Test (2010) 19: 187–208
DOI 10.1007/s11749-009-0147-9

ORIGINAL PAPER

Reference Bayesian methods for recapture models with heterogeneity

Alessio Farcomeni · Luca Tardella

Received: 4 May 2006 / Accepted: 11 April 2009 / Published online: 5 May 2009

Abstract In the context of capture–recapture experiments heterogeneous capture probabilities are often perceived as one of the most challenging features to be incorporated in statistical models. In this paper we propose within a Bayesian framework a new modeling strategy for inference on the unknown population size in the presence of heterogeneity of subject characteristics. Our approach is attractive in that parameters are easily interpretable. Moreover, no parametric distributional assumptions are imposed on the latent distribution of individual heterogeneous propensities to be captured. Bayesian inference based on marginal likelihood by-passes some common identifiability issues, and a formal default prior distribution can be derived. Alternative default prior choices are considered and compared. Performance of our formal default approach is favorably evaluated with two real data sets and with a small simulation study.

Keywords Capture–recapture models · Heterogeneity · Bayesian inference · Population size · Default prior · Model choice

Mathematics Subject Classification (2000) 62F15 · 62G05

1 Setup

The classical framework for dealing with a statistical model for a capture–recapture experiment in a closed population setting typically starts from modeling the binary matrix X of individual recapture histories through the probabilities $P_{ij} = \Pr(X_{ij} = 1)$ that the ith individual is caught at the jth recapture occasion.

A. Farcomeni (✉) · L. Tardella
Sapienza–University of Rome, Rome, Italy
e-mail: alessio.farcomeni@uniroma1.it

A fairly general model framework which dates back at least to Otis et al. (1978) is usually denoted as $\mathcal{M}_{bth}$ and assumes

$$P_{ij} = \begin{cases} P_{1ij}, & \text{until first capture,} \\ P_{2ij}, & \text{for any later capture.} \end{cases}$$

The idea is that P_{ij} may possibly depend on behavioral response to (first) capture, which modifies the probability of capture from P_{1ij} to P_{2ij}, individual propensity of the ith animal and the specific features of the jth trapping occasion. If we let $P_{1ij} = P_{2ij}$, we obtain the so called $\mathcal{M}_{th}$ model, which allows only for time/occasion effect and individual heterogeneity.

There is a huge literature nowadays on the subject. The reader may refer to the most recent overviews like Schwarz and Seber (1999), Pollock (2000), Chao (2001). In particular the Bayesian approach seems to have received a lot of attention in terms of new proposals and models as reviewed in Tardella (2007). However, the impact of prior inputs on the final results has not been focused thoroughly, both in the informative and noninformative contexts. One remarkable exception is the recent work of Wang et al. (2007) on the choice of a suitable default prior for $\mathcal{M}_t$ model. We agree with the spirit of their work and believe that comparative merits of Bayesian techniques should be verified also in the absence of precise information on the parameters involved in the model.

In this paper, instead of reparameterizing linearly the probabilities P_{ij} in the logarithmic or logit scales, as in the more traditional fashion (Huggins 1991; Evans et al. 1994; Coull and Agresti 1999), we keep them in the original scales. In this way we show how one can extend the approach of Tardella (2002), where F, the latent distribution summarizing individual heterogeneity, is not assumed to belong to a parametric family.

More formally let

$$P_{ij} = \begin{cases} \beta + (1-\beta)\alpha[\gamma_j + (1-\gamma_j)\delta_j\theta_i], & \text{until first capture,} \quad (1) \\ \gamma_j + (1-\gamma_j)\delta_j\theta_i, & \text{for any later capture.} \quad (2) \end{cases}$$

The two sets of parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_j, \ldots, \delta_J)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_j, \ldots, \gamma_J)$ allow for a flexible framework which describes how the subject-specific capture propensity $\theta_i \sim F$ interacts with the jth trapping occasion. All the involved parameters γ_j, δ_j, and θ_i have support in $(0, 1)$.

In order to fully understand the basic idea of the new parameterization, one can focus just on (2):

$$P_{ij} = \gamma_j + (1-\gamma_j)\delta_j\theta_i \quad \forall j = 1, \ldots, J, \tag{3}$$

which corresponds to the submodel $\mathcal{M}_{th}$. The individual parameter θ_i can be directly interpreted as a probability. We now explain in more detail the meaning of the parameters defining model $\mathcal{M}_{th}$ as above. The individual propensity θ_i is (linearly) reduced by a factor termed δ_j which depends upon the trapping occasion and might make the unobserved probability P_{ij} of being captured less than one even in the

presence of the greatest individual propensity. On the other hand, there is a parameter $\gamma_j \in (0, 1)$, also depending upon trapping occasion, which affects the individual propensity θ_i raising by a constant value the combined $\delta_j \theta_i$. The unobserved probability P_{ij} can then be greater than zero even in the presence of the smallest individual propensity. In any case the resulting probability P_{ij} in (3) is always a valid probability in (0, 1) for all possible values of the parameters $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, and $\boldsymbol{\theta}$. The proposed model represents also a first-order approximation of the more conventional logarithmic or logistic links.

We can also give a probabilistic explanation of the new parameterization related to the wildlife contexts: each animal can be captured either with probability γ_j, a baseline probability of being captured just by chance, or as a result of attractive/repulsive features of the trapping device/occasion, which combines independently random trap effectiveness represented by δ_j with the individual propensity θ_i. More formally, we can think as though there were two latent classes with a different probabilistic explanation of the trapping occurrence. An animal at trapping occasion j can happen to fall into the first class with probability γ_j, and in that case it is captured with certainty independently of its individual propensity to be trapped. Otherwise, with complementary probability $1 - \gamma_j$, the animal is then captured with a conditional probability $\theta_i \delta_j$. That is,

$$\begin{aligned} P_{ij} &= \Pr(X_{ij} = 1 | \text{class 1}) \Pr(\text{class 1}) + \Pr(X_{ij} = 1 | \text{class 2}) \Pr(\text{class 2}) \\ &= \Pr(X_{ij} = 1) = 1 \cdot \gamma_j + \theta_i \delta_j (1 - \gamma_j). \end{aligned}$$

In the full model $\mathcal{M}_{bth}$, there are two additional parameters α and β in (1) allowing to model behavioral effects with different probabilities before and after first capture. As an anonymous referee pointed out, the proposed parameterization for model $\mathcal{M}_{bth}$ rules out the possibility of undetectable individuals, which represents an obvious source of nonidentifiability in N.

Within this new modeling framework, we aim at developing effective tools for a full Bayesian analysis particularly in the absence of precise prior information on the heterogeneity of the individual propensities and on the trapping characteristics. We point out that our new modeling framework allows us to derive and advocate a formal noninformative default prior input (Bernardo 1979, 2005, Berger 2006), namely a Jeffreys prior. For a statistical model with potential lack of robustness, the availability of a Jeffreys prior can be an important achievement also in light of its optimal local robustness properties (Gustafson 1996).

The remainder of the paper is organized as follows: in the next section we discuss identification of our newly proposed parameterization and show the likelihood structure. In Sect. 3 we consider four alternative priors for our model, starting with the Jeffreys prior on the whole parameter vector, and illustrate how one can carry out model fitting and model comparison in a Bayesian framework. In Sect. 4 we first show results of a small simulation study in order to compare alternative prior distributions and highlight advantages of more formal default prior input, and then we entertain two real data examples. In Sect. 5 we provide concluding remarks.

2 Identifiability and likelihood evaluation

We now give a full account of the likelihood structure. In fact, we have to overcome a couple of identifiability issues: the first one is related to the $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ components, while the second one is related to the unrestricted distribution F for θ_i. In fact, Link (2003) points out that in the particular case of $\mathcal{M}_h$ model, without parametric restriction on F, the conditional likelihood yields a nonidentifiable model. Indeed, overcoming this second source of nonidentifiability has been thoroughly discussed in Farcomeni and Tardella (2008) for $\mathcal{M}_h$. At the end of this section we extend the main arguments for $\mathcal{M}_{th}$.

