geiger, D. & Heckerman, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Annals of Statistics* **25**, 1344-1369.

Hook, E. and Regal, R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Review* **17**, 243-64.

Jonhson, N. L., and Kotz, S. (2000). *Distributions in Statistics. Continuous Mutivariate Distributions.* Wiley, New York.

Kai Wang N., Guo-Liang T., Man-Lai T. (2011). *Dirichlet and related distributions*. Wiley, New York.

King, R. and Brooks, S.P. (2001) On the Bayesian estimation of population size. *Biometrika* **88**, 841-851.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford.

Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a closed of population. *Biometrika* **84**, 19-31.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician*. VoL. 47, No 1, 21-35.

Pollock, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for wildlife populations: past, present, and future. *J. Am. Statist. Asso.* **86**, 225-238.

Seber, G. A. F. (1982) *The estimation of animal abundance and related parameters*, 2nd edition. London: Griffin.

## Appendix A

From the *agregation property* we deduce that

$$(\theta_{1-j,+,+}, \theta_{jk+}, \theta_{j,1-k,+}) \sim \text{Dirichlet}(a_{1-j,+,+}, a_{jk+}, a_{j,1-k,+})$$

for all $j, k \in \{0, 1\}$. The second part of the Proposition uses the following property: if

$$(\alpha_0, \alpha_1, \ldots, \alpha_k) \sim \text{Dirichlet}(b_0, b_1, \ldots, b_k)$$

then

$$\frac{\alpha_j}{\sum_{i=1}^{k} \alpha_i} \sim \text{beta } (b_j, \sum_{i=j+1}^{k} b_i)$$

for all $j = 1, ..., k$: see Jonhson & Kotz (2004). From this property, we deduce that if $(\alpha_0, \alpha_1, \alpha_2) \sim \text{Dirichlet}(a_0, a_1, a_2)$ then $\alpha_1/(\alpha_1 + \alpha_2) \sim$ beta $(a_1, a_2)$. Consequently, one has $\theta_{jk+}/\theta_{j++} \sim \text{beta}(a_{jk+}, a_{j++} - a_{jk+})$ since $\theta_{jk+} + \theta_{j,1-k,+} = \theta_{j++}$ and $a_{j,1-k,+} = a_{j++} - a_{jk+}$.

## Appendix B

Proposition 4 is first proved with model $[12, 13]$, then with model $[1, 23]$. Moreover, let us recall that:

$$\theta_{\text{Sat}} \sim \text{Dirichlet}(a_{000}, a_{001}, a_{010}, a_{011}, a_{100}, a_{101}, a_{110}, a_{111}).$$

**I.** The model $[12, 13]$

The graph associated with the model $m = [12, 13]$ includes two cliques $\{C_1, C_2\}$, namely $C_1 = \{1, 2\}$ and $C_3 = \{1, 3\}$, and one separator $S = \{1\}$. One has

$$\theta_{C_1} = \{\theta_{00+}, \theta_{01+}, \theta_{10+}, \theta_{11+}\} \quad \text{and} \quad \theta_{C_2} = \{\theta_{0+0}, \theta_{0+1}, \theta_{1+0}, \theta_{1+1}\}$$

and $\theta_S = \{\theta_{0++}, \theta_{1++}\}$.

• It is assumed that the clique parameter $(\theta_{C_1}, \theta_{C_2})$ follows a compatible hyper-Dirichlet distribution.

The assumption of compatibility and the *agregation* property of the Dirichhlet distribution implies that:

$$\theta_{C_1} \sim \text{Dir}(a_{00+}, a_{01+}, a_{10+}, a_{11+}) \quad \theta_{C_2} \sim \text{Dir}(a_{0+0}, a_{0+1}, a_{1+0}, a_{1+1}) \qquad (1).$$

