HAL open science

Estimating the size of small populations from incomplete lists via graphical models

Jérôme Dupuis

To cite this version:

Jérôme Dupuis. Estimating the size of small populations from incomplete lists via graphical models. 2020. hal-02925222

HAL Id: hal-02925222

https://hal.science/hal-02925222

Preprint submitted on 28 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimating the size of small populations from incomplete lists via graphical models

Jérôme A. Dupuis

IMT, Université Paul Sabatier, Toulouse, France

Abstract

We consider the problem of estimating the size N of a closed population from q incomplete lists. Estimation of N is based on capture-recapture type models. We use graphical models to deal with possible dependencies between lists. The current parametrization involves clique probabilities which have no simple concrete meaning and are delicate to manipulate in a Bayesian context insofar as hyper-Dirichlet distributions are used as priors. Our parametrization involves marginal and conditional capture probabilities. We develop our approach with $q = 3$. We show that there is a one-to-one and onto correspondence between both parametrizations and that placing hyper-Dirichlet distributions on the clique parameter boils down to place independent beta distributions on the capture parameters. When N is small, the non informative Bayesian analysis encounters difficulties. The posterior distribution of N may not exist for a particular graphical model: we give a necessary and sufficient condition of existence for each. Moreover, it is highly desirable that the priors on capture are compatible across the different models. Now, due to the small size of N, fulfilling this requirement demands a particular attention. We conclude by extending our approach to $q = 4$ lists.

Key Words. Bayesian model averaging; Capture-recapture; Graphical models; Hyper-Dirichlet distribution; Incomplete lists; Population size estimation; Small populations.

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 $q \geq 2$ 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; Chao *et al.*, 2001). 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), what underlines the fact that the value $q = 3$ is an important particular case on which we will focus in this paper.

Two approaches have been proposed for modelling some possible dependences between the lists: the one of King & Brooks (2001) which uses log-linear models, and the one of Madigan & York (1997) which uses decomposable graphical models. The paper takes place in the latter. Graphical models are a particularly attractive tool to formulate in a rigorous way all the conditional (or marginal) independence assumptions between the different random variables. Moreover, this tool allows to visualize such assumptions what makes it very popular among the researchers; eg Dupuis (1995), Hojsgaard, Edwards, and Lauritzen (2012). The graph of a graphical model m is deter-

mined by its (maximal) cliques, and Madigan & York (1997) parametrize any model by the corresponding cliques probabilities. This way of proceeding is quite natural. However, most of researchers are used to work with capture parameters which are quantities having a very concrete meaning (contrary to clique probabilities). Hence the interest to examine in which extend it is possible to reparametrize the approach of Madigan & York (1997) via capture probabilities. We show that it is effectively the case and that both parametrizations produce the same bayesian inference on N. Accordingly, the theoretical results established in this paper will apply without any modification to the Madigan & York's approach. As in Madigan and York (1997), inference on N is based on a Bayesian model averaging which includes all the possible decomposable graphical models (for fixed q).

The paper focuses on populations of which the size N is small; see for example Wang *et al.* (2007) for motivations. In human populations, estimating the number of people affected by a rare disease typically enters in this framework. When N is small, it may occur that one (or more) count associated with a particular capture-recapture history (different from the one of an individual never captured) is null or very low. In such circumstancies, statistical difficulties may occur when one wishes to perform a non informative Bayesian analysis of the data.

- A first difficulty, not mentioned in Madigan and York (1997), is related to the existence of the posterior distribution of N when an improper prior is placed on N (Jeffreys or uniform). Sufficient and necessary conditions of existence are thus stated for each graphical model: our results extend the result of Wang *et al.* (2007) which concerned only the simplest model

(namely the independant model). A analoguous result of existence is also stated for the Bayesian averaging model procedure.

- A second difficulty is related to the priors put on the capture parameters. When inference involves several candidate models, it is strongly desirable that the priors are compatible across the different models (eg Dawid and Lauritzen, 1993). Here, the prior distribution on any capture parameter of any sub-model is derived from the prior distribution put on the parameter of the saturated model (afterwards denoted by $\boldsymbol{\theta}_{\text{Sat}}$). This strategy is the one adopted by Madigan and York (1997) to derive the prior distributions on the clique probabilities. Standard non informative priors on $\boldsymbol{\theta}_{\text{Sat}}$ are the uniform distribution on the 2^q-simplex and the Jeffreys prior. Now, we note that Madigan and York (1997) advice to use the uniform prior. But this prior induces informative priors on the marginal capture probabilities since they follow a beta $(2^{q-1}, 2^{q-1})$ which cannot be considered as non informative when $q \geq 3$ (especially when N is small).

- In fact these two difficulties are closely linked, since the prior adopted for $\boldsymbol{\theta}_{\text{Sat}}$ plays a part in the condition of existence of the posterior distribution of N (see Section 7). In this paper, we propose a distribution on $\boldsymbol{\theta}_{\text{Sat}}$ which induces non informative priors on marginal and conditional capture probabilities, and ensures - whatever the graphical model considered - the existence of the posterior mean of N when the Jeffreys prior is adopted for N.

In conclusion, we briefly indicate how to extend our approach to the particular case $q = 4$, and how to take into account a possible individual heterogeneity at the capture level (see Section 6).

2 Data description

For any individual i of the population of interest, we denote by $\mathbf{x}_i$ the vector $(x_{ir}; r = 1, 2, 3)$ where $x_{ir} = 1$ if individual i appears on list r 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). The set of all these histories is denoted by $\mathcal{H}$, and the set $\mathcal{H}$ minus the history 000 is denoted by $\mathcal{H}^*$. We denote by n_h the number of individuals whose history is h. Note that the count n_{000} is not observable and that $d = \sum_{h \in \mathcal{H}^*} n_h$ represents the number of individuals appearing in at least one list. Data is denoted by $\mathbf{y}$; thus, one has $\mathbf{y} = \{n_h; h \in \mathcal{H}^*\}$.

