Computational Statistics and Data Analysis 55 (2011) 655–666

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Using capture-recapture data and hybrid Monte Carlo sampling to estimate an animal population affected by an environmental catastrophe

Guoqi Qian *, Ning Li, Richard Huggins

Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia

ARTICLE INFO

Article history:
Received 31 August 2009
Received in revised form 4 April 2010
Accepted 7 June 2010
Available online 16 June 2010

Keywords:
Acceptance-rejection sampling
Capture recapture
Gibbs sampler
Population size
Sequential importance sampling

ABSTRACT

We propose a dynamic model for the evolution of an open animal population that is subject to an environmental catastrophe. The model incorporates a capture-recapture experiment often conducted for studying wildlife population, and enables inferences on the population size and possible effect of the catastrophe. A Bayesian approach is used to model unobserved quantities in the problem as latent variables and Markov chain Monte Carlo (MCMC) is used for posterior computation. Because the particular interrelationship between observed and latent variables negates the feasibility of standard MCMC methods, we propose a hybrid Monte Carlo approach that integrates a Gibbs sampler with the strategies of sequential importance sampling (SIS) and acceptance-rejection (AR) sampling for model estimation. We develop results on how to construct effective proposal densities for the SIS scheme. The approach is illustrated through a simulation study, and is applied to data from a mountain pygmy possum (*Burramys Parvus*) population that was affected by a bushfire.

1. Introduction

Open population models are important in ecological and environmental research. Early work on open population inference was largely conducted within the maximum likelihood framework. With advances in simulation techniques, implementation of the integrals required for calculating posterior distribution has become increasingly available. This leads to more studies being conducted via a Bayesian approach. For example, Crome et al. (1996) presented a Bayesian analysis to assess the impacts of logging rain forest on fauna using Before-After-Control-Impact-Pairs experimental data. Brooks et al. (2000) applied Bayesian methods to the Cormack–Jolly–Seber (CJS) alike models to estimate survival and recovery/capture probabilities from band-return and capture-recapture data. Link and Barker (2005) presented a hierarchical extension of the (CJS) models for estimating the correlation between biological birth rate and survival rate. Lee et al. (2006) analyzed data collected from a two-release experiment. A comprehensive review on open population models via a Bayesian approach can be found in Buckland et al. (2007).

Recently Huggins (2007) proposed a martingale estimation equation approach to obtain the size of an open population, where the population size is assumed to be a deterministic function of certain covariates and parameters, and linear models are used to estimate the effects of the covariates on the population size. This approach uses full individual capture histories in its bootstrap estimation of the standard deviations of parameter estimators. Sometimes individuals may be difficult to distinguish from each other in a capture-recapture experiment. For example, in the study of goose population (Macinnes, 1966; Malecki et al., 1981) the data are usually collected by sighting the neck or foot bands of the geese. An individual's capture history is hence not available in this case. Instead, summary data, i.e., the numbers of marked and unmarked animals

---

* Corresponding author. Tel.: +61 3 8344 4899; fax: +61 3 8344 4599.
E-mail addresses: qguoqi@unimelb.edu.au, g.qian@ms.unimelb.edu.au (G. Qian).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.009

captured on each occasion, are often obtainable. In this paper we aim to develop a population model and an estimation approach for this type of data. We assume the population under study is open, and has suffered from an environmental catastrophe during the data collection period. Further, suppose that the population size before the catastrophe has a stationary distribution with mean N, the catastrophe has an effect to reduce the mean population size by $N\delta$ for some $0 \leq \delta \leq 1$, and the population size eventually returns to the stationary distribution after the catastrophe. To model a stationary population it is reasonable to assume that the mean number of births and immigrants is the same as the mean number of deaths and emigrants.

In Section 2 we develop a dynamic population model to delineate the evolutions of population sizes and capture-recapture data. The model consists of latent variables, observed capture-recapture variables and the parameters of interest. In Section 3 we develop MCMC methods for computing the posterior distributions. In computing the conditional distribution of the latent variables, because the involved distributions are so convoluted that feasible proposal densities with sizeable acceptance probabilities cannot be constructed, we develop a sequential importance sampling (SIS) scheme (cf. Doucet et al. (2001), Liu (2001) and Chapter 14 of Robert and Casella (2004)). Simulation studies are conducted and reported in Section 4. In Section 5 the model and algorithm are applied to capture-recapture data on mountain pygmy possums (*Burramys Parvus*) whose habitat was affected by a bushfire during the data collection period. Further discussions are given in Section 6.

2. The model and notation

Consider a capture-recapture experiment conducted at times $t_1, t_2, \ldots, t_I$, with t_0 denoting an initial time point. The number of occasions, I, is usually small.

