STAT-221: Pset 4

Kevin Kuate Fodouop
Harvard University

Abstract

In this homework we use Markov Chain Monte Carlo (MCMC) methods to lead inference on the unknown number of experiments N from binomial observations. Different potential priors are examined theoretically, and MCMC analysis is performed to reproduce Raftery's results (1988), in terms of posterior distribution for N, pointwise and interval estimation.

We are given n observations $Y_1, Y_2, ..., Y_n$ drawn from a distribution

$$Y_i \sim Bin(N, \theta)$$

with N and θ unknown parameters. With $N \sim Poiss(\mu)$, we define $\lambda = \theta\mu$, and specify distributions on (λ, θ) in a Bayesian framework. λ will be conveninent to draw inference on, as it is the mean of the observed data. It is also more reasonable to assume a prior independence between λ and θ than between λ and μ (as a prior on λ is more informative).

question 1.1 We define our prior $p(\lambda, \theta) \propto \lambda^{-1}$. Hence λ and θ are independent a priori, $p(\lambda) \propto \lambda^{-1}$ and $p(\theta) \propto 1$. We compute the induced prior on (N, θ).

$$\begin{aligned}
p(N,\theta) &= p(N|\theta)p(\theta) \\
&= \int_0^{+\infty} p(N|\theta,\lambda)p(\theta)p(\lambda)d\lambda \\
&\propto \int_0^{+\infty} \frac{\left(\frac{\lambda}{\theta}\right)^N}{N!} e^{\frac{\lambda}{\theta}} \frac{1}{\lambda} d\lambda \\
&= \frac{1}{\theta}\frac{1}{N!}\int_0^{+\infty} \left(\frac{\lambda}{\theta}\right)^{N-1} e^{\frac{\lambda}{\theta}} d\lambda \\
&= \frac{1}{\theta}\frac{1}{N!}\int_0^{+\infty} u^{N-1}e^u du \times \theta \\
&= \frac{1}{N!}\Gamma(N) \\
&= \frac{1}{N}
\end{aligned}$$

Hence we have a prior distribution $p(N, \theta) \propto \frac{1}{N}$. This is the standard vague prior for N (inverse prior), multiplied by a uniform prior on θ. It puts higher weights on small values of N, and is an improper prior.

question 1.2 $p(\lambda, \theta)$ is an improper prior as $\int_0^{+\infty} \frac{1}{\lambda} d\lambda = [log(\lambda)]_0^{+\infty} = +\infty$.

question 1.3 $Y_i|\theta, \mu \sim Poiss(\theta\mu$ (chicken and egg problem). This is derived by using

$$p(Y_i|\theta,\mu) = \sum_{N=0}^{+\infty} p(Y_i|\theta,N)p(N|\mu)dN$$

and simplifying the obtained expression using the exponential series decomposition. So we have

$$p(Y_i|\theta,\mu) = \frac{1}{Y_i!}\,(\theta\mu)^{Y_i}\,e^{\theta\mu}$$

So we have a log-likelihood

$$\mathcal{L}(\theta,\mu) = Y_i log(\theta\mu) - \theta\mu - log(Y_i!)$$

And second derivatives

$$\frac{\partial^2\mathcal{L}(\theta,\mu)}{\partial\theta^2} = -\frac{Y_i}{\theta^2}$$
$$\frac{\partial^2\mathcal{L}(\theta,\mu)}{\partial\mu^2} = -\frac{Y_i}{\mu^2}$$
$$\frac{\partial^2\mathcal{L}(\theta,\mu)}{\partial\theta\partial\mu} = -1$$

And the determinant of the Fisher Information matrix is

$$\begin{aligned} det(I(\theta,\mu)) &= -\frac{\mu}{\theta} \times -\frac{\theta}{\mu} - 1 \\ &= 0 \end{aligned}$$

Hence our information matrix is not invertible. This is due to the fact that our model is not identifiable (we only can get inference on the product $\theta\mu$ from the data). After transformation the Fisher information matrix will still be singular, so that $p(\lambda,\theta)$ is not non-informative in Jeffrey's sense. Rafltery's prior is informative in the way it privileges small values of λ. Hence for same values of $S = N\theta$ (which generates same value of the log-likelihood), our prior categorizes higher N as less likely.

question 1.4 We try several implementation of MCMC, diagnostic plots of which are documented in appendix:

1. `mcmc.mh2step`: Samples first λ from its posterior distribution, sample θ from its posterior and derives μ from it. From μ sample $N \sim Poiss(\mu)$, truncated to be more tha y_{max}.