The first identifiability issue arises from the fact that with $a_j = \gamma_j$ and $b_j = (1 - \gamma_j)\delta_j$, it can be shown that there are at least two sets of values, $\{a_j, b_j\}_{j=1}^J$ and $\{a'_j, b'_j\}_{j=1}^J$ (see Appendix A for details), which lead to identical capture probabilities $P_{ij} = P'_{ij}$. This specific issue can be circumvented by constraining $(\gamma, \delta) \in (0, 1)^{2J}$ as $\min(\gamma_1, \ldots, \gamma_J) = 0$ and $\max(\delta_1, \ldots, \delta_J) = 1$ (see Appendix A for details). While the meaning of the detection probabilities P_{ij} is not affected by the new identifiability constraints leading to γ^* and δ^*, the latter identified parameters are chosen just as a technical device and can then be seen as extremal versions of the original parameters interpreted in the previous section as baseline probabilities for latent classes. Note also that the identifiability issue discussed here for γ's and δ's is distinguished from the identifiability matters related to N. The identifiability issue related to $\mathcal{M}_{bth}$ can be handled similarly.

To simplify notation we will hereafter focus only on submodel $\mathcal{M}_{th}$, which corresponds to the sole use of (3). The whole $\mathcal{M}_{bth}$ as in (1)–(2) is conceptually just a simple extension.

We now derive the analytic expression of the likelihood function. More precisely, we derive the so-called *marginal* likelihood since in our model $\mathcal{M}_{th}$ the unobserved individual propensity to be trapped θ can be integrated out. The capture history of the ith individual $(x_{i1}, \ldots, x_{iJ})$ will be equivalently expressed in terms of the subset of trapping occasions when the ith animal is trapped and denoted with G_i, the subset of indexes $g \in \{1, 2, \ldots, J\}$ such that $x_{ig} = 1$ for the ith individual. The likelihood factor corresponding to the ith individual is then

$$
\begin{aligned}
P(G_i; \boldsymbol{\gamma}, \boldsymbol{\delta}, F) &= L(\boldsymbol{\gamma}, \boldsymbol{\delta}, F; G_i) \\
&= \int_{[0,1]} \prod_{j=1}^{J} P_j(\theta)^{x_{ij}} \left(1 - P_j(\theta)\right)^{1 - x_{ij}} dF(\theta) \qquad (4) \\
&= \int_{[0,1]} \prod_{g \in G_i} \left(\gamma_g + (1 - \gamma_g)\delta_g \theta\right) \cdot \prod_{k \in G_i^c} \left(1 - \gamma_k - (1 - \gamma_k)\delta_k \theta\right) dF(\theta) \qquad (5) \\
&= \int_{[0,1]} \sum_{r=0}^{J} \psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i)\theta^r \, dF(\theta) = \sum_{r=0}^{J} \psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i) m_r(F), \qquad (6)
\end{aligned}
$$

where G_i^c corresponds to the subsets of occasions where the ith animal is not trapped, $\psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i)$ are suitable coefficients of the polynomial expression in θ contained in (5), $m_r(F) = \int_{[0,1]} \theta^r \, dF(\theta)$ is the rth moment of the unknown distribution F. Details for obtaining the polynomial coefficients are in the Appendix B. Sufficient statistics for model $\mathcal{M}_{th}$ turn out to be all counts n_G of the observed capture histories. In fact,

$$L(\boldsymbol{\gamma}, \boldsymbol{\delta}, F; \boldsymbol{G}) = \prod_{i=1}^{N} P(G_i; \boldsymbol{\gamma}, \boldsymbol{\delta}, F) = \prod_{G \in \mathcal{G}} \left(\sum_{r=0}^{J} \psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; \boldsymbol{G}) m_r(F) \right)^{n_G}, \quad (7)$$

where $\mathcal{G}$ is the collection of all (2^J) capture history subsets. The particular case of $G = \emptyset$ corresponds to those $n_\emptyset = N - n$ unobserved animals with binary capture history $(0, 0, \ldots, 0)$ which have been never captured, while n denotes the number of distinct (observed) animals trapped at least once.

We now come to the second identifiability issue mentioned at the beginning of this section and which was in fact raised by some of the referees. The reference approach could be questionable in the case of nonidentifiability of the parameters, as it occurs in the particular case of $\mathcal{M}_h$ model when the conditional likelihood is used (Link 2003) and no restriction on F is made. We point out that this is not the case for our Bayesian analysis where the marginal (rather than conditional) likelihood is considered. As a first hint of the difference of the two likelihood approaches, one can just take the counterexample in Link (2003), where it is shown that the conditional likelihood is the same when the F distribution have the same first J moments up to a proportionality constant. Then one can verify by simple computations that this is no longer true for the marginal likelihood. Indeed, in Farcomeni and Tardella (2008) the identifiability of the marginal likelihood is formally proved when the model is expressed in terms of N and the moment-based parameterization of detection probabilities $\theta_i \sim F$, with no distributional assumption on F. Similar arguments can be used for extending the result to $\mathcal{M}_{bth}$ and hence proving that there are no other sources of nonidentifiability. This could sound puzzling for those who often appeal to the well-known result of Sanathanan (1972), which states that marginal likelihood and conditional likelihood are asymptotically equivalent. The puzzle is solved in Farcomeni and Tardella (2008) pointing out that Sanathanan's Theorem 2 requires assumption A2 (on p. 148), which is not met in our nonparametric setting. Rigorous proofs and a more extensive discussion can be found in Farcomeni and Tardella (2008). Of course we have to acknowledge, as a referee pointed out, that the identified model can still yield a nearly flat likelihood and hence leading to annoying near-nonidentifiability issues. The use of the Bayesian framework can alleviate somehow this problem.

3 Default priors and Bayesian analysis

3.1 Default priors

In order for a Bayesian analysis to be carried out, one needs in the first place to elicit prior distributions on the unknown parameters of the model. In model $\mathcal{M}_{th}$ one must specify a full joint prior distribution on the following unknown parameters:

1. The parameter of main interest N
2. The first J moments of F denoted as $\boldsymbol{m} = (m_1(F), \ldots, m_r(F), \ldots, m_J(F))$
3. The $\boldsymbol{\gamma}$ parameters
4. The $\boldsymbol{\delta}$ parameters.

We will provide empirical evidence that the critical elicitation concerns not only the identifiable features of F but also the remaining $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ parameters. In (7) it is shown that these identifiable features correspond to the first J moments $(m_1, \ldots, m_r, \ldots, m_J)$ of F, where $m_r = \int \theta^r \, dF(\theta)$ stands for the ordinary rth moment of F. However, we will also consider a reparameterization of the first J moments in terms of the so-called canonical moments $(c_1, \ldots, c_r, \ldots, c_J) \in (0, 1)^J$. Canonical moments date back at least to the work of Skibinsky (1967), who introduced the terminology and studied in depth their properties in a sequel of papers (Skibinsky 1968, 1970, 1986). There is also a wonderful monograph (Dette and Studden 1997) illustrating many useful applications of canonical moments in several branches of statistics. For our purposes, it suffices to mention the basic fact that the rth canonical moment represents the relative position of the rth moment within the interval of all admissible values of the rth moment given fixed values of the first $r-1$ moments. More formally, we denote an admissible $(r-1)$-tuple of first moments with $\boldsymbol{m}_{(r-1)}$ and with $\mathcal{F}_{\boldsymbol{m}_{(r-1)}}$ the nonempty class of distributions F supported on $[0, 1]$ which share the same first $r-1$ moments $\boldsymbol{m}_{(r-1)}$. Also we denote

$$m_r^-(\boldsymbol{m}_{(r-1)}) = \inf_{F \in \mathcal{F}_{\boldsymbol{m}_{(r-1)}}} \int_{[0,1]} \theta^r \, dF(\theta),$$

$$m_r^+(\boldsymbol{m}_{(r-1)}) = \sup_{F \in \mathcal{F}_{\boldsymbol{m}_{(r-1)}}} \int_{[0,1]} \theta^r \, dF(\theta)$$

the extremes of the interval, so that we can write (implicitly) the one-to-one mapping $\eta(\boldsymbol{c})$ as

$$m_r = (1 - c_r) \, m_r^-(m_{(k-1)}) + c_r \, m_r^+(m_{(k-1)}),$$

where $m_1^-(m_{(0)}) = 0$ and $m_1^+(m_{(0)}) = 1$. See Dette and Studden (1997) for in depth computational details of the one-to-one mapping between the ordinary moment space and the canonical moment space.