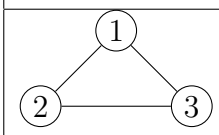
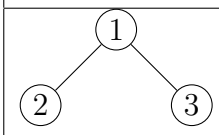
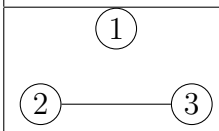
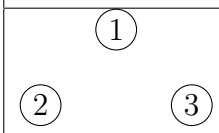
3 Assumptions, models and parameters

We assume that the N random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_N$ are independent and identically distributed. Therefore, the probabilistic assumptions concerning the components of the random vector $\mathbf{X}_i = (X_{ir}; r = 1, 2, 3)$ do not depend on i; and, for convenience, we will afterwards omit index i in X_{ir}. The assumptions on X_1, X_2, and X_3 are formulated via graphical models. As Madigan and York (1997), we consider eight models: the saturated model and seven sub-models obtained by removing one or several arrows in the graph of the saturated model.

- 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.

- A model which assumes a conditional independence assumption between two nodes of the graph is said of type I. The model which assumes that

Table 1: Characteristics of each type of graphical model ($q = 3$ lists)

model	name	assumption	factorization
	satured	no	$p(x_1, x_2, x_3)$
	type I	$X_2 \perp X_3 \mid X_1$	$p(x_1)p(x_2\mid x_1)p(x_3\mid x_1)$
	type II	$X_1 \perp (X_2, X_3)$	$p(x_1)p(x_2, x_3)$
	independent	$\perp (X_1, X_2, X_3)$	$p(x_1)p(x_2)p(x_3)$

$X_2 \perp X_3 | X_1$ is denoted by $[12, 13]$. The two other models of type I are denoted by $[23, 21]$ and $[31, 32]$ (with obvious notation).

- A model which assumes a marginal independence assumption between one node and the two others is said of type II. The model which assumes that $X_1 \perp (X_2, X_3)$ is denoted by $[1, 23]$; The two other models of type II are denoted by $[2, 13]$ and $[3, 21]$ (with obvious notation).

- The independant model denoted by $[1, 2, 3]$. It assumes that the three random variables X_1, X_2, and X_3 are independant.

The graphs of the above models appears in Table 1. In column 1, X_1, X_2, X_3 are, for convenience, respectively represented by 1, 2, 3. Note that all these graphs are decomposable. Let $p(x_1, x_2, x_3)$ be denote the probability that $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$ where x_1, x_2, x_3 belongs to $\{0, 1\}$. For each sub-model we can deduce from its graph a specific factorization of $p(x_1, x_2, x_3)$

given in Table 1. Each factorization induces a natural parametrization of the sub-model in terms of marginal and conditional capture-recapture probabilities. We thus introduce the following notation. Marginal capture probabilities are denoted by θ_r which represents the probability that an individual appears on list $r \in \{1, 2, 3\}$. Conditional capture probabilities are denoted as follows: $\theta_{s|r}$ represents the probability that an individual appears on list s given that he appears on list $r \neq s$, and $\theta_{s|\bar{r}}$ represents the probability that he appears on list s given that he 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 θ_1, θ_2, θ_3. A Model of type I includes five parameters. For example, the parameters of model $[12, 13]$ are θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$ and $\theta_{3|\bar{1}}$. As far as models of type II are concerned, two parametrizations are possible (both are natural). 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 θ_1, θ_2, $\theta_{3|2}$, $\theta_{3|\bar{2}}$, and parameters θ_1, θ_3, $\theta_{2|3}$ and $\theta_{2|\bar{3}}$ in the second case.

The above notation concern the parameters of sub-models. The saturated model is parametrized by the θ_h's where θ_h denotes the probability that an individual has h as history; thus, one has: $\boldsymbol{\theta}_{\text{Sat}} = (\theta_{111}, \ldots, \theta_{000})$. We stress that this additional notation indexes θ by a capture-recapture history, contrary to the one adopted in sub-models which indexes θ by the lists.

4 Priors

4.1 Priors on the size N of the population.

We will use the greek letter π to designate any prior or posterior density, continuous as well as discrete. Moreover, for a graphical model m, the set of all the capture probabilities (marginal and conditional) is denoted by $\boldsymbol{\theta}_m$.

First, we assume that, for all models m, N and θ_m are a priori independent, thus one has: $\pi(N, \boldsymbol{\theta}_m) = \pi(N)\pi(\boldsymbol{\theta}_m)$. 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; Dupuis and Schwarz, 2007). Note that both are improper and the existence of the posterior distribution of N is thus not guaranteed (hence the study made in Section 6).

4.2 Prior distributions on the capture probabilities

For each sub-model m, we assume that the elementary parameters present in sub-model m are a priori independent; we mean by elementary parameter any capture probability (marginal as well as conditional) present in model m. For example, if $m = [12, 13]$, we assume that θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$ and $\theta_{3|\bar{1}}$ are a priori independant. As in the paper of Madigan & York (1997), we adopt for $\boldsymbol{\theta}_{\text{Sat}}$ a Dirichlet distribution with parameters $(a_{000}, \ldots, a_{111})$, afterwards denoted by $\mathscr{D}(a_{000}, \ldots, a_{111})$; thus, one has:

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

where the a_h are all stricly positive. In a non informative set-up, a standard choice is the uniform distribution. The main alternatives to the uniform prior

are the Jeffreys prior and the Perks prior which respectively corresponds to a $\mathscr{D}(1/2, \ldots, 1/2)$ and to a $\mathscr{D}(1/2^3, \ldots, 1/2^3)$.

