problem set no. 4 — due tuesday 11/04

problem 1. partially reproduce Markow chain Monte Carlo analysis in Raftery (1988); implement an alternative inference strategy based on analytic calculations and/or numerical integration to verify the MCMC results.

problem background. consider observations $Y_1, \ldots, Y_n$ sampled from a

$$\text{Binomial } (N, \theta)$$

where both N and θ are unknown (constant) parameters. this is the model. a full Bayes strategy to make inference on the parameters would require positing a (prior) distribution on (N, θ). specifying a convenient family of distributions on (N, θ) is difficult, however, partly because of the discreteness of N.

Raftery (1988) posits a Poisson distribution with unknown mean μ for the parameter N. then, instead of specifying a distribution on (N, θ), Raftery defines $\lambda \equiv \theta \cdot \mu$ and specifies a distribution on (λ, θ). to understand why this is useful, note that the distribution on N effectively adds a source of variability into the model. the new (constant) parameters are (θ, μ). the mean of the observations is

$$\mathbb{E}[Y_i \mid \theta, \mu] = \theta \cdot \mu,$$

in other words, the mean of the observations is λ.

overview. you will have to develop two alternative inference strategies for (N, θ); implement your algorithms in `R`; run the MCMC analysis and marginal posterior calculations on Odyssey; report results in latex and provide code. to grade your solution, we will compile your latex and run your code.

question 1.1. (15 point) a suggested distribution is $p(\lambda, \theta) \propto \lambda^{-1}$. what is the rationale behind this choice? to answer, transform and determine the induced distribution on (N, θ). comment on your findings.

question 1.2. (5 point) is $p(\lambda, \theta)$ a proper distribution? that is, does it integrate to 1?

question 1.3. (5 point) is $p(\lambda, \theta)$ non-informative in the sense of Jeffreys? recall that Jeffreys' suggested non-informative prior for $p(\text{observations} \mid \text{constants}) \propto \sqrt{\det I(\text{constants})}$. derive $p(Y_i \mid \theta, \mu)$, compute its Fisher information matrix $I(\theta, \mu)$ and transform.

question 1.4. (40 point) develop an MCMC algorithm to sample from $p(N, \theta \mid Y_{1:n})$, using the prior suggested in (1.1). perform the analysis using both the waterbuck and the impala data used in Raftery (1988), where $n = 5$, using 10 chains. display a scatterplot

of the posterior simulations of (N, θ) with posterior contours overlaid for the 10 chains in each of the two data sets. (20 plots total). additional MCMC diagnostic plots are encouraged but not required. for `R` users: to plot posterior contours you should consider the functions `kde2d` (in `library(MASS)`) to perform bivariate kernel-density estimation and `contour(...,add=TRUE)` to add contour lines to an existing plot (you may need to change colour and line-width [e.g., `col='red',lwd=2.5`] to make the contours visible). the `coda` library also contains excellent resources for plotting, diagnostics and summaries of MCMC chains.

question 1.5. (30 point) derive the marginal posterior distribution for N, again using the prior in (1.1). since you will later need to compute probabilities with this marginal distribution, you will need to find the normalizing constant for the density (this can be done numerically for the impala and waterbuck datasets if you wish). report the normalizing constant for both datasets.

question 1.6. (5 point) compare the posterior probability that $N > 100$ you obtain with MCMC versus the probability obtained analytically (and possibly with numerical integration), in the 10 experiments (chains) for the impala and for the waterbuck data sets separately. compute the mean and standard deviation of the 10 posterior probability estimates you obtained with MCMC on the two data sets. comment on your findings in general and with respect to the answer obtained in (1.5).

* * *

what to submit. submit 1 zip file. the zip file should contain: the `R` and `slurm` scripts to run the MCMC analysis on Odyssey (two files); the `R` and `Slurm` scripts to calculate marginal probabilities in (1.5) (two file); the 20 figures described in 1.4 (20 files); a latex (or word) document with your answers to the remaining questions, including a brief description of your MCMC design and implementation on Odyssey, and a subset of the relevant figures.

please read the following instructions carefully and make sure your submission conforms to the guidelines. this is a computing exercise; if your code does not run it's a problem. any submitted homeworks that do not conform to the specifications will lose points.

email a ZIP file named as `myfasusername_ps4.ZIP` to stat221.harvard.fall2014@gmail.com (where `myfasusername` is replaced by your actual FAS username), and make sure to include "stat221" in the subject. The ZIP file should contain two `R` scripts: named as `myfasusername_mcmc.R` and `myfasusername_marginal.R` (again, where `myfasusername` is replaced by your actual FAS username) and your Slurm scripts. your scripts must run on Odyssey and produce 20 plots as described above, **assuming all input/output files are in the same folder**. the ZIP file should contain a PDF file named as `myfasusername_ps4.pdf`, as well as the TEX or DOC file with the solution, named as `myfasusername_ps4.tex` or

`myfasusername_ps4.doc`. if you have any questions about code formatting and organization then please ask Panos. happy coding!