Although there is substantial literature on the advantages of the reference approach, we believe it is not wise to neglect subjective elicitation when prior information is actually available. Moreover, Link (2003) argues that only with explicit hypotheses on the characteristics of the distribution of the heterogeneous probabilities one can produce reasonably safe estimates on the unknown population size. However, we will argue that:

– On one hand, it often happens that neither those explicit hypotheses nor other sources of prior information are genuinely available. In that case it is particularly important to rely on a formal default approach. In any case one must recognize that, despite the fact that the Bayesian approach has received increasing attention in terms of new proposals and models, even in the applied field the use of genuine

prior information is the exception rather than the rule. Overall prior information is usually elicited only vaguely.
– On the other hand, the inferential difficulties for the inference on N highlighted in Link (2003) are related just to the inference based on the conditional likelihood and do not apply to the likelihood inference or Bayesian inference based on the marginal (or complete) likelihood as in (7).

We stress the fact that the main reason leading us to propose here a new alternative modeling framework as in (3) is that one can rely on formal noninformative (reference) specification, which we describe below.

Jeffreys prior is defined as

$$\pi(\phi) = \sqrt{|I(\phi)|}, \tag{8}$$

where $|\cdot|$ denotes a matrix determinant computation, ϕ is the vector of continuous parameters involved in the model, and $I(\phi)$ is the Fisher information matrix, whose generic element can be given under standard regularity conditions by

$$I(\phi)_{ij} = E\left[\left(\frac{\partial}{\partial\phi}\log L(\phi)\right)^{\mathrm{T}}\left(\frac{\partial}{\partial\phi}\log L(\phi)\right)\right]_{ij} = -E\left[\frac{\partial^2 \log L}{\partial\phi_i\,\partial\phi_j}\right], \tag{9}$$

and $\log(L)$ is the log-likelihood. A Jeffreys prior $\pi(\phi)$ on a vector of parameters ϕ has many interesting properties, the main one being its invariance to reparameterizations. Refer to Bernardo and Smith (1994) for in-depth account. In the following we will also propose other default choices and compare them with the Jeffreys prior for all the continuous parameters involved in the model.

Our marginal likelihood structure in (7) contains no integral part and is explicitly derived in terms of moments of F. This simplifies the task of deriving a default prior for the parameters at stake, in particular, for those related to the distribution F. Tedious derivation of the Jeffreys prior corresponding to this model for the moments $m_r(F)$, $r = 1, \ldots, J$, and for the whole set of parameters is detailed in Appendix C.

Besides the formal Jeffreys prior, denoted as $\pi_1(\cdot)$, three other alternative types of prior inputs will be derived (see Appendix C for further details). More precisely, we will consider:

– *Complete Jeffreys*: a fully formal joint reference prior on $(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{m})$, or, equivalently, on $(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{c})$, which we denote by

$$\pi_1(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{c}) \propto \sqrt{|J(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{c})|}$$

and which we will finally advocate as a more convenient and neat noninformative choice.
– *Conditional Jeffreys*

$$\pi_2(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{c}) \propto \pi_{CJ}(\boldsymbol{c}|\boldsymbol{\delta}, \boldsymbol{\gamma}),$$

which again independently combines naive uniform noninformative prior on $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ with a conditional reference prior (Berger and Bernardo 1992) on the truncated

moment space, equivalently rewritten with the appropriate Jacobian in terms of the corresponding distribution on the canonical moments $\boldsymbol{c}$.

– *Naive uniform with ordinary moments*

$$\pi_3(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{c}) \propto \prod_{i=1}^{J} \big(c_i(1-c_i)\big)^{J-i},$$

which independently combines naive noninformative priors on $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ as uniform distributions and a naive uniform prior on the constrained space of the first moments of F, again rewritten here in terms of the corresponding distribution on the canonical moments $\boldsymbol{c}$.

– *Naive uniform with canonical moments*

$$\pi_4(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{c}) \propto 1,$$

which independently combines naive uniform noninformative priors on $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ with a naive uniform prior on $\boldsymbol{c}$, within the unconstrained space $(0,1)^J$ of the first J canonical moments of F.

All proposed priors have the common feature of being independent of N. Notice that only the first prior π_1 is fully invariant under reparameterization, and the uniform priors π_3 and π_4 represent alternative ways of specifying uniform measures within the moment space.

The involved expression for π_1 did not allow us to ascertain whether the Jeffreys prior is a proper distribution. However, in Appendix C we are able to give at least weak conditions (easily met in practice) under which the posterior is guaranteed to be proper. We anticipate that those conditions are easily met in the datasets considered in Sect. 4. In the particular case of submodel $\mathcal{M}_h$ propriety of the Jeffreys prior has been already shown in Tardella (2002).

For N, we will closely follow considerations in Tardella (2002) and specify a Rissanen prior, which is always proper and is given by $\pi(N) \propto 2^{-\log^* N}$, where $\log^* N$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2\{\log_2 N\}, \ldots\}$. Such a distribution was originally derived from formal information-theoretic considerations by Rissanen (1983) as a universal prior on the positive integers. Other priors, charging less heavily the right tail of the distribution of N will be considered such as $\pi(N) \propto N^{\lambda}$ with $\lambda \in \{-2, -1, 0\}$ truncating to a convenient upper value of N to avoid impropriety. The priors used in this paper for N are always proper.

3.2 Bayesian model estimation

Once the model has been chosen and a prior distribution has been elicited on the full parametric space, one needs to update it through the Bayes theorem. The usual way to perform this step is to rely on the powerful MCMC machinery (Robert and Casella 1999; Gilks et al. 1996; Liu 2001) by which one can get a sequence of simulated samples from, approximately, the posterior distribution and then approximate integral (posterior) quantities via empirical averages. We omit details of the basic MCMC

strategy and its theoretical background and refer to some classical textbooks on that subject: Gilks et al. (1996), Robert and Casella (1999), Liu (2001). Computations are handled by a Gibbs sampler. When the full conditional is not recognized as a standard distribution, we have used ARMS (Gilks et al. 1995) routine to draw from the full conditional. In the likelihood evaluation an identifiability step is added so that only the restricted parameter space is relevant in the posterior exploration. This is quite easy and involves a simple transformation, which is derived in Appendix A. Code for model estimation written in `C` or in `R` (R Development Core Team 2007) is available from any of the authors upon request.

An important detail to be provided is that as far as the identifiable features of F are concerned, namely $(m_1(F), \ldots, m_r(F), \ldots, m_J(F))$, all the simulation task is carried out in the unconstrained space of the so-called canonical moments of the distribution F, supported in $[0, 1]$, rather than in the constrained space of the first moments of F. We point out that we make use of canonical moments just for technical reasons due to the difficulties in dealing with (and simulating within) the constrained convex body of ordinary moment space, rather than for their interpretation. For more details on canonical moments and an explicit form of the one-to-one mapping via the inverse of the so-called quotient-difference algorithm, one can refer to Dette and Studden (1997) and Tardella (2002) for their usefulness in the Bayesian analysis of capture–recapture models. In particular, in our case one can do all the computations in the unconstrained parameter space and finally reparameterize back into the space of the ordinary moments with little extra effort. Of course all the prior distributions which have been elicited in the original moment parameterization, once reparameterized, need to include the Jacobian of the mapping $c \to \boldsymbol{m} = \eta(\boldsymbol{c})$. This Jacobian is available in closed form as

$$\prod_{i=1}^{J} \big(c_i(1-c_i)\big)^{J-i} \tag{10}$$

(see Chang et al. 1993).