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

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

Recall that θ_{jkl} denotes the probability that an individual has jkl as history. Notations θ_{j+l}, θ_{+kl}, θ_{+k+} and θ_{++l} are defined similarly.

As far as a Bayesian 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. In particular, it is natural to require that the prior distribution put on any fixed elementary parameter is the same from model to model. A simple way to fulfill this requirement is to derive its density from the one put on $\boldsymbol{\theta}_{\text{Sat}}$, considering that all elementary parameters express in function of the θ_h's; for example, one has:

$$\theta_3 = \sum_{j,k \in \{0,1\}} \theta_{jk1}, \quad \theta_{2|1} = \frac{\theta_{110} + \theta_{111}}{\theta_{1++}} \quad \text{and} \quad \theta_{2|\bar{1}} = \frac{\theta_{010} + \theta_{011}}{\theta_{0++}}.$$

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

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

Proof. The first part of the Proposition is an immediate consequence of the *agregation property* of the Dirichlet distribution. See Appendix A for the second part.

Similar results of course hold for θ_{+k+}, θ_{++l}, and for all the possible ratios similar as the one appearing in Proposition 1. Fixing the hyperparameters a_h

thus induces a prior distribution for all the marginal and conditional capture probabilities. If we put a uniform distribution on $\boldsymbol{\theta}_{\text{Sat}}$, the conditional capture probabilities follow a Beta $(2, 2)$ and the marginal probabilities follow a Beta $(4, 4)$. These beta distributions cannot be considered as non informative. As far as marginal probabilities are concerned, the Jeffreys prior suffers from the same drawback than the uniform prior (but to a lesser extent). On the contrary, if we put a $\mathscr{D}(1/4, \ldots, 1/4)$ on $\boldsymbol{\theta}_{\text{Sat}}$, all the marginal and conditional capture probabilities follow non informative distributions, namely a uniform distribution for the former, and a Jeffreys distribution for the latter. The Perks prior also induces non informative priors on capture probabilities; but the Dirichlet $\mathscr{D}(1/4, \ldots, 1/4)$ will turn out to be preferable to ensure the existence of the posteriors (see Section 7).

5 The links with the Madigan & York's paper

Proposition 2 below clarifies the links between the parametrisation of Madigan and York (1997) which uses the notion of clique probability, and ours which involves marginal and conditional capture probabilities.

Proposition 2. For each fixed sub-model m, there is a one-to-one and onto correspondence between the clique probabilities parametrisation and the capture probabilities parametrisation.

Proof. The proof appears in Appendix B.

Proposition 3 below clarifies the links exiting between the prior adopted by Madigan and York (1997) for the clique parameter of model m and the prior we put on the capture parameter present in model m.

Proposition 3. Sub-model m being fixed, if one adopts the hyper-Dirichlet

distribution of Madigan and York 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. The proof appears in Appendix C.

In this Appendix we recall, for each model m, the density of the hyper-Dirichlet distribution adopted by Madigan and York (1997). In Proposition 3, we mean by *clique parameter* of model m the set of the clique probabilities present in model m (a similar definition is adopted for the term *capture parameter*). Moreover, we mean by *compatible* beta distributions that the prior distributions placed on the marginal and conditional capture probabilities are derived from the Dirichlet distribution placed on $\boldsymbol{\theta}_{\text{Sat}}$, as Madigan & York (1997) did for the clique probabilities.

6 Conditions of existence of posteriors.

The graphical model being fixed, we provide a necessary and sufficient condition of existence of different posteriors: namely, the posterior distribution of N, as well as the posterior mean and variance of N. We also give a necessary and sufficient condition of existence of the averaged-model posterior mean of N, that is of $E[N|\mathbf{y}]$. In this paper, we focus on the posterior mean which is the quantity the most often retained for estimating N; see eg George & Robert (1992), King & Brooks (2001); as well as Wang *et al.* (2007) for small populations. Note that Madigan & York (1997) use the absolute quadratic loss which yields an estimate of N different from the posterior mean.

6.1 Conditions of existence of $E[N|\mathbf{y}, m]$.

Model m and data $\mathbf{y}$ being given, we have:

$$\pi(N|\mathbf{y}, m) = \frac{p(\mathbf{y}|N, m)\pi(N)}{\sum_{N \geq d} \; p(\mathbf{y}|N, m)\pi(N)} \tag{5.1}$$

with

$$p(\mathbf{y}|N, m) = \Pr(\mathbf{Y} = \mathbf{y}|N, m) = \int_{\Theta_m} L(\boldsymbol{\theta}_m, N; \mathbf{y})\pi(\boldsymbol{\theta}_m)\, \mathrm{d}\,(\boldsymbol{\theta}_m), \tag{5.2}$$

where $L(\boldsymbol{\theta}_m, N; \mathbf{y})$ denotes the likelihood of $(\boldsymbol{\theta}_m, N)$ under model m. Note that the distribution of $N|\mathbf{y}, m$ will exist if and only if the integral appearing in (5.2) is finite (for all $N \geq d$) and the series of general term $p(\mathbf{y}|N, m)\pi(N)$ converges. For obtaining $L(\boldsymbol{\theta}_m, N; \mathbf{y})$ we have to compute $\Pr(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta}_m, N)$. The assumption of independence between the X_i's (see Section 3) implies that:

$$(N_{001}, \ldots, N_{111})|N, \boldsymbol{\theta}_m \sim \text{Multinomial}(N; \theta_{001}, \ldots, \theta_{111}),$$

from which we deduce that:

$$L(\boldsymbol{\theta}_m, N; \mathbf{y}) = \frac{N!}{(N-d)! \; \prod_{h \in \mathcal{H}^*} n_h!} \left[1 - \sum_{h \in \mathcal{H}^*} \theta_h\right]^{N-d} \prod_{h \in \mathcal{H}^*} \theta_h^{n_h}. \tag{5.3}$$

The expression of $L(\boldsymbol{\theta}_m, N; \mathbf{y})$ in function of the capture parameters present in model m, is now easily obtained by taking into account the factorization given in Table 1 (see Appendix D).