Moreover, the Madigan & York (1997) approach requires that the prior on $(\theta_{C_1}, \theta_{C_2})$ is such that $\theta_{C_1} \perp \theta_{C_2}|S$. In these conditions, the density of the hyper-Dirichlet placed on $(\theta_{C_1}, \theta_{C_2})$ is as follows:

$$\pi(\theta_{C_1}, \theta_{C_2}) \propto \frac{\prod_{j,k} \theta_{jk+}^{a_{jk+}-1} \prod_{j,l} \theta_{j+l}^{a_{j+l}-1}}{\prod_j \theta_{j++}^{a_{j++}-1}} \qquad (2).$$

Note that the parameters of the above hyper-Dirichlet distribution are all pairwise hyperconsistent (Dawid and Lauritzen, 1993), since the following

constraints:

$$\sum_k \alpha_{jk+} = \sum_l a_{j+l} = a_{j++},$$

are satisfied, for all fixed $j$. Due to these constraints, the hyper-Dirichlet distribution includes in fact only six independant parameters. Note that similar constraints exist on the clique parameters, since it is clear that:

$$\theta_{00+} + \theta_{01+} = \theta_{0+0} + \theta_{0+1} = \theta_{0++} \quad \text{and} \quad \theta_{10+} + \theta_{11+} = \theta_{1+0} + \theta_{1+1} = \theta_{1++}.$$

Considering that $\theta_{0++} + \theta_{1++} = 1$, one has in fact five independant (that is unconstrained) parameters, as in our parametrization; for example: $\theta_{1++}$, $\theta_{11+}$, $\theta_{01+}$, $\theta_{1+1}$, $\theta_{0+1}$.

Our parametrization and the one of Madigan & York are linked as follows:

$$\theta_{1++} = q_1,\ \theta_{11+} = q_1\mu_{2|1},\ \theta_{01+} = (1-q_1)\lambda_{2|1},\ \theta_{1+1} = q_1\mu_{3|1},\ \theta_{0+1} = (1-q_1)\lambda_{3|1}.$$

Considering the above equalities, it is immediate to check that there is a one-to-one and onto correspondence between both parametrisations. The density of $\theta_m$ is derived from the one of $(\theta_{1++}, \theta_{11+}, \theta_{01+}, \theta_{1+1}, \theta_{0+1})$. First, is easy to check that the Jacobian of the above transformation is equal to: $q_1^2(1-q_1)^2$. Using now (1), it comes that $\pi(\theta_m)$ is proportionnal to $T_1\, T_2\, T3$, where and

$$T_1 = q_1^{a_{1++}-1}(1-q_1)^{a_{0++}-1} \quad T_2 = \mu_{2|1}^{a_{11+}-1}(1-\mu_{2|1})^{a_{10+}-1}\lambda_{2|1}^{a_{01+}-1}(1-\lambda_{2|1})^{a_{00+}-1}$$

and

$$T_3 = \mu_{3|1}^{a_{1+1}-1}(1-\mu_{3|1})^{a_{1+0}-1}\lambda_{3|1}^{a_{0+1}-1}(1-\lambda_{3|1})^{a_{0+0}-1},$$

from which we immediately deduce that $q_1$, $\mu_{2|1}$, $\lambda_{2|1}$, $\mu_{3|1}$ and $\lambda_{3|1}$ follow independently beta distributions. Moreover, it is straightforward to check (using Proposition 1) that they are all compatible.

• It is now assumed that $q_1$, $\mu_{2|1}$, $\lambda_{2|1}$, $\mu_{3|1}$, $\lambda_{3|1}$ follow independently compatible beta distributions.

First, compatibility and Proposition 1 imply that $q_1 \sim \text{beta}(a_{1++}, a_{0++})$, $\mu_{2|1} \sim \text{beta}(a_{11+}, a_{10+})$, $\lambda_{2|1} \sim \text{Beta}(a_{01+}, a_{00+})$, $\mu_{3|1} \sim \text{beta}(a_{1+1}, a_{1+0})$, $\lambda_{2|1} \sim \text{beta}(a_{0+1}, a_{0+0})$. Re-finding the density of $(\theta_{C_1}, \theta_{C_2})$ (as it appears in the first part of the proof), from the one of $q_1$, $\mu_{2|1}$, $\lambda_{2|1}$, $\mu_{3|1}$, and $\lambda_{3|1}$, proceeds as in the first part of the proof; therefore, details are omitted. It is clear that the hyper-Dirichlet distribution of which the density is given by (2) is effectively compatible, since this hyper-Dirichlet has (by construction) its margins given by (1).