3.3 Bayesian model choice

We now complete the Bayesian toolkit for a default Bayesian analysis of capture–recapture data with effective ways of discriminating and choosing among submodels of $\mathcal{M}_{bth}$. One possibility would be to use Bayes factors (Kass and Raftery 1995), that is, the ratio of normalizing constants of the posterior distributions computed respectively under each hypothesis. However, the use of the Bayes factor requires that the prior distributions are proper, otherwise the ratio of integrated likelihoods may be arbitrarily altered. In the presence of improper priors one could use alternative Bayes factors.

Here, we are not able to ascertain in general the propriety of π_1 and π_2. Hence we will opt for the method of Ghosh and Norris (2005) based on Mean Square Predicted Error (MSPE), instead of Bayes factors. Let n_G be the counts for the observed capture histories, and $\hat{n}_G$ the posterior predictions. Both quantities are indicized in the set $\mathcal{G}$

of all the possible capture histories. The MSPE is equal to the posterior expectation of

$$\frac{\sum_{G\in\mathcal{G}}(\log(n_G+0.5)-\log(\hat{n}_G+0.5))^2}{|\mathcal{G}|}, \tag{11}$$

where $|\mathcal{G}|$ denotes cardinality of all capture histories.

One can compute an MSPE for each submodel of $\mathcal{M}_{bth}$ (and prior choice) and then select the model/prior combination with lowest MSPE.

Here we prefer MSPE to Bayes factors not exclusively for the possibility of dealing with improper priors. Indeed, it also naturally penalizes for model complexity, it was specifically devised for discriminating among recapture models, and further it is very easy to implement.

As a final remark, note that when the data do not express a decisive evidence in favor of one model, another possibility is to implement Bayesian model averaging (see Hoeting et al. 1999 for a review).

4 Applications and simulations

4.1 Simulations

In order to investigate the sensitivity to the prior choice and highlight possible advantages in using formal rules to derive a default prior, we have conducted a small simulation study with $J = \{4, 6\}$, $N = \{100, 1000\}$. We have simulated $B = 100$ datasets and compared the performance of Bayesian estimates under different settings. The fabricated data were generated randomly combining grids of J equally spaced δ_j and γ_j values, respectively, in the intervals $(0.1, 1)$ and $(0, 0.2)$. For instance, when $J = 4$, we set $\boldsymbol{\delta}$ equal to a random permutation of $(0.1, 0.4, 0.7, 1)$ and $\boldsymbol{\gamma}$ to a different random permutation of $(0, 1/15, 2/15, 0.2)$; and similarly for $J = 6$. We have also varied the order of magnitude of the true population size $N \in \{100, 1000\}$ and the latent F distribution by drawing θ_i alternatively from: a uniform, a $Beta\langle 0.5, 1\rangle$, which places most of the probability mass towards zero, and finally a $Beta\langle 0.5, 0.5\rangle$, which is known to cause problems with many estimators, although it may be argued that it corresponds to a not very common situation for real data applications.

We also note that when $N = 100$, the simulated data may originate sparse tables as in the real data example of Sect. 4.3.

For each simulation setting and $s \in \{1, \ldots, B\}$, we let the MCMC sampler run for 10000 iterations, discard the first 3000, and compute the Bayesian estimate $\hat{N}_s$. For each combination of default prior π_i $(i = 1, 2, 3, 4)$ and the three different F distributions above specified, we report the relative mean square error (*RMSE*), that is, $\sqrt{\frac{1}{B}\sum_{s=1}^{B}(\hat{N}_s/N-1)^2}$, where $\hat{N}_s$ is the posterior mean, and N is the known true value of the population size. The comparison of the performance of the Bayesian estimates can be deduced by the results displayed in Tables 1, 2, 3, and 4.

It can be seen that in almost all cases the lowest RMSE is achieved with π_1 and π_2, i.e., with those priors based on Jeffreys rule.

Table 1 *RMSE*, $N = 100$, $J = 4$, Rissanen prior on N

Prior	π_1	π_2	π_3	π_4
	compl. Jeffreys	cond. Jeffreys	unif. mom.	unif. can. mom.
F				
$Beta\langle 0.5, 0.5\rangle$	**0.101**	0.114	0.160	0.168
$Beta\langle 1, 1\rangle$	0.116	**0.072**	0.145	0.156
$Beta\langle 0.5, 1\rangle$	**0.117**	0.179	0.215	0.218

Table 2 *RMSE*, $N = 100$, $J = 6$, Rissanen prior on N

Prior	π_1	π_2	π_3	π_4
	compl. Jeffreys	cond. Jeffreys	unif. mom.	unif. can. mom.
F				
$Beta\langle 0.5, 0.5\rangle$	0.086	**0.068**	0.132	0.138
$Beta\langle 1, 1\rangle$	0.069	**0.054**	0.075	0.081
$Beta\langle 0.5, 1\rangle$	**0.082**	0.098	0.127	0.120

Table 3 *RMSE*, $N = 1000$, $J = 4$, Rissanen prior on N

Prior	π_1	π_2	π_3	π_4
	compl. Jeffreys	cond. Jeffreys	unif. mom.	unif. can. mom.
F				
$Beta\langle 0.5, 0.5\rangle$	0.091	**0.053**	0.063	0.062
$Beta\langle 1, 1\rangle$	0.068	0.053	**0.046**	0.052
$Beta\langle 0.5, 1\rangle$	0.080	**0.063**	0.095	0.101

Table 4 *RMSE*, $N = 1000$, $J = 6$, Rissanen prior on N

Prior	π_1	π_2	π_3	π_4
	compl. Jeffreys	cond. Jeffreys	unif. mom.	unif. can. mom.
F				
$Beta\langle 0.5, 0.5\rangle$	0.032	**0.029**	0.029	0.032
$Beta\langle 1, 1\rangle$	**0.023**	0.029	0.025	0.029
$Beta\langle 0.5, 1\rangle$	**0.031**	0.033	0.034	0.034

Basically, when N is small, J is small, or F puts much mass towards low capture propensity values, i.e., when the sample coverage is lower, the complete Jeffreys prior $\pi_1(\cdot)$ seems to outperform the others more often, while when the sample coverage is higher, the conditional Jeffreys $\pi_2(\cdot)$ seems to be preferable.

The RMSE obviously decreases with N and J, since more capture information allows for a better estimation of N. It can be also noted that the sensitivity to the

Table 5 Hepatitis data: estimates of N and Credibility Intervals obtained using a Rissanen prior on N and, conditionally on N, a Jeffreys prior on $(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{m})$ for model $\mathcal{M}_{th}$ and a Jeffreys prior on $\boldsymbol{m}$ for model $\mathcal{M}_h$

Panel A		Panel B	
Model $\mathcal{M}_{th}$		Model $\mathcal{M}_h$	
Post mean	667.23	Post. mean	629.19
Post median	624	Post. median	610
Post mode	593	Post. mode	579
$\hat{N}_{RMSE}$	716	$\hat{N}_{RMSE}$	652
95% CI	(453, 1170)	95% CI	(425, 881)
80% CI	(495, 885)	80% CI	(484, 803)
95% CI, one-side	1032	95% CI, one-side	858
80% CI, one-side	768	80% CI, one-side	740

prior choice decreases with N and J, and in fact when $J = 6$ and $N = 1000$, the four priors yield often to a very similar performance.