For obtaining $p(\mathbf{y}|N, m)$ one has to integrate $L(\boldsymbol{\theta}_m, N; \mathbf{y})$ over $\boldsymbol{\theta}_m$. This integral, afterwards denoted by $I_m(N)$, can be write down in a closed form (for each model m); see Appendix D. As stressed in this Appendix, $I_m(N)$ exists for all model m, for all $N \geq d$ and for all data set $\mathbf{y}$, because the a_h's are

strictly positive. $E[N|\mathbf{y}, m]$ will thus exist if and only if the series of general term $NI_m(N)\pi(N)$ converges. In Proposition 4 below, we provide a necessary and sufficient condition so that this series converges; in our statement, $\lambda = 1$ is associated with the Jeffreys prior $\pi(N) = 1/N$, and $\lambda = 0$ with the uniform prior $\pi(N) = 1$.

Proposition 4. The posterior mean of N exits:

- under the saturated model, if and only if, $\lambda + a - a_{000} > 2$,
- under $[12, 13]$, if and only if, $\lambda + n_{011} + (a - a_{000}) + a_{011} > 2$,
- under $[1, 23]$, if and only if, $\lambda + d_1 - n_{100} + (a + a_{1++} - a_{+00}) > 2$,
- under $[1, 2, 3]$, if and only if, $\lambda + (d_1 + d_2 + d_3) - d + a_{1++} + a_{+1+} + a_{++1} > 2$.

Proof. It appears in Appendix E.

- If the existence of the posterior distribution of N is of interest, replace 2 by 1 in the right member of inequalities. For the posterior variance of N, replace 2 by 3.

- 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}.

We now briefly comment Proposition 4.

- Except for the saturated model, we observe that the left term which appears in the inequalities decomposes in three parts: the first one (namely λ) is related to the prior put on N, the second one is related to the data, and the last one is related to the prior put on $\boldsymbol{\theta}_{\text{Sat}}$. In small populations, it may occur

that, for a given sub-model, the second term may be null and thus the choice of the prior on $\boldsymbol{\theta}_{\text{Sat}}$ becomes crucial to meet the corresponding condition of existence (it is the case for the two data sets considered in Section 7). For example, if one adopts the Perks prior for $\boldsymbol{\theta}_{\text{Sat}}$ and the Jeffreys prior for N, the posterior mean of N does not exist under the model [12, 13] when n_{011} is null (in this configuration, only the posterior distribution of N exists). If, always in this configuration, we replace the Perks prior by a $\mathscr{D}(1/4, \ldots, 1/4)$ thus the posterior mean of N will exist (recall that this Dirichlet ensures that the priors on the capture proabilites are all non informative, see section 4.2).

- For the saturated model, comments are reported in Section 6.2.

6.2 Condition of existence of $E[N|\mathbf{y}]$.

In a Bayesian model averaging procedure, N is typically estimated by the averaged-model posterior mean of N, that is by $E(N|\mathbf{y})$; see eg Hoeting *et al.* (1999). Now, one has:

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

where $p(m|\mathbf{y})$ represents the posterior probability of model m.

Proposition 5. $E(N|\mathbf{y})$ exists if and only if $\lambda + a - a_{000} > 2$ where $\lambda = 0$ when $\pi(N) = 1$ and $\lambda = 1$ when $\pi(N) = 1/N$.

Proof. See Appendix F.

We note that data play no part in the condition given by Proposition 4; the reason is that the condition $\lambda + a - a_{000} > 2$, which appears in Proposition 2, is the stongest one (see Appendix F). Consequently, the result of Proposition 5 is very general and applies to any type of data (relatively to

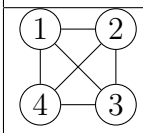
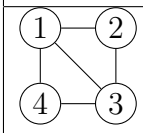
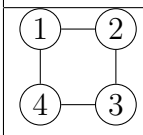
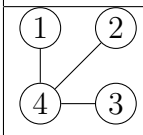
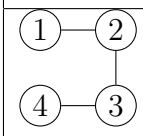
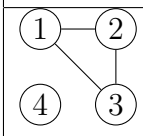
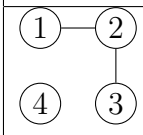
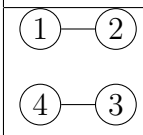
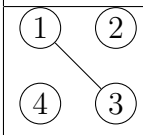
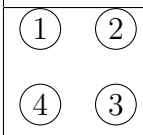
the sizes of the n_h's). That means that basing inference on a Bayesian model averaging procedure is possible only if appropriate priors are choose for N and $\boldsymbol{\theta}_{\text{Sat}}$. For example, if one adopts the Jeffreys prior for N, the Perks's prior cannot be used; on the opposite, the averaged-model posterior mean of N will exist in all the other cases (that is with the uniform and Jeffreys priors, as well as a $\mathscr{D}(1/4, \ldots, 1/4)$).