2. `mcmc.mh_Sexp`: Sample $E[S] = N\theta$ as a scaled beta using last value of N, and sample N using a truncated geomtric.

3. `mcmc.mdhir`: Samples N from truncated geometric, θ from its posterior beta distribution.

4. `mcmc.md_Sexplogonly`: Samples the same way as `mcmc.mh_Sexp`, but only use the log-likelihood ratio in the acceptance ratio (does not take into account the non-symmetry of the sampling scheme). Even though it seemed theoretically unsound, it was tried and implemented as it gave way better results than the other algorithms.

Our first algorithm drives high autocorrelation, and fails the halfwidth mean test of convergence (on both the `impala` and `waterbuck` data set). It also has quite value of rubin- gelman test (from 1.2 to 2.3), and is not very stable. Acceptance rate is around 30%.

The second one led to divergence in N for ununderstood reasons. There might be an issue with the acceptance rate definition, but it was not resolved. This algorithm had to be abandonned for this reason.

The third algorithm also leads to high correlation and high values of gelman test (values higher than 2), and has a small acceptance rate at around 10%.

The fourth variant of the second gives way better results. Acceptance rate is around 60% (good), it has highly decreasing autocorrelation, passes Heidelberg and Halfwidth Mean tests, and produces a rubin-gelman score of 1.01 and 1.00. Our only concern using this method is in the seemingly wrongly defined acceptance ratio.

No stable and seemingly theoretically valid algorithm could be set up. We chose to implement the first (high acceptance rate and valid acceptance rate formula) and the fourth algorithm (gives the best results) on Odyssey to produce the plots.

question 1.5 We derive the marginal posterior distributino for N. We have the complete posterior log-likelihood, for $N \geqslant y_{max}$ (otherwise null log-likelihood)

$$
\begin{aligned}
p(N,\theta|y) &\propto p(Y|N,\theta) \times p(N,\theta) \\
&= \prod_{i=1}^{n} C_N^{y_i} \theta^{y_i} (1-\theta)^{N-y_i} \times \frac{1}{N} \\
&= \frac{1}{N} \left[\prod_{i=1}^{n} C_N^{y_i} \right] \theta^S (1-\theta)^{nN-S}
\end{aligned}
$$

With $S = \sum_i y_i$. Integrating out θ (and as the prior on θ is uniform), for $N \geqslant y_{max}$

$$
\begin{aligned}
p(N|y) &\propto \frac{1}{N} \left[\prod_{i=1}^{n} C_N^{y_i} \right] \int_0^1 \theta^S (1-\theta)^{nN-S} d\theta \\
&= \frac{1}{N} \left[\prod_{i=1}^{n} C_N^{y_i} \right] \mathcal{B}(1+S, 1+nN-S) \quad (1)
\end{aligned}
$$

We compute the normalizing constants of (1) for both datasets in table 1,

$$
K = \left(\sum_{N=y_{max}}^{+\infty} \frac{1}{N} \left[\prod_{i=1}^{n} C_N^{y_i} \right] \mathcal{B}(1+S, 1+nN-S) \right)^{-1}
$$

We estimate the infinite sum involved in K. We implemented importance sampling, but the term in the sum becomes hard to compute for $N > 7 \times 10^3$ (product of an infinite and a zero value) where it has already converged reasonably towards zero. Hence we use an exact summation for those first values to approximate the constants. We check the distribution obtained using the computed constant indeed integrates to 1.