Our results agree with Wang et al. (2007), who show with a more extensive simulation study under an M_t model that the preferable default prior depends on the number of sampling occasions. According to our small study, it depends on J too but also on N and F, and the general recommendation is that in the presence of lower coverage one should use the Jeffreys prior π_1, while the conditional Jeffreys π_2 may be preferable when J is higher and data are not sparse. An interesting development not pursued here is to find out effective formal tools for appropriately selecting the most convenient prior.

4.2 Real dataset 1: *hepatitis A–Taiwan 1995*

We now analyze an epidemiological dataset in which capture–recapture techniques have been used for recovering the real spread of an outbreak of hepatitis A virus in a college in Northern Taiwan during the period April–May 1995. This is a challenging data set with a low coverage and particularly intriguing for an ex-post screening, which made available a more reliable evaluation of the real number of infected people as 545 (Chao et al. 2003), which can be considered at least as a close lower bound. The number of individuals observed at least in one of the three lists (capture occasions) is 271. The full dataset is reported in Chao et al. (2003).

We performed a fully Bayesian analysis of model M_h and M_{th} as described in (3). We have used a Rissanen prior on $N - n$ and, conditionally on N, alternative priors (π_1, π_2, π_3 and π_4) on the remaining continuous parameters. Posterior estimates are obtained with MCMC approximations.

The observed counts and alternative fits given by our model can be seen from Farcomeni and Tardella (2007), while Table 5 shows the estimated size of infected people for this data resulting from using Jeffreys priors. Indeed results displayed in Table 5 are well in agreement with the ex-post information and give a convincing representation of uncertainty in the light also of a reasonable fit. In order to perform a sensitivity analysis of our conclusions with respect to the choice of the prior, we fit

Table 6 Hepatitis data: effect of chosen prior on posterior inference

Prior on $(\boldsymbol{m}, \boldsymbol{\delta}, \boldsymbol{\gamma})$	Mean	Median	$HPD_{0.95}^{\text{low}}$	$HPD_{0.95}^{\text{upp}}$
π_1	667.23	624	453	1170
π_2	637.11	602	445	1031
π_3	602.20	574	417	968
π_4	630.37	584	421	1113

Table 7 Hepatitis data: mean and standard deviation for MSPE, with 2.5 percentile, median, and 97.5 percentile

Model	Prior	Mean	Sd	2.5%	Median	97.5%
M_{th}	Jeffreys	0.042	0.024	0.010	0.037	0.104
M_{th}	Cond. Jeff.	0.041	0.025	0.009	0.036	0.103
M_{th}	Unif. Mom.	0.056	0.035	0.010	0.047	0.137
M_{th}	Unif. Can. Mom.	0.041	0.026	0.008	0.035	0.109
M_h	Jeffreys	0.020	0.012	0.008	0.016	0.052
M_h	Unif. Mom.	0.020	0.012	0.008	0.017	0.051
M_h	Unif. Can. Mom.	0.020	0.012	0.008	0.017	0.052

our model with different combination of prior choices. We particularly focus on the $(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{m})$ parameters.

Table 6 summarizes the results showing posterior mean and median, together with the lower bounds ($HPD_{0.95}^{\text{low}}$) and upper bounds ($HPD_{0.95}^{\text{upp}}$) of the credibility interval for N corresponding to 0.95 posterior probability.

It can be seen that, for this dataset, the effect of changing prior input on $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ has a relatively moderate impact on the final estimate of N ranging from 602 to 667 with an increase of about 10%, while the relative change in spread of uncertainty in terms of HPD regions can be as relevant as 30%. This confirms the importance of relying on a theoretically well-grounded choice of default priors.

Model $\mathcal{M}_h$ yields slightly higher and more disperse posterior estimates than model M_{th} with slightly larger confidence intervals, which actually embrace well the ones obtained with M_{th}, substantially showing that our proposed Bayesian procedure is reasonably adaptive in nested models.

Finally we consider a formal model choice between M_{th} and M_h carried out with the approach Ghosh and Norris (2005) based on MSPE. Table 7 shows the mean, standard deviation, 2.5, 50, and 97.5 percentile of the MSPE for different models and different priors. Note that, under M_h model, there is only one prior derived from the Jeffreys rule. The MSPE criterion leads us to choose the M_h model with Jeffreys prior. Among the M_{th} models, π_1 and π_2 are among the favorite ones.

With this difficult dataset we also show how our results compare with more classical available options. Using the software CARE (Chao et al. 2003), we got several other estimates for the Taiwan hepatitis dataset, very few of which are in line with the ex-post verification that a lower bound for N is 545 (see $HPD_{0.95}^{\text{low}}$ in Table 8).

Table 8 Hepatitis data: comparison of estimates of N, model $\mathcal{M}_{th}$

Estimator	$\hat{N}$	$HPD_{0.95}^{\text{low}}$	$HPD_{0.95}^{\text{upp}}$
Post. mean $\mathcal{M}_{th}$	667	453	1170
Post. mean $\mathcal{M}_h$	629	425	881
Independent	388	352	442
12/3	416	365	494
12/23	527	412	735
Symmetry	1314	685	2899
Quasi-Symmetry	1313	685	2899
Chao-$\hat{N}$	971	369	5290
Chao-$\hat{N}_1$	508	442	600

Table 9 Snowshoe hare data: mean and standard deviation for MSPE, with 2.5 percentile, median, and 97.5 percentile

Model	Prior	Mean	Sd	2.5%	Median	97.5%
M_{th}	Jeffreys	0.374	0.041	0.305	0.370	0.467
M_{th}	Cond. Jeff.	0.376	0.040	0.308	0.372	0.463
M_{th}	Unif. Mom.	0.388	0.040	0.316	0.386	0.475
M_{th}	Unif. Can. Mom.	0.391	0.042	0.313	0.390	0.476
M_h	Jeffreys	0.415	0.028	0.376	0.411	0.482
M_h	Unif. Mom.	0.422	0.029	0.380	0.418	0.489
M_h	Unif. Can. Mom.	0.423	0.029	0.379	0.418	0.492

4.3 Real dataset 2: *snowshoe hare*

Capture–recapture data about snowshoe hares in six trapping occasions are reported originally in Cormack (1989) and later reanalyzed also in Coull and Agresti (1999) among others. This is a typical example where the matrix of observed counts is sparse. The observed data, together with fits given by our model, can be seen in Farcomeni and Tardella (2007). Preliminarily we discuss model choice. Table 9 shows the mean, standard deviation, 2.5, 50, and 97.5 percentile of the MSPE, computed as the posterior expectation of (11), corresponding to the alternative models and different priors. In this example the MSPE criterion leads us to choose the M_{th} model with the Jeffreys prior π_1.

We display in Table 10 the results of a full Bayesian analysis for the chosen model M_{th} as described in (3). We have used a Rissanen prior on $N - n$ and, conditionally on N, the four proposed priors. Posterior propriety is guaranteed since sufficient conditions (21) are met. We show also two-side and one-side credibility intervals (HPD) obtained from the posterior. We first look at different methods of estimation of the sample size (posterior mean, posterior median, posterior mode, $\hat{N}_{RMSE}$) and conclude that those are sufficiently stable and well in the center of the 95% HPD regions. By $\hat{N}_{RMSE}$ we mean the minimum of posterior loss associated with the loss function

Table 10 Snowshoe hare data: estimates of N and Credibility Intervals obtained using a Rissanen prior on N and, conditionally on N, a Jeffreys prior on $(\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{m})$

	Post. mean	Post. median	Post. mode	$\hat{N}_{RMSE}$	95% CI	80% CI	95% CI one-side	80% CI one-side
π_1	87.93	85	83	89	(75, 111)	(78, 100)	106	94
π_2	84.49	83	80	85	(73, 104)	(76, 95)	99	90
π_3	80.64	80	78	81	(72, 96)	(74, 89)	92	85
π_4	78.94	78	76	79	(71, 93)	(73, 87)	90	83

$l(a, N) = (a/N - 1)^2$. For this dataset, we can see a relatively mild effect of the prior choice on the estimate of N when one considers the posterior estimates, while more pronounced effects when considering HPD regions where Jeffreys prior π_1 results in a greater uncertainty with an HPD about 60% larger than the one corresponding to the prior input π_4. For the debated problem of interval estimates in a capture–recapture context, where often ad hoc solutions are employed, we believe that having larger intervals more honestly summarizes posterior uncertainty.