Assume that the captured animals are marked and released; the marks are recognizable on later occasions; individuals have the same probability p of being captured, and p does not vary with capture history; the mean departure rate from the population is the same as the mean arrival rate β, apart from the time period in which the catastrophic event occurred. At t_0 the initial population size N_0 has a $\text{Pois}(N)$ distribution. The population size $\{N_i\ (i = 0, 1, \ldots, c-1)\}$ before the catastrophe is a stationary process with the Poisson stationary distribution $\text{Pois}(N)$. The catastrophe occurs during the time period (t_{c-1}, t_c), and has the immediate effect of reducing the population size by an extra mean departure rate δ.

At a given time t_i the population consists of two independent components, that is, $N_i = Y_i + V_i$, where Y_i represents new births or new arrivals from other regions, and V_i represents survivors from time t_{i-1}. The N_i, Y_i and V_i are all unobservable and thus latent variables. Denote by θ the model parameters (N, β, δ, p). We assume the distributions $Y_i|\theta \sim \text{Pois}(N\beta)$ for all i and $V_i|(\theta, N_{i-1}) \sim \text{Bin}(N_{i-1}, 1-\beta)$ for all non-catastrophic time periods i; and $V_c|(\theta, N_{c-1}) \sim \text{Bin}(N_{c-1}, 1-\beta-\delta)$ with $(0 \leq \beta, \delta \leq 1, 0 \leq \beta + \delta \leq 1)$ for the catastrophic time period. Then it can be shown that $N_i|\theta \sim \text{Pois}(N)$ for $i \leq c-1$ and $N_i|\theta \sim \text{Pois}\left(N[1-\delta(1-\beta)^{i-c}]\right)$ for $i \geq c$.

Suppose we only observe the numbers of new captures, u_i, and the numbers of recaptures, m_i, on every occasion t_i. Denoted by D is the observed data. To embed the observed capture results $\{u_i, m_i\}$ into the population dynamic we further decompose the survivors V_i into latent variables W_i and (M_{i1}, M_{i2}) according to capture and survival status. The W_i is the number of individuals in population at t_{i-1} who had not been captured by t_{i-1} and had survived to t_i. While M_{i1} is the number of individuals who were first captured during $(t_{i-2}, t_{i-1}]$ and had survived to t_i $(2 \leq i \leq I)$, M_{i2} is the number of individuals who had been captured by t_{i-2} and had survived to t_i $(3 \leq i \leq I)$. And we define $M_{i1} = 0$ for $(i \leq 1)$ and $M_{i2} = 0$ for $(i \leq 2)$. All the latent variables are now denoted as $\tau = (N_0, Y_1, \ldots, Y_I, W_1, \ldots, W_I, M_{21}, \ldots, M_{I1}, M_{22}, \ldots, M_{I2})$. While Y_i was modelled with a $\text{Pois}(N\beta)$ distribution, we formulate W_i and (M_{i1}, M_{i2}) with the following probability distributions

$$\begin{aligned}
&W_1|(\theta, N_0) \sim \text{Bin}(N_0, 1-\beta)\\
&W_i|(\theta, Y_{i-1}, W_{i-1}, u_{i-1}) \sim \text{Bin}\,(Y_{i-1} + W_{i-1} - u_{i-1}, 1-\beta-\delta 1_c(i)) \quad \text{for } i > 1\\
&M_{i1}|(\theta, u_{i-1}) \sim \text{Bin}\,(u_{i-1}, 1-\beta-\delta 1_c(i))\\
&M_{i2}|(\theta, M_{(i-1)1}, M_{(i-1)2}) \sim \text{Bin}\left(M_{(i-1)1} + M_{(i-1)2}, 1-\beta-\delta 1_c(i)\right)
\end{aligned}$$

and model the observed data with probability distributions

$$\begin{aligned}
&u_i|(\theta, Y_i, W_i) \sim \text{Bin}(Y_i + W_i, p)\\
&m_i|(\theta, M_{i1}, M_{i2}) \sim \text{Bin}\,(M_{i1} + M_{i2}, p)
\end{aligned}$$

where $1_c(i)$ is the indicator function for the catastrophe that takes 1 if $i = c$ and zero otherwise. For convenience the notation is also listed in Appendix, and the model is diagramed in Fig. 1.

There are three components in our model: the observed data D, the latent variables τ, and the parameter θ. The goal is to infer θ from D. In the Bayes approach, the extra unknown may sometimes be estimated together with model parameters and the objective becomes to estimate $f(\theta, \tau|D)$, from which the marginal posterior distribution $f(\theta|D)$ can be obtained. Often, the difficulty in an inference is due to the lack of a link between the observed and the quantity to be inferred. If this missing part can be explicitly described in a model, the complexity of the inference may be reduced to pieces that can be handled more easily. This is the so called missing-data formulation. See Little and Rubin (1987) and Gelman and Meng (2004) for special discussion on this topic. The modelling approach we used in the above is such a formulation, in which the unobserved population composition sequence, τ, was explicitly modelled.