Table 1: *Normalizing constant of $p(N|y)$ defined in (1), for* `impala` *and* `waterbuck` *data sets*

Data Set	Constant	Distribution integral
impala	6267314	1
waterbuck	525394839	1

question 1.6 We compute the posterior probability $P(N > 100|y)$ using the normalized distribution computed in *1.5*, and compare to the estimates obtained with our 10 MCMC draws per data set. Results are displayed in table 2.

Table 2: *$P(N > 100|y)$ analytically and from MCMC.*

Data Set	Analytically	MCMC
impala	0.33	1
waterbuck	0.92	1

A Figures

A.1 Posterior log-likelihood plots

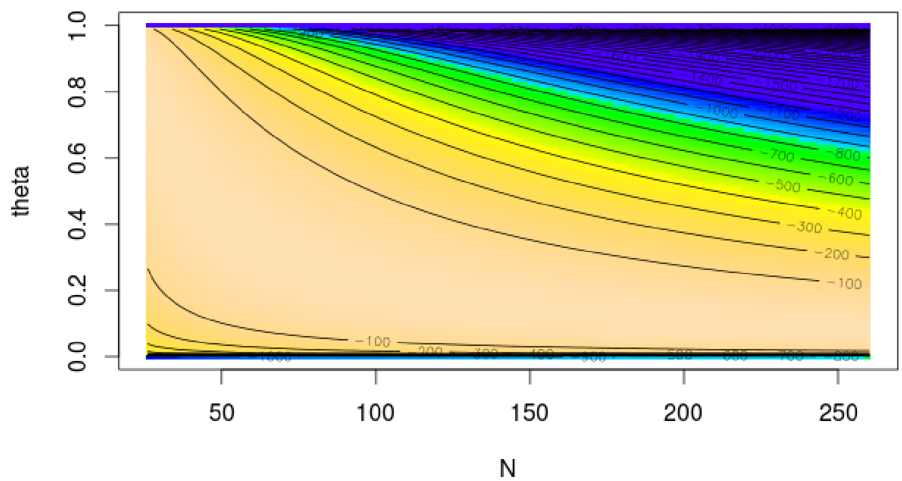

Figure 1: *Posterior log-likelihood plot of (N, θ) for the* `impala` *dataset.*

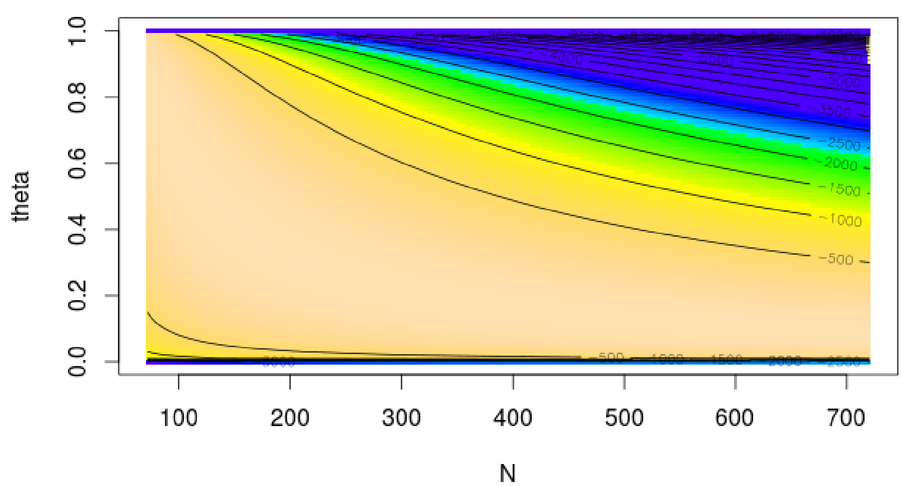

Figure 2: *Posterior log-likelihood plot of (N, θ) for the* `waterbuck` *dataset.*