4.4 Empirical findings

We briefly summarize the main conclusions of our applications:

- Our new model is not only easily interpretable but also performs well in real applications when compared with other classical approaches for estimating the unknown population size.
- With respect to prior choices there is some evidence of sensitivity of risk, posterior estimates, and in particular width of the HPD credibility intervals.
- On simulated data formal default priors are preferable to the other default choices. On both real data sets the MSPE method lead to choose the complete Jeffreys prior within the chosen submodel of M_{bth}.
- Bayesian analysis based on Jeffreys' priors for real data sets not only gives sensible answers but also results in larger HPD intervals. We believe that larger HPD intervals more honestly summarize actual uncertainty. For the Hepatitis data, the estimates are more convincing than those obtained by many other classical estimators.

5 Conclusions and further developments

We have shown how a formal default Bayesian analysis can be carried out for capture–recapture models where the three typical sources of variation for capture probabilities are present: individual heterogeneity as well as trapping occasion heterogeneity and behavioral effects. This has been made possible building up a new model framework (1)–(2) based on a linear reparameterization of the capture probabilities. The marginal likelihood (7) can be easily represented in terms of moments of the distribution F of the unobservable individual propensity to be trapped. The

whole parameterization has also interpretative advantages since all parameters are on the same scale and can be directly compared.

The importance of deriving a formal default prior distribution has been motivated by the seemingly weak tenability of the theory of precise measurement of Edwards et al. (1963) and also by results from a small simulation study. In fact, we have given some empirical evidence of nonrobustness of posterior conclusions coming out even when reasonably natural alternative flat priors are considered. We have also mentioned the optimal local robustness properties of the Jeffreys prior derived in Gustafson (1996).

Notice that individual propensity, behavioral effects, and occasion heterogeneity are strictly interwined in the likelihood. Independent priors on the corresponding parameters may then result in an unintentional bias in the final output. This is as another theoretical argument strongly favoring the use of the Jeffreys prior $\pi_1(\cdot)$.

Our default approach has been successfully tested with two well-known data sets. In particular, our results in the Hepatitis Data turn out to be somehow more convincing than those resulting from classical available techniques. The present work confirms the benefits of the Bayesian approach based on formal default priors already highlighted in Tardella (2002) and Wang et al. (2007) and encourage further extensions and investigation of formal default priors on the whole set parameters (including N) for wider classes of models.

Acknowledgements The authors are grateful to an Associate Editor and three referees for punctual comments and stimulating suggestions, which greatly improved the present work.

Appendix A: Identifiability of the parameterization in δ and γ

Our original parameters $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ can be reparameterized with a one-to-one correspondence $(\gamma_j, \delta_j) \longleftrightarrow (a_j, b_j)$ with $a_j = \gamma_j$ and $b_j = (1-\gamma_j)\delta_j$, so that $P_{ij} = a_j + b_j\theta_i$. The original unconstrained space for the couple $(\gamma_j, \delta_j) \in (0,1) \times (0,1)$ is turned into the product of simplexes $S = \{(a_j, b_j) : a_j \geq 0; b_j \geq 0; a_j + b_j \leq 1; j = 1, \ldots, J\}$ or, equivalently, $S = \{(a_j, b_j) : 0 \leq a_j \leq 1 - b_j \leq 1; j = 1, \ldots, J\}$. With this notation it is easy to argue that there are identifiability concerns when one combines the $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ parameters with the θ_i distribution F supported in $(0,1)$. In fact, the parameter vectors $(a_1, \ldots, a_j, b_1, \ldots, b_J, \theta_1, \ldots, \theta_N)$ and $(a'_1, \ldots, a'_j, b'_1, \ldots, b'_J, \theta'_1, \ldots, \theta'_N)$, where $a'_j = a_j - c\frac{b_j}{d}$, $b'_j = \frac{b_j}{d}$, and $\theta'_i = c + d\theta_i$, give the same capture probabilities P_{ij} and P'_{ij}:

$$P_{ij} = a_j + b_j\theta_i,$$

$$P'_{ij} = a'_j + b'_j\theta'_i = a_j - c\frac{b_j}{d} + \frac{b_j}{d}(c + d\theta_i) = a_j - c\frac{b_j}{d} + c\frac{b_j}{d} + b_j\theta_i = P_{ij}.$$

In order to guarantee $(a'_j, b'_j) \in S$, we need to consider $(a_j, b_j) \in S$ and only (c, d) values for which

$$0 \leq a'_j \leq 1 - b'_j \leq 1, \quad j = 1, \ldots, J.$$

The last condition holds if and only if for each $j = 1, \ldots, J$,

$$\text{(i)} \quad a_j - c\frac{b_j}{d} \geq 0, \qquad \text{(ii)} \quad a_j - c\frac{b_j}{d} \leq 1 - \frac{b_j}{d},$$

or, equivalently,

$$\text{(i)} \quad \frac{c}{d} \leq \frac{a_j}{b_j}, \qquad \text{(ii)} \quad d \geq \frac{b_j}{1 - (a_j - \frac{c^*}{d^*}b_j)}.$$

If we start from $(a_j, b_j) \in S$, we will end up with $(a'_j = a_j - c\frac{b_j}{d}, b'_j = \frac{b_j}{d}) \in S$, i.e.,

$$0 \leq a'_j \leq 1 - b'_j \leq 1, \quad j = 1, \ldots, J,$$

if and only if for each $j = 1, \ldots, J$,

$$\text{(i)} \quad a_j - c\frac{b_j}{d} \geq 0, \qquad \text{(ii)} \quad a_j - c\frac{b_j}{d} \leq 1 - \frac{b_j}{d},$$

or, equivalently,

$$\text{(i)} \quad \frac{c}{d} \leq \frac{a_j}{b_j}, \qquad \text{(ii)} \quad d \geq \frac{b_j}{1 - (a_j - \frac{c^*}{d^*}b_j)}.$$

Hence, in order to enforce identifiability, we restrict the parameter space by fixing (c^*, d^*) such that

$$\frac{c^*}{d^*} = \min\left\{\frac{a_1}{b_1}, \ldots, \frac{a_J}{b_J}\right\}, \quad d^* = \max\left\{\frac{b_1}{1 - (a_1 - \frac{c^*}{d^*}b_1)}, \ldots, \frac{b_J}{1 - (a_J - \frac{c^*}{d^*}b_J)}\right\},$$

or, equivalently,

$$\frac{c^*}{d^*} = \min\left\{\frac{a_1}{b_1}, \ldots, \frac{a_J}{b_J}\right\}; \quad d^* = \max\left\{\frac{1}{\frac{1-a_1}{b_1} - \frac{c^*}{d^*}}, \ldots, \frac{1}{\frac{1-a_J}{b_J} - \frac{c^*}{d^*}}\right\}. \tag{12}$$

The original parameter space is then restricted as follows:

$$S^* = \big\{(a_j^*, b_j^*) : 0 \leq a_j^* \leq 1 - b_j^* \leq 1,\ j = 1, \ldots, J;$$
$$\min\{a_1^*, \ldots, a_J^*\} = 0,\ \max\{a_1^* + b_1^*, \ldots, a_J^* + b_J^*\} = 1\big\}.$$

