Bayesian Methods for Semiparametric Recapture Models

Metodi Bayesiani per Modelli Semiparametrici di Cattura-Ricattura

Alessio Farcomeni[1], Luca Tardella[1],
Dipartimento di Statistica, Probabilità e Statistiche Applicate; Università "La Sapienza"
e-mail: alessio.farcomeni@uniroma1.it

Riassunto: Si propone un nuovo modello di cattura-ricattura alternativo ai classici modelli basati su riparametrizzazione logistica o log-lineare per sviluppare una procedura inferenziale bayesiana derivante da una distribuzione a priori di tipo Jeffreys. Il modello consente di incorporare sia l'eterogeneità delle probabilità di cattura degli individui sia l'eterogeneità nelle diverse occasioni e si basa su una parametrizzazione alternativa per la quale è possibile derivare in forma chiusa la verosimiglianza marginale. Vengono riportate alcune evidenze empiriche sulla validità dell'approccio su dati di letteratura.

Keywords: Capture-recapture, Model $\mathcal{M}_{th}$, Jeffreys Prior, Marginal Likelihood.

1. Introduction

The classical framework for dealing with a statistical model for capture-recapture experiment in a closed population setting typically starts from modelling the binary matrix of individual recapture history through the probabilities P_{ij} that the individual with label i is caught at the recapture occasion labelled j.

There is a huge literature nowadays on the subject testifying a lot of recent advances and interest in new techniques. 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. However, the impact of prior inputs on the final results has not been focussed thoroughly and only few articles address formally the prior elicitation step emphasizing for instance the role that a non-informative elicitation may play. Comparative merits of Bayesian techniques as a general tool should always be verified also in the absence of a precise information on the parameters involved in the model.

In this work instead of reparametrizing linearly those 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 scale appropriately and easily handling the ensuing bounded parameter space of the coefficients of the linear link. In this way we show how one can extend the approach of Tardella (2002) where the distribution F of the possibly heterogeneous probabilities θ_i is not assumed to belong to a parametric family but can be handled through its first J moments.

More formally let

$$P_{ij} = \gamma_j + (1 - \gamma_j)\delta_j\theta_i \tag{1}$$

The two sets of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ allow for a flexible framework within which the subject-specific capture probability can be reduced by a multiplicative factor and/or increased with the addition of a constant quantity.

The individual propensity parameter named θ_i can take any value in $(0,1)$ and can then be directly interpreted as a probability. The individual probability can be reduced (linearly) by a factor termed δ_j which depends upon the trapping occasion and might make the unobserved probability P_{ij} less than one even in the presence of the greatest individual propensity. Conversely, there is a parameter $\gamma_j \in (0,1)$ depending upon trapping occasion which affect the individual propensity θ_i raising by a constant value the combined propensity $\delta_j\theta_i$ and making the resulting probability P_{ij} in (1) always a valid probability in $(0,1)$ for all possible values of $\theta_i \in (0,1)$, $\delta_j \in (0,1)$ and $\gamma_j \in (0,1)$.

2. Likelihood Evaluation

If we denote the observed data or individual capture history $(x_{i1},...,x_{iJ})$ equivalently in terms of the subset of trapping occasions when the i-th animal is trapped denoted as G_i i.e. the subset of indexes $g \in \{1,2,...,J\}$ such that the i-th individual has $x_{ig}=1$; the sufficient statistics for this model turn out to be all counts n_G of the observed capture histories; in fact

$$\begin{aligned} L(\boldsymbol{\gamma},\boldsymbol{\delta},F;\boldsymbol{G}) &= \int_{[0,1]} \prod_{g\in G_i} (\gamma_g + (1-\gamma_g)\delta_g\theta) \cdot \prod_{k\in G_i^c} (1-\gamma_k-(1-\gamma_k)\delta_k\theta)\, dF(\theta) \\ &= \prod_{G\in\mathcal{G}} \left(\sum_{r=0}^{J} \psi_r(\boldsymbol{\gamma},\boldsymbol{\delta};\boldsymbol{G}) m_r(F) \right)^{n_G} \end{aligned} \tag{2}$$

where $\mathcal{G}$ is the collection of all (2^J) capture histories among which the particular case $\tilde{G}=\emptyset$ corresponding to those unobserved animals which have been never captured; $\psi_r(\boldsymbol{\gamma},\boldsymbol{\delta};G_i)$ are suitable coefficients for the polynomial (in θ) expression involved in the first line of the expression above, which is in fact a polynomial of order J and $m_r(F)=\int_{[0,1]}\theta^r dF(\theta)$ is the r-th central moment of the unknown distribution F.

This simplified explicit likelihood structure allows us to derive a default prior of the parameters at stake, in particular those related to the distribution F of individual propensity to be trapped. This basically justify the use of (1) instead of the most traditional logistic or loglinear reparameterizations.

3. Prior Determination

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 our model one must specify a full joint prior distribution on $(N,\boldsymbol{m},\boldsymbol{\delta},\boldsymbol{\gamma})$; respectively the parameter of main interest N, the first J moments of F denoted as $\boldsymbol{m}=(m_1(F),...,m_r(F),...,m_J(F))$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ parameters. We will provide empirical evidence how critical the elicitation step is not only for the identifiable features of F but also the remaining $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$. We stress on the fact that one of the main reasons leading us to propose here a new modelling framework as in (1) is that one can rely on formal noninformative (reference) specification. For the parameter N we will closely follow considerations in Tardella (2002) and specify a Rissanen prior as $\pi(N) \propto 2^{-\log^* N}$, where $\log^* N$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2\{\log_2 N\},...\}$. Such prior is proper, and it was originally derived from formal information-theoretic considerations by Rissanen (1983) as a universal prior on the positive integers. For the remaining part of the parameter space we considered

a formal noninformative choice such as a Jeffreys prior. Also we have compared the resulting posterior outputs with those implied by alternative less formal naive choices for each component parameter, based on flat or non-informative distributions.