7 Generalizations and extensions.

• The parametrization in terms of marginal and conditional capture probabilities can be extended to $q \geq 4$ lists. In practice, the number of lists rarely exceeds 4, and we will thus limit ourself (for brievety) to the case $q = 4$. Table 6 provides, for each type of graphical model, the corresponding conditional and marginal independence assumptions. All these models are decomposable, except the third (from top). For each model (except the third) we provide the factorization of $p(x_1, x_2, x_3, x_4)$ derived from its graph. The third model has no factorization and will be typically removed from a Bayesian averaging model. As for the case $q = 3$, the definition of the parameters is, for each decomposable sub-model, derived from its factorization (details are omitted). Note that the second and sixth models from top, involve conditional capture probabilities where the conditioning is on two lists.

Concerning the averaged-model posterior mean of N, it is easy to check that it exists if and only if $\lambda + a - a_{0000} > 2$ where the convention concerning λ is unchanged (see Section 6). The proof proceeds as for $q = 3$ and is omitted for concision. Concerning the prior placed on $\boldsymbol{\theta}_{\text{Sat}}$ we advocate to use a $\mathscr{D}(1/2^{q-1}, \ldots, 1/2^{q-1})$ with $q = 4$, for two reasons. First, this choice

Table 2: Characteristics of each type of graphical model ($q = 4$)

model	assumptions	factorization
	none	$p(x_1, x_2, x_3, x_4)$
	$2 \perp 4 \mid (1,3)$	$p(x_2 \mid x_1, x_3)p(x_4 \mid x_1, x_3)p(x_1, x_3)$
	$1 \perp 3 \mid (2,4)\,;\ 2 \perp 4 \mid (1,3)$	no factorization
	$\perp (1,2,3) \mid 4$	$p(x_1 \mid x_4)p(x_2 \mid x_4)p(x_3 \mid x_4)p(x_4)$
	$1 \perp 3 \mid 2;\ 2 \perp 4 \mid 3;\ 1 \perp 4 \mid (2,3)$	$p(x_1 \mid x_2)p(x_4 \mid x_3)p(x_2, x_3)$
	$4 \perp (1,2,3)$	$p(x_1, x_2, x_3)p(x_4)$
	$4 \perp (1,2,3)\,;\ 1 \perp 3 \mid 2$	$p(x_4)p(x_1 \mid x_2)p(x_3 \mid x_2)p(x_2)$
	$(1,2) \perp (3,4)$	$p(x_1, x_2)p(x_3, x_4)$
	$2 \perp 4\,;\ (1,3) \perp (2,4)$	$p(x_1, x_3)p(x_2)p(x_4)$
	$\perp (1,2,3,4)$	$p(x_1)p(x_2)p(x_3)p(x_4)$

ensures the existence of $E[N|\mathbf{y}]$ when $\pi(N) = 1/N$. Second, it ensures that all the marginal and conditional capture probabilities follow non informative prior distributions. Indeed, it is easy to check that: the marginal capture probalities follow a uniform distribution, the conditional probabilities capture follow a Jeffreys distribution when the conditioning is on one list, and a Beta $(1/4, 1/4)$ when the conditioning is on two lists. The use of the uniform distribution will suffer, in a way more accute than for $q = 3$, from the drawback mentioned in Section 4.2 since the prior on a marginal capture probability will now follows a Beta$(8, 8)$ (which is far from being able to be considered as non informative). A similar remark applies to the Jeffreys prior (though to a lesser extent).

• The approach developped in this paper assumes that the capture probabilities do not depend on individual characteristics. To take into account some discrete individual characteristics (for example, the sex) two approaches have been considered. Hook and Regal (1985) propose to conduct separate analyses (one for men and another one for women). Madigan and York (1997) will include the individual variable (in our example, the sex) in the graph, at the same level as the lists.

• In public health, it is important to know if a given disease falls within the field of rare diseases, or not. For a population of interest $\mathcal{P}$, a disease is qualified of *rare*, if its prevalence (equal to the ratio of the number N of people belonging to $\mathcal{P}$ and affected by this disease over the size of $\mathcal{P}$) is smaller than a threshold fixed by the competent authorities. Because the size of $\mathcal{P}$ is known in practice, this issue can be addressed by the test:

$$\mathrm{H}_0 : N \leq N_0 \qquad {}'\mathrm{vs} \qquad \mathrm{H}_1 : N > N_0$$

where N_0 is the product of the above threshold by the size of $\mathcal{P}$. We will typically conclude for H_0, if $\Pr(N \leq N_0|\mathbf{y})$ is enough high (typically 0.95); see Robert (2007). Now, computing $\Pr(N \leq N_0|\mathbf{y})$ is straightforward since

$$\Pr(N \leq N_0|\mathbf{y}) = \sum_m p(m|\mathbf{y}) \Pr(N \leq N_0|m, \mathbf{y})$$

and $\Pr(N \leq N_0|m, \mathbf{y})$ can be easily computed by the Gibbs sampling implementing for obtaining $E(N|m, \mathbf{y})$ because $\Pr(N \leq N_0|m, \mathbf{y}) = E[\mathbb{1}_{(N \leq N_0)}|m, \mathbf{y}]$.

References

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

Chao, A. *et al.* (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medecine* **20**, 3123-3157.

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

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 Schwarz, J.C. (2007). A Bayesian approach to the multistate Jolly-Seber capture-recapture model. *Biometrics* **63**, 1015-1022.

Hojsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical Models with R*. Springer.

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

Kotz, S., Balakrishnan N., and Jonhson, N. (2004). *Continuous Mutivariate 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.

Wang, X., He C., and Sun D. (2007). Bayesian population estimation for small capture-recapture data using noninformative priors. *Journal of Statistical Planning and Inference*, **137**, 1099-1118.

Robert, C.P. (2007). *The Bayesian Choice: From Decision-Theoretic Fundations to Computational Implementation.* Springer-Verlag.