For any $2J$-tuple $(a_1, \ldots, a_j, b_1, \ldots, b_J)$ with $(a_j, b_j) \in S$, we can now find a unique equivalent (in terms of the resulting $P_{ij} = a_j + b_j\theta_i$) $(a_1^*, \ldots, a_j^*, b_1^*, \ldots, b_J^*)$ with $(a_j^*, b_j^*) \in S^* \subset S$, i.e., the representative of the equivalence class, by setting $a_j^* = a_j - \frac{c^*}{d^*}b_j$ and $b_j^* = \frac{b_j}{d^*}$ with (c^*, d^*) defined as in (12). In fact, let us admit that $\bar{j}$ is such that $\frac{a_{\bar{j}}}{b_{\bar{j}}} = \frac{c^*}{d^*} = \min\{\frac{a_1}{b_1}, \ldots, \frac{a_J}{b_J}\}$; this implies that $a_j^* = a_j - \frac{a_{\bar{j}}}{b_{\bar{j}}}b_j \geq$

$a_j - \frac{a_j}{b_j} b_j \geq 0$ and that $a^*_{\bar{j}} = a_{\bar{j}} - \frac{a_{\bar{j}}}{b_{\bar{j}}} b_{\bar{j}} = 0 = \min\{a^*_j\}$. Analogously, let us admit that $\bar{h}$ is such that

$$\frac{b_{\bar{h}}}{1 - (a_{\bar{h}} - \frac{c^*}{d^*} b_{\bar{h}})} = d^* = \max\left\{ \frac{b_1}{1 - (a_1 - \frac{c^*}{d^*} b_1)}, \ldots, \frac{b_J}{1 - (a_J - \frac{c^*}{d^*} b_J)} \right\};$$

this implies that

$$a^*_j + b^*_j = a_j - \frac{c^*}{d^*} b_j + b_j \frac{1 - (a_{\bar{h}} - \frac{c^*}{d^*} b_{\bar{h}})}{b_{\bar{h}}} \leq a_j - \frac{c^*}{d^*} b_j + b_j \frac{1 - (a_j - \frac{c^*}{d^*} b_j)}{b_j} = 1$$

and also that $a^*_{\bar{h}} + b^*_{\bar{h}} = 1 = \max\{a^*_1 + b^*_1, \ldots, a^*_J + b^*_J\}$.

The above constraints and equivalence mapping can be easily mapped into the original parameterization: for any $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in (0,1)^{2J}$, we end up identifying $(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) \in (0,1)^{2J}$ where $\min\{\gamma^*_1, \ldots, \gamma^*_J\} = 0$ and $\max\{\delta^*_1, \ldots, \delta^*_J\} = 1$.

Appendix B: Derivation of coefficients $\psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i)$

We now show how to obtain $\psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i)$. Consider the expression of the integrand function in (4)

$$\prod_{j \in G} \big(\gamma_j + (1 - \gamma_j)\delta_j \theta\big) \prod_{j \notin G} \big(1 - \gamma_j - (1 - \gamma_j)\delta_j \theta\big), \tag{13}$$

which we rewrite as $\prod_{j=1}^{J} (a_j + b_j \theta)$ as follows:

$$a_j = \begin{cases} \gamma_j, & j \in G, \\ 1 - \gamma_j, & j \notin G, \end{cases} \qquad b_j = \begin{cases} (1 - \gamma_j)\delta_j, & j \in G, \\ -(1 - \gamma_j)\delta_j, & j \notin G. \end{cases} \tag{14}$$

The expression (13) can then be easily recognized as a polynomial of order J formalized as $\sum_{r=0}^{J} \psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i)\theta_i^r$. A useful pseudo-code for recursively obtaining $\psi_r(\boldsymbol{\gamma}, \boldsymbol{\delta}; G_i)$ as a function of $\boldsymbol{a}$ and $\boldsymbol{b}$ is:

```
/* initialization */
 psi[0]= tempsi[0]=a[1] ;
 psi[1]= tempsi[1]=b[1] ;
 for(i=2; i<=J; i++){psi[i]=tempsi[i]=0.0;}
/* end initialization */

for(i=1; i<=(dim-1); i++){
   for(degree=0; degree<=i; degree++){
     tempsi[degree]=psi[degree]*a[i+1] ;}
   for(degree=1; degree<=(i+1); degree++){
     tempsi[degree]=tempsi[degree] + psi[(degree-1)]*b[i+1] ;}
   for(degree=0; degree<(i+1); degree++){psi[degree]=tempsi[degree];}
}
```

Appendix C: Derivation and propriety of the Jeffreys priors

Along the strategy suggested in (8), we now illustrate the derivation of the Jeffreys prior $\pi_1(\cdot)$ and of the conditional Jeffreys $\pi_2(\cdot)$. We start from the last one.

The loglikelihood is $\log L = \sum_{G\in\mathcal{G}} n_G \log(\sum_{r=0}^{J} \psi_r(G) m_r)$, and for deriving $\pi_2(\cdot)$, we consider it only as a function of the J moments so that the first partial derivatives are

$$\frac{\partial \log L}{\partial m_r} = \sum_{G\in\mathcal{G}} n_G \frac{\psi_r(G)}{(\sum_{r=0}^{J} \psi_r(G) m_r)},$$

and the matrix of second derivatives has the following generic entry:

$$\frac{\partial^2 \log L}{\partial m_r \partial m_s} = -\sum_{G\in\mathcal{G}} n_G \frac{\psi_r(G)\psi_s(G)}{[(\sum_{r=0}^{J} \psi_r(G) m_r)]^2}.$$

Taking expectations, we get the matrix with entries

$$-\sum_{G\in\mathcal{G}} \frac{\psi_r(G)\psi_s(G)}{[(\sum_{r=0}^{J} \psi_r(G) m_r)]} = -\sum_{G\in\mathcal{G}} \frac{\psi_r(G)\psi_s(G)}{P(G)}. \tag{15}$$

Once the information matrix has been computed, the only numerical derivation needed is related to the computation of the determinant. The information matrix $I(\phi) = [I(\phi)_{rs}]$ is computed with entries $I(\phi)_{rs}$ as in (15), at the parameter values ϕ sampled at the current iteration of the MCMC algorithm, and then, the Jeffreys prior is set as the square root of the determinant of the information matrix. The quantities $P(G)$ and $\psi_r(G)$ have to be computed in any case for evaluating the likelihood, so there is almost no additional computational effort.

The conditional reference prior $\pi_2(\boldsymbol{m}|\boldsymbol{\gamma}, \boldsymbol{\delta})$ is indeed completed with a uniform distribution on the $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in [0, 1]^2$, and it is then given by

$$\pi_2(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{m}) \propto \sqrt{\left| -\left(\frac{\partial^2 \log L}{\partial m_r \partial m_s} \right)_{rs} \right|}. \tag{16}$$

The computation of the Jeffreys prior $\pi_1(\cdot)$ for the whole set of parameters can be derived using the first formula in (9). In fact, when I is expressed in terms of multinomial cell probabilities P, it can be written as $I(P) = E[(\frac{\partial}{\partial P} \log L(P))^{\mathrm{T}} (\frac{\partial}{\partial P} \log L(P))]$ and has the following diagonal form with generic (G_1, G_2) entry:

$$I(P)_{G_1,G_2} = \begin{cases} \frac{1}{P(G)}, & \text{if } G_1 = G_2 = G, \\ 0, & \text{if } G_1 \neq G_2. \end{cases} \tag{17}$$

Whenever the Fisher information matrix I is considered as a function of the multinomial cell probabilities, namely $\phi = (\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{m}) = \phi(P)$, one can write $I(\phi) = (\frac{\partial P}{\partial \phi})^{\mathrm{T}} I(P(\phi)) (\frac{\partial P}{\partial \phi})$. Again, the Jeffreys prior on the complete set of parameters $\pi_1(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{m})$ can be computed easily deriving the partial derivatives $\frac{\partial P}{\partial \phi}$, where P

is linear with respect to each ϕ_i component. The only numerical step needed is related with the computation of the determinant of the information matrix at the current parameter set.