We are not endorsing our reference approach under any circumstance. In fact, when prior info on average characteristics or even on the shape of F is present, one must use that information. 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. That is the reason why one should be interested in a formal evaluation of the possibly large uncertainty on estimates on the unknown population size when those explicit hypotheses are not at all guaranteed or known in advance. Hence a reasonable reference approach is particularly important when prior specification on F in the absence of specific parametric assumptions may lead to non robust inference.

Once a prior distribution has been elicited on the full parametric space one needs to update it through the Bayes theorem and usual way to perform this step is to rely on the powerful MCMC tools (Robert and Casella (1999); Gilks *et al.* (1996); Liu (2001)) by which one can get a sequence of simulations as if they were a random sample from the posterior distribution and then approximate integral quantities via empirical averages of a function of the simulated sequence. We omit details of the MCMC strategy.

4. A Real Dataset: *Hepatitis A - Taiwan 1995*

This dataset is based on an epidemiological study for which capture-recapture techniques have been used for recovering the real spread of an outbreak of hepatitis A virus experienced 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 has made it possible to have a more reliable evaluation of the real number of infected people. In fact, after gathering of supplementary information it has been argued in Chao *et al.* (2003) that 545 is a reasonable ex-post lower bound for the true value of N.

Table 1 summarizes the results with different choices of the prior, showing posterior mean and median, together with the lower bounds ($HPD^{low}_{0.95}$) and upper bounds ($HPD^{upp}_{0.95}$) 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 strong impact on the final estimate of N ranging from 561 to 610 with an increase of about 10%. This is another instance of how crucial the option of a theoretically well grounded default choice can be. Notice that the direction of the change depends on the choice of the prior chosen for F and in fact with the choice of the joint Jefferys prior on $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ the width of the interval estimate is reduced. Again, the change in spread of the final estimate can be as relevant as 100 units corresponding to more than 15% of the estimated population size.

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 while our estimates comparatively best fit that evidence (see Table 2).

Table 1: Hepatitis Data: *Effect of chosen prior on posterior inference.*

Prior on $(\boldsymbol{m}, \boldsymbol{\delta}, \boldsymbol{\gamma})$	Mean	Median	$HPD_{0.95}^{low}$	$HPD_{0.95}^{upp}$
$(\boldsymbol{m}, \boldsymbol{\delta}, \boldsymbol{\gamma}) \sim Jeffreys$	600.85	584.12	449.00	847.24
$m \sim Unif$, $(\delta, \gamma) \sim U[0,1]$	561.22	538.51	416.68	849.88
$m \sim Jeffreys$, $(\delta, \gamma) \sim U[0,1]$	581.72	557.29	431.08	874.98
$m \sim Unif$, $(\delta, \gamma) \sim Beta < 0.5, 0.5 >$	584.49	551.04	426.40	908.12
$m \sim Jeffreys$, $(\delta, \gamma) \sim Beta < 0.5, 0.5 >$	610.02	582.78	441.05	923.92

Table 2: Hepatitis Data: *Estimates of N with classical estimators and our proposal*

Estimator	$\hat{N}$	$HPD_{0.95}^{low}$	$HPD_{0.95}^{upp}$
post. mean	600	449	855
Independent	388	352	442
12/3	416	365	494
12/23	527	412	735
Symmetry	1314	685	2899
Chao-$\hat{N}$	971	369	5290
Chao-$\hat{N}_1$	508	442	600

References

ANNE CHAO (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics*, **6, 2**, 158–175.

ANNE CHAO, HSIN-CHOU YANG, AND PAUL S. F. YIP (2003). The use of capture-recapture methodology in epidemiological surveillance. In: *Advanced medical statistics*, 711–739. World Sci. Publishing, River Edge, NJ.

BRENT A. COULL AND ALAN AGRESTI (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, **55**, 294–301.

M. A. EVANS, D. G. BONETT, AND L. L. MCDONALD (1994). A general theory for modeling capture-recapture data from a closed population. *Biometrics*, **50**, 396–405.

W. R. GILKS, S. RICHARDSON, AND D. J. SPIEGELHALTER, eds. (1996). *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall, London.

R. M. HUGGINS (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, **47**, 725–732.

WILLIAM A. LINK (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, **59, 4**, 1123–1130.

JUN S. LIU (2001). *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New York.

KENNETH H. POLLOCK (2000). Capture-recapture models. *Journal of the American Statistical Association*, **95**, 293–296.

JORMA RISSANEN (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, **11, 2**, 416–431.

CHRISTIAN P. ROBERT AND GEORGE CASELLA (1999). *Monte Carlo statistical methods*. Springer-Verlag Inc.

C.J. SCHWARZ AND G.A.F SEBER (1999). Estimating animal abundance: review iii. *Statistical Science*, **14**, 427–456.

LUCA TARDELLA (2002). A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika*, **89, 4**, 807–817.