Appendix A

From the *agregation property* we deduce that

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

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

$$(\alpha_0, \alpha_1, \ldots, \alpha_k) \sim \mathscr{D}(b_0, b_1, \ldots, b_k)$$

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 Kotz *et al.* (2004). From this property, we deduce that if $(\alpha_0, \alpha_1, \alpha_2) \sim \mathscr{D}(a_0, a_1, a_2)$ then $\alpha_1/(\alpha_1+\alpha_2) \sim \text{beta } (a_1, a_2)$. 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

We limit ourself to one model by type. The proof for the other models of same type being similar, it has been omitted for concision.

1. <u>The model [12, 13]</u>.

The graph of model $m = [12, 13]$ includes two maximal cliques, we denote by C_1 and C_2, where $C_1 = \{X_1, X_2\}$ and $C_2 = \{X_1, X_3\}$. As Madigan and York (1997), we denote the corresponding clique probabilities as follows:

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

Due to the following constraints:

$$\theta_{00+} + \theta_{01+} = \theta_{0+0} + \theta_{0+1} = \theta_{0++}, \qquad \theta_{10+} + \theta_{11+} = \theta_{1+0} + \theta_{1+1} = \theta_{1++}.$$

and $\theta_{0++} + \theta_{1++} = 1$, the clique parameter $(\theta_{C_1}, \theta_{C_2})$ involves in fact five unconstrained parameters, for example: θ_{1++}, θ_{11+}, θ_{01+}, θ_{1+1}, θ_{0+1} (all the other parameters being redundant). Recall that the capture probabilities parametrisation involves five parameters, namely: θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$, $\theta_{3|\bar{1}}$. Both parametrizations are linked as follows:

$$\theta_{1++} = \theta_1,\ \theta_{11+} = \theta_1\theta_{2|1},\ \theta_{01+} = (1-\theta_1)\theta_{2|\bar{1}},\ \theta_{1+1} = \theta_1\theta_{3|1},\ \theta_{0+1} = (1-\theta_1)\theta_{3|\bar{1}}.$$

Considering the above equalities, it is immediate to check that there is a one-to-one and onto correspondence between both parametrisations.

2. <u>The model [1, 23]</u>.

The graph of [1, 23] includes two maximal cliques $\{C_1, C_2\}$ where $C_1 = \{X_1\}$ and $C_2 = \{X_2, X_3\}$. We have thus two clique probabilities:

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

where

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

Due to these constraints, we have in fact four unconstrained parameters, for example: θ_{1++}, θ_{+01}, θ_{+10}, θ_{+11}. As far as the capture probabilities parametrisation is concerned, we have the choice between both parametrisations. Assume that we adopt the following parametrisation θ_1, θ_2, $\theta_{3|2}$ and $\theta_{3|\bar{2}}$. The clique probabilities parametrisation and the capture probabilities parametrisation are linked as follows:

$$\theta_{1++} = \theta_1,\ \ \theta_{+01} = (1-\theta_2)\theta_{3|\bar{2}},\ \ \theta_{+10} = \theta_2(1-\theta_{3|2}),\ \ \theta_{+11} = \theta_2\theta_{3|2}.$$

It is clear that there is a one-to-one and onto transformation between both parametrisations.

3. The independant model.

The graph of the independent model includes three cliques $\{C_1, C_2, C_3\}$ where $C_1 = \{X_1\}$ and $C_2 = \{X_2\}$ and $C_3 = \{X_3\}$. We have thus three clique probabilities:

$$\theta_{C_1} = (\theta_{0++}, \theta_{1++}), \quad \theta_{C_2} = (\theta_{+0+}, \theta_{+1+}), \quad \text{and} \quad \theta_{C_3} = (\theta_{++0}, \theta_{++1})$$

where

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

The clique probabilities parametrisation thus involves only three parameters; for example: θ_{1++}, θ_{+1+} and θ_{++1}. Since, the capture probabilities parametrisation involves the parameters $\theta_1, \theta_2, \theta_3$ it is clear that there is a one-to-one and onto transformation between both parametrisations.

4. The saturated model.

The graph of the saturated model includes one (maximal) clique. One has only one clique probability which coincides with $\boldsymbol{\theta}_{\text{Sat}}$ defined in Section 2.

Appendix C

For concision, we limit ourself to one model by type.

1. The model [12, 13].

The direct sens. The hyper-Dirichlet put on the clique parameter $(\theta_{C_1}, \theta_{C_2})$ is such that $\theta_{C_1} \perp \theta_{C_2} | \theta_S$ where $\theta_S = \{\theta_{0++}, \theta_{1++}\}$. Its density is:

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

See eg Dawid and Lauritzen (1993) for details. Margins satisfy:

$$\theta_{C_1} \sim \mathscr{D}(a_{00+}, a_{01+}, a_{10+}, a_{11+}) \quad \text{and} \quad \theta_{C_2} \sim \mathscr{D}(a_{0+0}, a_{0+1}, a_{1+0}, a_{1+1})$$

and are compatible with the prior placed on θ_{sat} since

$$\boldsymbol{\theta}_{\text{Sat}} \sim \mathscr{D}(a_{000}, a_{001}, a_{010}, a_{011}, a_{100}, a_{101}, a_{110}, a_{111}).$$

Recall now that the clique probabilities parametrisation involves in fact only five unconstrained parameters, for example: θ_{1++}, θ_{11+}, θ_{01+}, θ_{1+1} and θ_{0+1}; see Appendix B. Linked to this parametrisation, we introduce the transformation:

$$\Phi: \ \left(\theta_1, \theta_{2|1}, \theta_{2|\bar{1}}, \theta_{3|1}, \theta_{3|\bar{1}}\right) \ \longmapsto \ \left(\theta_1, \theta_1\theta_{2|1}, (1-\theta_1)\theta_{2|\bar{1}}, \theta_1\theta_{3|1}, (1-\theta_1)\theta_{3|\bar{1}}\right).$$