A.2 MCMC algorithms diagnostics

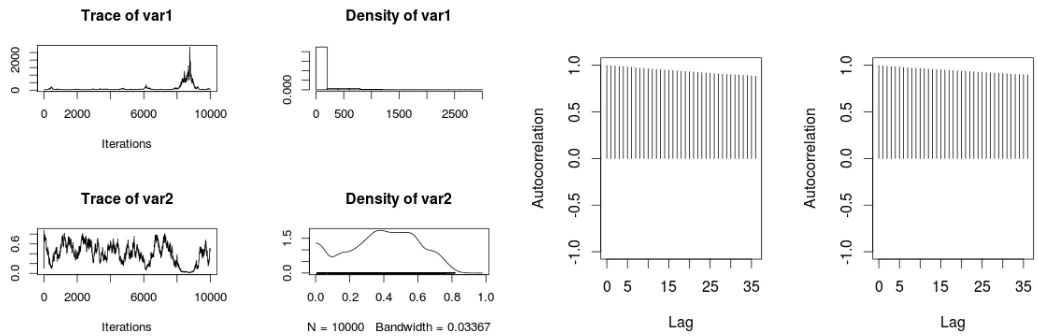

Figure 3: *Trace and autocorrelation plots for* `mcmc.mh2step` *(*`impala` *data set). Autocorrelation remains very high even after consequent lag.*

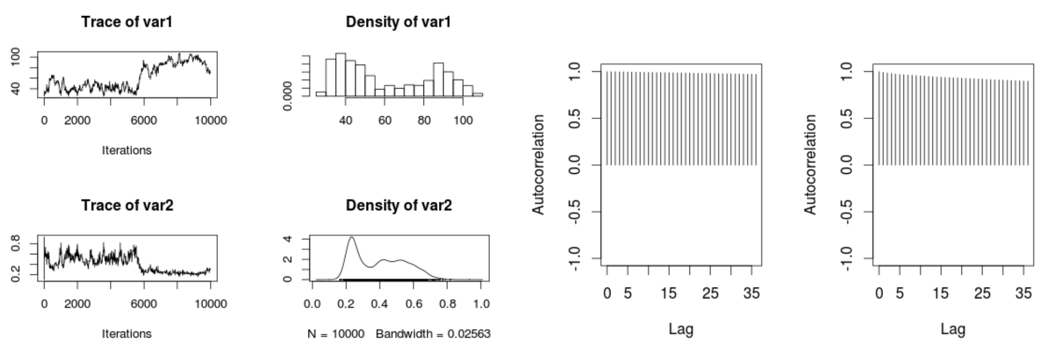

Figure 4: *Trace and autocorrelation plots for* `mcmc.mhdir` *(*`impala` *data set). Autocorrelation is even higher.*

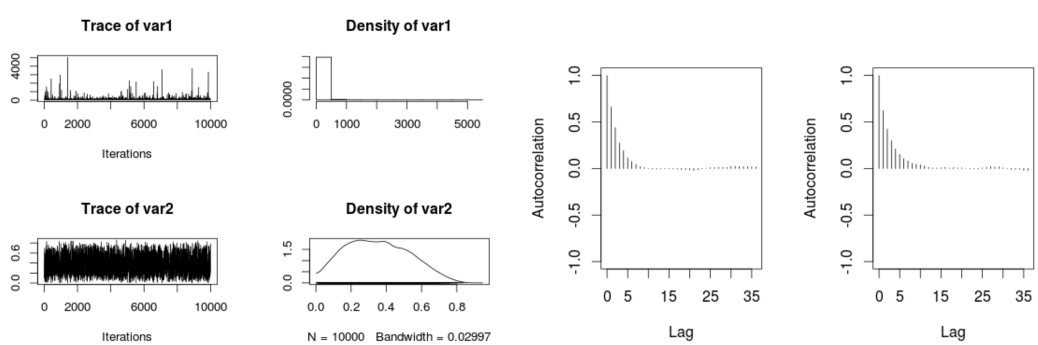

Figure 5: *Trace and autocorrelation plots for* `mcmc.mh_Sexplongly` *(*`impala` *data set). Autocorrelation decreases fast, and trace plot shows great exploration of the parameter space.*

A.3 Posterior contour plots