**II.** The model $[1, 23]$

In the graph associated with the model $[1, 23]$, there are two cliques $\{C_1, C_2\}$, namely $C_1 = \{1\}$ and $C_3 = \{2, 3\}$, and no separator.

• We first prove the direct sens of Proposition 3. Because the priors distributions placed on the cliques probabilities are assumed to be compatible, one has:

$$\theta_{C_1} = (\theta_{1++}, \theta_{0++}) \sim \text{Dirichlet}(a_{1++}, a_{0++})$$

and

$$\theta_{C_2} = (\theta_{+00}, \theta_{+01}, \theta_{+10}, \theta_{+11}) \sim \text{Dirichlet}(a_{+00}, a_{+01}, a_{+10}, a_{+11}).$$

Moreover, the Madigan & York (1997) approach requires that the prior on $(\theta_{C_1}, \theta_{C_2})$ is such that $\theta_{C_1} \perp \theta_{C_2}$ (considering the graph of $[1, 23]$). We have to prove that $q_1$, $q_2$, $\mu_{3|2}$ and $\lambda_{3|2}$ follows independently compatible beta distributions. If the other parametrisation is of concern, one has to prove the same property, but with $q_1$, $q_3$, $\mu_{2|3}$, $\lambda_{2|3}$. The fact that $q_1$ follows a

compatible beta distribution is immediate. From now, we work with the first parametrisation, but similar developments hold with the other. The fact that $q_2$, $\mu_{3|2}$ and $\lambda_{3|2}$ follow also compatible beta distributions is a consequence of Proposition 2 applied to $\theta_{C_2}$. The fact that $q_2$, $\mu_{3|2}$ and $\lambda_{3|2}$ are independent is an immediate consequence of the characterisation of the Dirichlet distribution via global parameter independence (namely $q_2 \perp (\mu_{3|2}, \lambda_{3|2})$ in our case), and local parameter independence (namely $\mu_{3|2} \perp \lambda_{3|2}$ in our case); see Geiger and Heckerman (1997). We are indeed exactly in the framework considered by these authors; namely a two-way contingency table with $J \times K$ cells (in our case $J = K = 2$). Remain to show that $q_1 \perp (q_2, \mu_{3|2}, \lambda_{3|2})$. Now, on the one hand, one has $\theta_{+11} \perp (\theta_{+01}, \theta_{+10}, \theta_{+11})$ (which is deduced from $\theta_{C_1} \perp \theta_{C_2}$); and, on the other hand, there exits a one-to-one and onto transformation between $(\theta_{+01}, \theta_{+10}, \theta_{+11})$ and $(q_2, \mu_{3|2}, \lambda_{3|2})$ since $q_2 = \theta_{+10} + \theta_{+11}$, $\mu_{3|2} = \theta_{+11}/\theta_{+1+}$, and $\lambda_{3|2} = \theta_{+01}/(1 - \theta_{+1+})$. From both hands, we deduce that $q_1 \perp (q_2, \lambda_{3|2}, \mu_{3|2})$.

• To prove the converse, we use again the characterisation of Geiger and Heckerman (1997). Before, let us clarify one point. When model $[1, 23]$ is under consideration, two parametrisations are possible. Because the model averaging procedure has to work with both, it is logic to require that the two coresponding parameters sets satisfy the assumptions of prior independence. This way of seeing allows us to use the the characterisation of Geiger and Heckerman to establish the converse. This characterisation says that if: 1) $q_2$, $\mu_{3|2}$, $\lambda_{3|2}$ are mutually independent; 2)$q_3$, $\mu_{2|3}$, $\lambda_{2|3}$ are mutually independent; 3) each parameter has a strictly positive pdf; thus $\theta_{C_2}$ follows a Dirichlet distribution.