Starting from the density of $(\theta_{C_1}, \theta_{C_2})$, we deduce that $\pi(\boldsymbol{\theta}_m)$ is proportional to $J_\Phi \frac{T_2 T_3}{T_1}$ where the Jacobian J_Φ is equal to $\theta_1^2(1-\theta_1)^2$ and where the term T_1 is equal to $\theta_1^{a_{1++}-1}(1-\theta_1)^{a_{0++}-1}$, T_2 is equal to

$$\left[(1-\theta_1)(1-\theta_{2|\bar{1}})\right]^{a_{00+}-1} \left[(1-\theta_1)\theta_{2|\bar{1}})\right]^{a_{01+}-1} \left[\theta_1(1-\theta_{2|1})\right]^{a_{10+}-1} \left[\theta_1\theta_{2|1}\right]^{a_{11+}-1}$$

and T_3 is equal to

$$\left[(1-\theta_1)(1-\theta_{3|\bar{1}})\right]^{a_{0+0}-1} \left[(1-\theta_1)\theta_{3|\bar{1}})\right]^{a_{0+1}-1} \left[\theta_1(1-\theta_{3|1})\right]^{a_{1+0}-1} \left[\theta_1\theta_{3|1}\right]^{a_{1+1}-1}.$$

Gathering the terms θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$, and $\theta_{3|\bar{1}}$, it is easy to check that $\pi(\boldsymbol{\theta}_m)$ is proportionnal to $D_1\, D_{2|1}\, D_{2|\bar{1}}\, D_{3|1}\, D_{3|\bar{1}}$ where $D_1 = \theta_1^{a_{1++}-1}(1-\theta_1)^{a_{0++}-1}$,

$$D_{2|1} = \theta_{2|1}^{a_{11+}-1}(1-\theta_{2|1})^{a_{10+}-1}, \qquad D_{2|\bar{1}} = \theta_{2|\bar{1}}^{a_{01+}-1}(1-\theta_{2|\bar{1}})^{a_{00+}-1}$$

and

$$D_{3|1} = \theta_{3|1}^{a_{1+1}-1}(1-\theta_{3|1})^{a_{1+0}-1}, \qquad D_{3|\bar{1}} = \theta_{3|\bar{1}}^{a_{0+1}-1}(1-\theta_{3|\bar{1}})^{a_{0+0}-1},$$

from which we immediately deduce that θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$, $\theta_{3|\bar{1}}$ follow independently beta distributions. Moreover, it is immediate to check (using Proposition 1) that they are all compatible.

The converse. It is now assumed that θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$, $\theta_{3|\bar{1}}$ follow independently compatible beta distributions. Compatibility and Proposition 1 imply: $\theta_1 \sim \text{beta}(a_{1++}, a_{0++})$, $\theta_{2|1} \sim \text{beta}(a_{11+}, a_{10+})$, $\theta_{2|\bar{1}} \sim \text{Beta}(a_{01+}, a_{00+})$, $\theta_{3|1} \sim \text{beta}(a_{1+1}, a_{1+0})$ and $\theta_{3|\bar{1}} \sim \text{beta}(a_{0+1}, a_{0+0})$. Re-finding the density of $(\theta_{C_1}, \theta_{C_2})$ - as it appears in (1) - from the one of θ_1, $\theta_{2|1}$, $\theta_{2|\bar{1}}$, $\theta_{3|1}$, and $\theta_{3|\bar{1}}$ proceeds similarly as the direct sense of the proof; therefore, details are omitted.

1. <u>The model [1, 23]</u>.

The hyper-Dirichlet put on the clique parameter $(\theta_{C_1}, \theta_{C_2})$ where

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

is such that $\theta_{C_1} \perp \theta_{C_2}$ where θ_{C_1} follows a $\text{Beta}(a_{1++}, a_{0++})$ and θ_{C_2} follows a $\mathscr{D}(a_{+00}, a_{+01}, a_{+10}, a_{+11})$. Consequently, one has:

$$\pi(\theta_{C_1}, \theta_{C_2}) \propto \left[\theta_{1++}^{a_{1++}-1} (1 - \theta_{1++})^{a_{0++}-1}\right] \prod_{k,l \in \{0,1\}} \theta_{+kl}^{a_{+kl}-1}. \tag{2}$$

We have to prove that (2) is equivalent to θ_1, θ_2, $\theta_{3|2}$ and $\theta_{3|\bar{2}}$ follows independently compatible beta distributions. If the other parametrisation is of concern, one has to prove the same equivalence with θ_1, θ_3, $\theta_{2|3}$, $\theta_{3|\bar{2}}$. From now, we work with the former parametrisation (but similar developments hold with the latter). Recall that the clique probabilities parametrisation involves in fact only four unconstrained parameters: θ_{1++}, θ_{+01}, θ_{+10}, θ_{+11} and that there is a one-to-one and onto transformation between both parametrisations (see Appendix B). Considering that $\theta_{C_1} \perp \theta_{C_2}$ and that $\theta_{C_1} \sim \text{Beta}(a_{1++}, a_{0++})$ it is clear that to prove Proposition 2, we have to prove that $\theta_{C_2} \sim \mathscr{D}(a_{+00}, a_{+01}, a_{+10}, a_{+11})$ if and only if θ_2, $\theta_{3|2}$ and $\theta_{3|\bar{2}}$ follows independently compatible beta distributions.

