

# 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  $Y_1, Y_2, \dots, Y_n$  drawn from a distribution

$$Y_i \sim \text{Bin}(N, \theta)$$

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

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

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

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  $\theta$ . 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*  $Y_i|\theta, \mu \sim \text{Poiss}(\theta\mu)$  (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

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

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. Raftery's prior is informative in the way it privileges small values of  $\lambda$ . 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  $\lambda$  from its posterior distribution, sample  $\theta$  from its posterior and derives  $\mu$  from it. From  $\mu$  sample  $N \sim \text{Pois}(\mu)$ , truncated to be more than  $y_{\max}$ .
2. `mcmc.mh_Sexp`: Sample  $E[S] = N\theta$  as a scaled beta using last value of  $N$ , and sample  $N$  using a truncated geometric.
3. `mcmc.mdhir`: Samples  $N$  from truncated geometric,  $\theta$  from its posterior beta distribution.

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 understood reasons. There might be an issue with the acceptance rate definition, but it was not resolved. This algorithm had to be abandoned 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%.

No stable algorithm could be set up. To privilege high acceptance rate and better results on the tests (as well as better visual posterior distribution), we chose to implement the first algorithm 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  $N \geq y_{max}$  (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  $\theta$  (and as the prior on  $\theta$  is uniform), for  $N \geq y_{max}$

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

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

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

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.

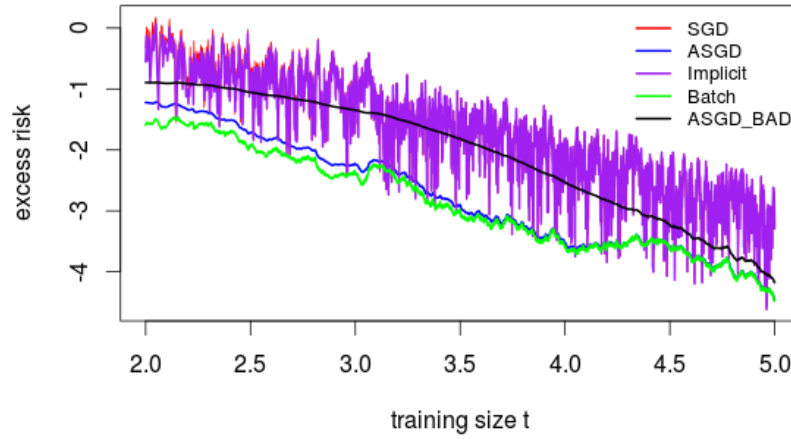
**Table 1:** Normalizing constant of  $p(N | y)$  defined in (1), for *impala* and *waterbuck* data sets

Data Set	Constant	Distribution integral
impala	6267314	1
waterbuck	525394839	1

*question 1.6*

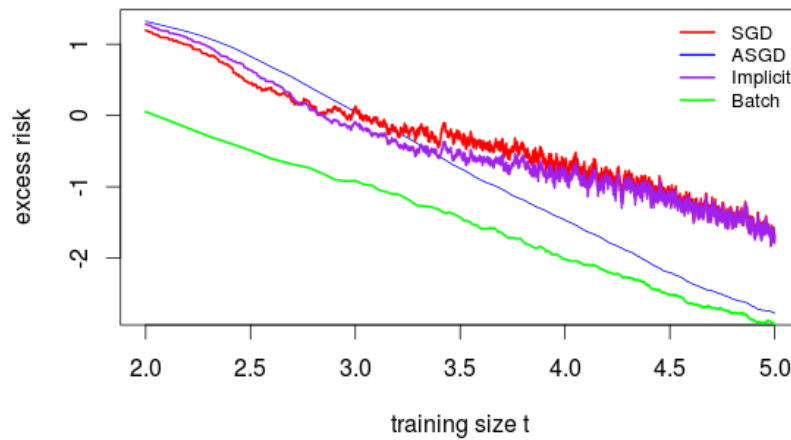
## A Figures

### A.1 SGD methods on Normal model

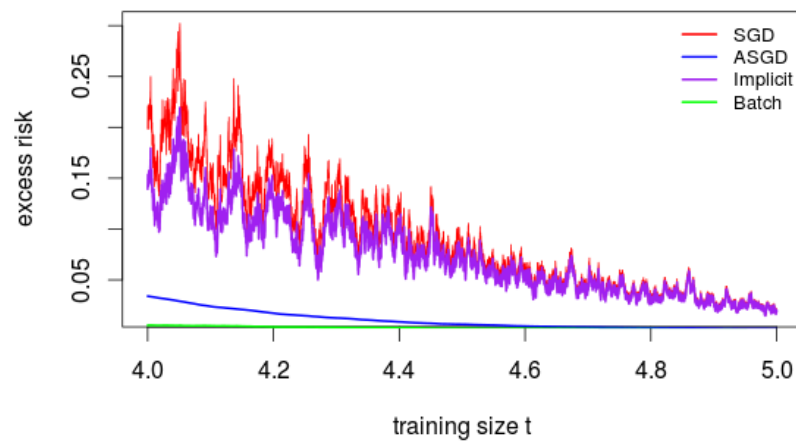


**Figure 1:** Excess risk for different update methods. SGD and Implicit methods exposes a very unstable behaviors. General relative behaviors of methods are nonetheless reproduced compared to Xu. Excess risk and training size  $t$  on  $\log_{10}$  scale.

### A.2 SGD methods on regression



**Figure 2:** Excess risk for different update methods. Like in Xu, SGD starts with better performance than ASGD but eventually performs worse for high training size (Implicit does slightly better than SGD). Batch is doing better overall. Excess risk and training size  $t$  on  $\log_{10}$  scale.



**Figure 3:** *Excess risk zoom on training size superior to  $10^4$ .*