The third point is fulfilled since a beta distribution is assumed for all the parameters. We now have to prove that this Dirichlet is effectively compatible; in other terms, one has to prove that if $\theta_{C_2} \sim$ Dirichlet $(\alpha_{00}, \alpha_{01,10}, \alpha_{11})$, thus $\alpha_{00} = a_{+00}$,$\alpha_{01} = a_{+00}$, $_{+10} = a_{+10}$, and $\alpha_{11} = a_{+11}$. Applying Proposition 2 to the above distribution of $\theta_{C_2}$, one derives that $\mu_{3|2} \sim \text{beta}(\alpha_{11}, \alpha_{10})$ and $\mu_{3|2} \sim \text{beta}(\alpha_{11}, \alpha_{10})$. Now, $\mu_{3|2}$ and $\lambda_{3|2}$ are assumed to follow Beta compatible distributions, thus $\mu_{3|2} \sim \text{beta}(\text{a}_{+11}, a_{+10})$ and $\mu_{3|2} \sim \text{beta}(\text{a}_{+11}, a_{+10})$, hence the conclusion. Remain to prove that $\theta_{C_1} \perp \theta_{C_2}$; it is done as in the direct sens of the prove, except that parts between $(\theta_{+01}, \theta_{+10}, \theta_{+11})$ and $(q_2, \mu_{3|2}, \lambda_{3|2})$ are reversed.

## Appendix C

Recall that $q_1 \sim \text{beta}(a_{1++}, a - a_{1++})$, $q_2 \sim \text{beta}(a_{+1+}, a - a_{+1+})$, and that $q_3 \sim \text{beta}(a_{++1}, a - a_{++1})$. Consequently, the fact of having the distribution of each $q_j$ is equivalent to have $a_{1++}$, $a_{+1+}$, $a_{++1}$ and $a$ which are, for convenience, respectively denoted by: $c_1$,$c_2$,$c_3$. Assume now that, for each fixed pair $\{r, s\}$ where $r$ and $s$ are distinct and belong to $\{1, 2, 3\}$, one has the prior mean of any one parameter among $\{\mu_{r|s}, \mu_{s|r}, \lambda_{r|s}, \lambda_{r|s}\}$. For example, assume that $\text{E}(\mu_{2|1})$, $\text{E}(\mu_{3|2})$, and $\text{E}(\mu_{1|3})$ are available; for convenience, they are respectively denoted by $c_4$,$c_5$,$c_6$. All this prior leads to seven independent equations, namely:

$a_3 + a_5 + a_6 + a_7 = ac_1$;$a_2 + a_3 + a_6 + a_7 = ac_2$;$a_1 + a_3 + a_5 + a_7 = ac_3$

$a_6 + a_7 = ac_1c_4$;$a_3 + a_7 = ac_2c_5$;$a_5 + a_7 = ac_3c_6$

$a_0 + a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 = a$

For example, equation (1) is deduced from $\text{E}(q_1) = a_{1++}/a = c_1$, hence $a_{1++} = ac_1$; equation (4) is deduced from $\text{E}(\mu_{2|1}) = a_{11+}/a_{1++} = c_4$, hence

$a_{11+} = ac_1c_4$. We need an additional equation which is independent of the seven above equations. Having $E(1-\theta_{000})$, afterwards denoted by $c_0$, is a possibility; it leads to the additional equation: $(a-a_0)/a = c_0$ that is: $a - a_0 = ac_0$. Finally we obtain the following system: $MA = C$ where $A^t = (a_0, a_1, \ldots, a_7)$, $C^t = (c_0, c_1, \ldots, a_7)$ and

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

It is easy to show (though it is particularly tedious) that $\det(M)$=1. The system has thus a unique solution, namely $M^{-1}C$.

## Appendix D

In this Appendix we provide the expressions of the likelihood and of $I_m(N)$ for each model $m$.