Assume that $\theta_{C_2} \sim \text{Dirichlet}(a_{+00}, a_{+01}, a_{+10}, a_{+11})$ and consider the transformtion:

$$\Psi : \ \Big(\theta_2,\, \theta_{3|2},\, \theta_{3|\overline{2}}\Big) \ \longmapsto \ \Big((1-\theta_2)\theta_{3|\overline{2}},\, \theta_2(1-\theta_{3|2}),\, \theta_2\theta_{3|2}\Big).$$

Starting from the density of θ_{C_2}, we deduce that $\pi(\theta_2, \theta_{3|2}, \theta_{3|\overline{2}})$ is proportional to $J_\Psi P$ where the product P is equal to

$$\big[(1-\theta_2)(1-\theta_{3|\overline{2}})\big]^{a_{+00}-1} \big[(1-\theta_2)\theta_{3|\overline{2}}\big]^{a_{+01}-1} \big[\theta_2(1-\theta_{3|2})\big]^{a_{+10}-1} \ \big[\theta_2\theta_{3|2}\big]^{a_{+11}-1}$$

and where the Jacobian J_Ψ is equal to $\theta_2(1-\theta_2)$. By gathering the terms θ_2,$\theta_{3|2}$ and $\theta_{3|\overline{2}}$ we deduce that $\pi(\theta_2, \theta_{3|2}, \theta_{3|\overline{2}})$ is proportional to:

$$\theta_2^{a_{+1+}-1}(1-\theta_2)^{a_{+0+}-1} \ \theta_{3|2}^{a_{+11}-1}(1-\theta_{3|2})^{a_{+10}-1} \ \theta_{3|\overline{2}}^{a_{+01}-1}(1-\theta_{3|\overline{2}})^{a_{+00}-1}$$

what means that θ_2, $\theta_{3|2}$ and $\theta_{3|\overline{2}}$ follows independently compatible beta distributions. The converse follows similar lines and is omitted for brievety.

3. <u>The independent model</u>.

The proof is trival since the hyper-Dirichlet distribution put on the clique parameter (C_1, C_2, C_3) is the product of 3 independant beta distributions.

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:

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h\in\mathcal{H}^*} n_h!} \ \frac{\Gamma(N-d+a_{000})}{\Gamma(N+a)} \ \prod_{h\in\mathcal{H}^*} \Gamma(n_h + a_h)$$

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

$$L(\boldsymbol{\theta}_m, N; \mathbf{y}) \propto \frac{N!}{(N-d)!} \theta_1^{d_1} \, (1-\theta_1)^{N-d_1} \, E_{2|1} \, E_{2|\bar{1}} \, E_{3|1} \, E_{3|\bar{1}}$$

where

$$E_{2|1} = \theta_{2|1}^{n_{11+}} \, (1-\theta_{2|1})^{n_{10+}} \qquad E_{2|\bar{1}} = \theta_{2|\bar{1}}^{n_{01+}} \, (1-\theta_{2|\bar{1}})^{N-d+n_{001}}$$

and

$$E_{3|1} = \theta_{3|1}^{n_{1+1}} \, (1-\theta_{3|1})^{n_{1+0}}, \quad E_{3|\bar{1}} = \theta_{3|\bar{1}}^{n_{0+1}} \, (1-\theta_{3|\bar{1}})^{N-d+n_{010}}$$

Morever, one has:

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h \in \mathcal{H}^*} n_h!} B_1 B_{2|1} \, B_{2|\bar{1}} \, B_{3|1} \, B_{3|\bar{1}}$$

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

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

and

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

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

$$L(\boldsymbol{\theta}_m, N; \mathbf{y}) \propto \frac{N!}{(N-d)!} \theta_1^{d_1} (1-\theta_1)^{N-d_1} \theta_2^{d_2} (1-\theta_2)^{N-d_2} E_{3|2} \, E_{3|\bar{2}},$$

where

$$E_{3|2} = \theta_{3|2}^{n_{+11}} \, (1-\theta_{3|2})^{n_{+10}} \quad E_{3|\bar{2}} = \theta_{3|\bar{2}}^{n_{+01}} \, (1-\theta_{3|\bar{2}})^{N-d+n_{100}}$$

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:

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h \in \mathcal{H}^*} n_h!} B_1 \, B_2 \, B_{3|2} \, B_{3|\bar{2}}$$

where B_1 is defined above, $B_2 = \mathrm{B}(d_2 + a_{+1+}, N - d_2 + a_{+0+})$ and

$$B_{3|\bar{2}} = \mathrm{B}(n_{+01}+a_{+01}, N-d+n_{100}+a_{+00}) \quad B_{3|2} = \mathrm{B}(n_{+11}+a_{+11}, n_{+10}+a_{+10})$$

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

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

and

$$I_m(N) = \frac{N!}{(N-d)! \prod_{h \in \mathcal{H}^*} n_h!} B_1 B_2 B_3$$

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:

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

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:

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

Since, one has:

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

it comes:

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

The result follows from $(N+v)^{v-u} \sim N^{v-u}$.

• 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 $w_N = N^{d-t} T_1 T_2 T_3$ where

$$T_1 = \frac{\Gamma(N - d_1 + a_{0++})}{\Gamma(N+a)} \quad \text{and} \quad T_2 = \frac{\Gamma(N - d + n_{001} + a_{00+})}{\Gamma(N - d + n_{001} + n_{01+} + a_{0++})}$$

and $T_3 = \Gamma(N - d + n_{001} + a_{0+1})/\Gamma(N - d + n_{010} + n_{0+1} + a_{0++})$.

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

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

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:

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

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

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

we deduce that

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

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

Appendix F

The averaged posterior mean of N exists if and only if the posterior mean of N exists under each model m. 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 $a - a_{000}$.