Tardella (2002) shows that the Jeffreys prior for model $\mathcal{M}_h$ is proper. We now consider whether the proposed priors are proper also for $\mathcal{M}_{th}$. Since $P(G)$ can vanish, $\pi_1(\cdot)$ is not bounded, and we cannot exclude a priori that it is improper. Unfortunately, it is very hard to prove or disprove that π_1 is proper. In any case, we are able at least to show that $\pi_1(\cdot)$ leads to a valid Bayesian inference as long as it leads to a proper posterior.

Since $I(P)$ is diagonal, the generic entry of the Fisher information matrix $I(\phi)$ for the reparameterized vector ϕ can then be written as

$$I(\phi)_{r,s} = \sum_{G \in \mathcal{G}} \frac{\frac{\partial P(G)}{\partial \phi_r} \frac{\partial P(G)}{\partial \phi_s}}{P(G)},$$

so that its determinant can be decomposed as a sum of expressions which can vanish (at most) with the order of $\frac{1}{\prod_{g=1}^{p} P(G_g)}$, where p is the number of free parameters, namely $p = 3J - 2$. The posterior distribution is given by the determinant multiplied by the likelihood function and hence proportional to

$$\frac{\prod_{G \in \mathcal{G}} P(G)^{n_G}}{\prod_{g=1}^{p} P(G_g)^{\frac{1}{2}}}. \tag{18}$$

Indeed, from (14) and the expression of $\psi_r(G)$ one can easily realize that for any fixed $\boldsymbol{m}$ in the interior of the moment space, $P(G)$ can vanish only when either one of the following holds true:

$$\begin{cases} (1 - \gamma_j) \to 0, & \text{when } j \notin G, \\ (\gamma_j, \delta_j) \to (0, 0), & \text{when } j \in G. \end{cases} \tag{19}$$

In fact, the above vanishing behavior of each $\psi_r(G)$ coefficients is shared with all other $\psi_r(G')$ coefficients corresponding to $P(G')$ with G' sharing the same capture outcome of G at the jth occasion, so that in order to control the ratio (18), one can try directly to verify whether the rate of the numerator can lead to

$$0 \leq \frac{\prod_{G \in \mathcal{G}} P(G)^{n_G}}{\prod_{g=1}^{p} P(G_g)^{\frac{1}{2}}} \leq K, \tag{20}$$

which would guarantee that the posterior is proper since the support of the parameter space of ϕ is bounded. Similar arguments apply for $m_1 \to 0$, in which case the numerator of the ratio always dominates the infinitesimal behavior of the denominator, provided that $n \geq p/2$. Suppose that $G \in \mathcal{G}$ is such that $1 \notin G$, i.e., there is a 0 corresponding to the first capture occasion. From (19) and the ensuing arguments we can derive that $\prod_{G \in \mathcal{G}} P(G)^{n_G}$ has an infinitesimal behavior with $(1 - \gamma_1) \to 0$, which is of order $O(1 - \gamma_1)^M$ with $M = \sum_{G: 1 \notin G} n_G$. Similar arguments can be invoked when $1 \in G$, in which case $\prod_{G \in \mathcal{G}} P(G)^{n_G}$ has an infinitesimal behavior with

$(\gamma_1, \delta_1) \to (0, 0)$, which is basically of order $O(\gamma_1 \delta_1)^M$ with $M = \sum_{G:1\in G} n_G$, and similar relations can be derived for all indexes $j > 1$.

At this point we are ready to state that both the following

$$\sum_{G:j\in G} n_G \geq \frac{p}{2} \quad \forall j = 1, \ldots, J; \qquad \sum_{G:j\notin G} n_G \geq \frac{p}{2} \quad \forall j = 1, \ldots, J \tag{21}$$

are sufficient conditions to guarantee that the posterior corresponding to π_1 is proper. Similarly one can adapt the arguments to assess propriety of π_2.

References

Berger JO (2006) The case for objective Bayesian analysis. Bayesian Anal 1:385–402

Berger JO, Bernardo JM (1992) Ordered group reference priors with application to the multinomial problem. Biometrika 79(1):25–37

Bernardo JM (1979) Reference posterior distributions for Bayesian inference (C/R pp 128–147). J R Stat Soc Ser B 41:113–128

Bernardo JM (2005) Reference analysis. In: Handbook of statistics. Bayesian thinking, modeling and computation, vol 25. Elsevier, Amsterdam

Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, New York

Chang F-C, Kemperman JHB, Studden WJ (1993) A normal limit theorem for moment sequences. Ann Probab 21:1295–1309

Chao A (2001) An overview of closed capture–recapture models. J Agric Biol Environ Stat 6(2):158–175

Chao A, Tsay PK, Lin S-H, Shau W-Y, Chao D-Y (2003) Tutorial in biostatistics: the applications of capture–recapture models to epidemiological data. Stat Med 20:3123–3157

Cormack RM (1989) Log-linear models for capture–recapture. Biometrics 45:395–413

Coull BA, Agresti A (1999) The use of mixed logit models to reflect heterogeneity in capture–recapture studies. Biometrics 55:294–301

Dette H, Studden WJ (1997) The theory of canonical moments with applications in statistics, probability, and analysis. Wiley, New York

Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psychol Rev 70:193–242

Evans MA, Bonett DG, McDonald LL (1994) A general theory for modeling capture-recapture data from a closed population. Biometrics 50:396–405

Farcomeni A, Tardella L (2007) Reference Bayesian methods for recapture models with heterogeneity. Technical report, Dept of Statistics, Sapienza–University of Rome

Farcomeni A, Tardella L (2008) Identifiability of population size from capture–recapture data with heterogeneous detection probabilities. Technical report, Dept of Statistics, Sapienza–University of Rome

Ghosh SK, Norris JL (2005) Bayesian capture–recapture analysis and model selection allowing for heterogeneity and behavioral effects. J Agric Biol Environ Stat 10:35–49

Gilks WR, Best NG, Tan KKC (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. Applied Stat 44:455–472. (Corr: 46:541–542 (1997) with RM Neal)

Gilks WR, Richardson S, Spiegelhalter DJ (eds) (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London

Gustafson P (1996) Local sensitivity of posterior expectations. Ann Stat 24:174–195

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14:382–417

Huggins RM (1991) Some practical aspects of a conditional likelihood approach to capture experiments. Biometrics 47:725–732

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Link WA (2003) Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. Biometrics 59:1123–1130

Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York

Otis DL, Burnham KP, White GC, Anderson DR (1978) Statistical inference from capture data on closed animal populations. Wildlife monographs

Pollock KH (2000) Capture–recapture models. J Am Stat Assoc 95:293–296
R Development Core Team (2007) R: A Language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria
Rissanen J (1983) A universal prior for integers and estimation by minimum description length. Ann Stat 11:416–431
Robert CP, Casella G (1999) Monte Carlo statistical methods. Springer, Berlin
Sanathanan L (1972) Estimating the size of a multinomial population. Ann Math Stat 43:142–152
Schwarz CJ, Seber GAF (1999) Estimating animal abundance: review III. Stat Sci 14:427–456
Skibinsky M (1967) The range of the $(n+1)$-th moment for distributions on [0, 1]. J Appl Probab 4:543–552
Skibinsky M (1968) Extreme nth moments for distributions on [0, 1]. J Appl Probab 5:693–701
Skibinsky M (1970) A characterization of hypergeometric distributions. J Am Stat Assoc 65:926–929
Skibinsky M (1986) Principal representations and canonical moment sequences for distributions on an interval. J Math Anal Appl 120:95–118
Tardella L (2002) A new Bayesian method for nonparametric capture–recapture models in presence of heterogeneity. Biometrika 89:807–817
Tardella L (2007) Bayesian capture–recapture. ISBA bulletin, 14(1). Annotated bibliography, available at http://www.bayesian.org/bulletin/0703.pdf
Wang X, He CZ, Sun D (2007) Bayesian population estimation for a capture–recapture model using non-informative priors. J Stat Plan Inference 137:1099–1118