• Under the saturated model $m = [123]$, the likelihood is the one given by (5.2) and one has:

$$I_m = \frac{N!}{(N-d)!\prod_{h=1}^{7} n_h!} \; \frac{\Gamma(N-d+a_{000})}{\Gamma(N+a)} \; \prod_{h=1}^{7} \Gamma(n_h + a_h)$$

• Under the model $m = [12, 13]$, the likelihood is proportional to:

$$L(\theta_m, N; \mathbf{y}) \propto \frac{N!}{(N-d)!} q_1^{d_1} (1-q_1)^{N-d_1} \, E_{2|1} \, E_{3|1}$$

where

$$E_{2|1} = \lambda_{2|1}^{n_{01+}} (1-\lambda_{2|1})^{N-d_1-n_{01+}} \mu^{n_{11+}} (1-\mu_{2|1})^{n_{10+}}$$

and

$$E_{3|1} = \lambda_{3|1}^{n_{0+1}} (1-\lambda_{3|1})^{N-d_1-n_{0+1}} \mu_{3|1}^{n_{1+1}} (1-\mu_{3|1})^{n_{1+0}}$$

Morever, one has:

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h=1}^{7} n_h!} B_1 B_{2|1} B_{3|1}$$

where $B_1 = \mathrm{B}(d_1 + a_{1++}, N - d_1 + a_{0++})$ and,

$$B_{2|1} = \mathrm{B}(n_{01+} + a_{01+}, N - d + n_{001} + a_{00+}) \; \mathrm{B}(n_{11+} + a_{11+}, n_{10+} + a_{10+})$$

and

$$B_{3|1} = \mathrm{B}(n_{0+1} + a_{0+1}, N - d + n_{010} + a_{0+0}) \; \mathrm{B}(n_{1+1} + a_{1+1}, n_{1+0} + a_{1+0})$$

- Under the model $m = [1, 23]$, one has:

$$L(\theta_m, N; \mathbf{y}) \propto \frac{N!}{(N-d)!} q_1^{d_1} (1-q_1)^{N-d_1} q_2^{d_2} (1-q_2)^{N-d_2} E_{3|2},$$

where

$$E_{3|2} = \lambda_{3|2}^{n_{+01}} (1-\lambda_{3|2})^{N-d-n_{100}} \mu_{3|2}^{n_{+11}} (1-\mu)_{3|2}^{n_{+10}},$$

and when one adopts the factorization: $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_2)$. Morever, one has:

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h=1}^{7} n_h!} B_1 B_2 B_{3|2}$$

where $B_1$ is defined above, and

$$B_2 = \mathrm{B}(d_2 + a_{+1+}, N - d_2 + a_{+0+}) \quad B_{3|2} = \mathrm{B}(n_{+01} + a_{+01}, N - d + n_{100} + a_{+00})$$

and $\mathrm{B}(n_{+11} + a_{+11}, n_{+10} + a_{+10})$.

- Under the independent model $m = [1, 2, 3]$, one has:

$$L(\theta_m, N; \mathbf{y}) \propto \frac{N!}{(N-d)!} \prod_{j=1}^{3} q_j^{d_j} (1 - q_j)^{N - d_j}$$

and

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h=1}^{7} n_h!} B_1 B_2 B_3$$

where $B_1$, $B_2$ are defined above and $B_3 = \mathrm{B}(d_3 + a_{++1}, N - d_3 + a_{++0})$.

## Appendix E

Recall that the posterior distribution of $N$ (under model $m$) will be defined if and only if the series of general term $p(\mathbf{y}|N, m)\pi(N)$ is convergent where the expression of $p(\mathbf{y}|N, m) = I_m(N)$ is given in Appendix E. The results concerning the existence of $N|\mathbf{y}, m$ under models $m = [12, 13]$ and $m = [123]$ (proved below) use the following result:

$$\frac{\Gamma(N + v)}{\Gamma(N + u)} \sim \exp(v - u)\ N^{v-u}.$$

where $u$ and $v$ denote reals which do not depend on $N$. To obtain this equivalent, we start from the well known equivalent:

$$\Gamma(N) \sim \sqrt{2\pi} N^{N - \frac{1}{2}} \exp(-N).$$

from which we deduce that:

$$\frac{\Gamma(N + v)}{\Gamma(N + u)} \sim \frac{(N + v)^{N + v - 1/2}}{(N + u)^{N + u - 1/2}} \exp^{-(v-u)}.$$

