Bayesian Case Studies
Practical 4

Capture-Recapture. Metropolis-within-Gibbs. Model misspecification.

Robin Ryder

Aim: Sampling from the posterior in the capture-recapture model.

Reference: *Bayesian Essentials with R* (Marin & Robert), chapter 5.

Data: European dipper dataset, from `http://bit.ly/MASH-BCS`.

1 Basic capture-recapture

We consider a population with unknown size N; we wish to infer N. To this end, we capture n_1 individuals from the population, mark them, then perform a second capture of n_2 individuals: of these, we observe that m_2 are marked. We denote by p the probability for an individual to be captured at each step of the data collection procedure:

$$n_1 \sim \mathcal{B}(N, p) \qquad m_2|n_1 \sim \mathcal{B}(n_1, p) \qquad n_2 - m_2 \sim B(N - n_1, p).$$

We choose an independent prior $\pi(N, p) = \pi(N)\pi(p)$ where $\pi(p)$ is the $\mathcal{U}([0, 1])$ distribution. We assume that the population is unchanged between the two steps of the procedure.

We consider a dataset on a population of birds called European dippers (*Cinclus cinclus*) in Southern France in the 1980s. The initial observations are $n_1 = 22$, $n_2 = 60$, $m_2 = 11$.

1. Write the likelihood of (p, N). Deduce the conditional posterior distribution of $p|N, n_1, n_2, m_2$.
2. Find a sufficient statistic of dimension 2.
3. We choose a hyperparameter S and use the prior $\pi(N) = \frac{1}{S}\mathbb{I}_{\{N \leq S\}}$. Calculate the marginal posterior distribution of N and compute the mean and variance of the posterior $\pi(N|n_1, n_2, m_2)$. Give a 95% confidence interval.
4. Examine the influence of the hyperparameter S.
5. Extend this model to the case with 3 samplings.
6. Perform in-model validation: simulate synthetic data from the model for values N and p of your choosing, and verify that you are able to estimate to estimate N correctly.

7. One year lapses between each sampling. Think about possible misspecifications of the model. How can we handle them?

2 Open population

We now consider three samplings, and we no longer assume that the population is unchanged: an unknown number r_1 of marked individuals are removed (eg they die) between the first and second samplings; an unknown number r_2 of marked individuals are removed between the second and third samplings. Each individual dies with unknown probability q. We observe three quantities: the number of captured individuals at sampling 1 ($n_1 = 22$), the number of marked individuals recaptured at sampling 2 ($m_2 = 11$) and at sampling 3 ($m_3 = 6$).

8. Write the corresponding model. We choose an improper prior $\pi(N) \propto 1/N$.

9. Compute the conditional distributions for a Gibbs' sampler. Are they easy to sample from?

10. (Optional) Given a (possibly unnormalized) density g which is difficult to sample from, a sample Y can be simulated from g using the Accept-Reject algorithm: find a density f and a constant M such that $\forall x, g(x) \leq Mf(x)$ then:

    a) Generate $X \sim f$ and $U \sim \mathcal{U}([0,1])$.

    b) If $u \leq g(x)/Mf(x)$ then accept $Y = x$; else repeat.

    Use this method to simulate from the posterior distribution for the open population model. You may want to use a suitable chosen binomial distribution for f.

11. An alternative when one of the conditionals is complex in a Gibbs sampler is to replace the simulation from the conditional by a single Metropolis-Hastings step. Implement this Metropolis-within-Gibbs method, and compare its efficiency with the previous question.

12. Think about possible misspecifications of the model. How can we handle them?

3 Arnason-Schwarz model

The Arnason-Schwarz model allows experimenters to register the zone where an individual was recorded. It is useful to understand migrations. For an individual i, we have two description vectors:

- $\boldsymbol{z}_i = (z_{it}, t = 1 \dots T)$ describes the location of the individual at each time t;
- $\boldsymbol{x}_i = (x_{it}, t = 1 \dots T)$, a binary vector describing whether the individual was captured at each time t.

The variable z_{it} can take the values $1 \dots R$ where R is the number of locations, or the value † if the individual is dead. The parameters of interest are now

- the capture probabilities $p_r = P[x_{it} = 1 | z_{it} = r]$ for $r = 1 \dots R$ (we assume $p(\dagger) = 0$);
- the migration rates $q_{rs} = P[z_{i,t+1} = s | z_{it} = r]$.

13. Discuss appropriate prior distributions for the parameters.

14. The value of z_{it} is not observed when $x_{it} = 0$. Does this hinder your inference procedure?

15. Write a Gibbs sampler to infer the population size for the European dipper dataset.