Now, it is easy to check that:

$$\frac{(N+v)^{N+v-1/2}}{(N+u)^{N+u-1/2}} = \Big[1-\frac{u-v}{N+u}\Big]^{N+u-1/2}(N+v)^{v-u}.$$

Since, one has:

$$\Big[1-\frac{u-v}{N+u}\Big]^{N+u-1/2} = \exp\Big[(N+u-1/2)\log\Big(1-\frac{u-v}{N+u}\Big)\Big]$$

it comes:

$$\Big[1-\frac{u-v}{N+u}\Big]^{N+u-1/2} \sim \exp(v-u).$$

• We first consider the model $m = [12, 13]$. We have to examine the series of general term $I_m\ \pi(N)$ where $I_m = \frac{N!}{(N-d)!}\prod_{h=1}^{7} n_h! B_1 B_{2|1} B_{3|1}$ and $\pi(N) = 1/N^t$. It is straightforward to see that one has actually to examine the convergence of the series of general term $w_N = N^{d-t} T1T2T3$ where

$$T1 = \frac{\Gamma(N-d_1+a_{0++})}{\Gamma(N+a)} \quad \text{and} \quad T2 = \frac{\Gamma(N-d+n_{001}+a_{00+})}{\Gamma(N-d+n_{001}+n_{01+}+a_{0++})}$$

and $T3 = \Gamma(N-d+n_{001}+a_{0+1})/\Gamma(N-d+n_{010}+n_{0+1}+a_{0++})$.

Using now the above equivalent of $\Gamma(N+v)/\Gamma(N+u)$, one finds that:

$$w_N \sim c_1\ N^{d-t} N^{-d_1+a_{0++}-a} N^{-n_{01+}-a_{01+}} N^{-n_{0+1}-a_{0+1}}.$$

where $c$ denotes a constant. By observing that $d-d_1-n_{01+}-n_{0+1} = -n_{011}$ and that $-a+a_{0++}-a_{01+}-a_{0+1} = -a+a_{000}-a_{011}$ it comes that:

$$w_N \sim c\, N^{-(t+n_{011}+a-a_{000}+a_{011})}.$$

The posterior distribution of $N$ thus exits if and only if: $t+n_{011}+a-a_{000}+a_{011} > 1$.

• We now consider the saturated model [123]. One has actually to examine the convergence of the series of general term:

$$w_N = \frac{N!}{(N-d)} \frac{\Gamma(N-d+a_{000})}{\Gamma(N+a)} \frac{1}{N^t}.$$

Considering that

$$\frac{\Gamma(N-d+a_{000})}{\Gamma(N+a)} \sim \exp(a-a_{000}-d)\, N^{-(d+a-a_{000})},$$

we deduce that

$$w_N \sim \exp(a-a_{000}-d) N^{-(t+a-a_{000})}.$$

The posterior distribution of $N$ thus exits if and only if: $t+a-a_{000} > 1$.

## Appendix F

First the posterior distribution of $N$ exists if and only if the posterior distribution of $N$ exists under each model $m$, as it clearly appears in (5.1). Now, the condition which ensures the existence of the posterior distribution of $N$ under the saturated model is the strongest. Indeed, it is clear that, on one hand, $(d_1+d_2+d_3)-d \geq 0$, $d_1 - n_{100} \geq 0$ (idem for $d_2 - n_{010}$, and for $d_3 - n_{001}$) and that, on the other hand, $a_{1++}+a_{+1+}+a_{++1}$, $(a+a_{1++}-a_{100})$, and $(a-a_{000})+a_{011}$ are all strictly greater than $a-a_{000}$.