Bayesian Computation: A Pedagogical View

Jim Albert
Emeritus Professor, BGSU

February 26, 2021

How Do I Like Retirement?

Actually ...

- Haven't done much traveling (haven't been to the beach)
- Playing a lot of tennis
- Keeping up my "Exploring Baseball with R" blog
- Some statistical stuff (research, etc.)

Outline

Introduction: Why Bayes?

Teaching Bayes

Bayesian Computational Methods

Wrap-Up

The Plan

- Look back at my efforts in Bayesian pedagogy
- Historical review of Bayesian computation
- Use a Bayesian multilevel model to illustrate computational methods
- Available software
- What does the future of Bayes (and the teaching of Bayes) look like?

Two Views of Statistical Inference

Frequentist:

- 95% confidence intervals, 5% tests of significance
- Evaluate methods by their long-run performance in repeated sampling

Bayesian:

- Use subjective probability to express uncertainty about unknowns
- 90% credible interval – 90% is the probability that the unknown parameter is contained in the computed interval

Why Bayes?

- Statistics is “Learning from Data”
- Bayesian paradigm provides an attractive way of implementing inference
- Express beliefs about parameter using a Prior
- Observe data and update one’s beliefs by Bayes’ rule (Posterior)

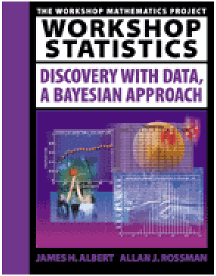
What's Wrong with Frequentist Thinking?

- Certainly, Frequentist methods are very useful
- But the logic of Frequentist inference is not natural
- Have to think about repeated sampling
- Easy to misinterpret confidence intervals and tests of significance

My Teaching of Bayes

- Taught a masters-level course MATH 6480 many times
- Inspired by Don Berry and 1960's texts, I introduced Bayes for intro-stats (MATH 1150)
- Just completed a "Probability and Bayesian Modeling" text (with Monika Hu) assuming calculus - MATH 4410-4420

Baby Bayes (with Allan Rossman)



- A "workshop style" Bayesian text for intro-stats (MATH 1150)
- Currently free to download.
- Students did a Bayesian sample survey project

Student Survey Project

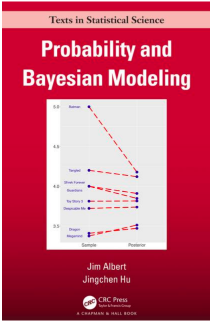
- Consider p, the proportion of BGSU students who sleep at least 7 hours a night
- Place a prior on values $p = 0, 0.1, 0.2, ..., 0.9, 1$
- Take a survey of 20 students.
- Update opinion by Bayes' rule.

Bayesian Calculator for a Proportion

	Prior	Likelihood	Product	Posterior
p=0	0.028	0	0	0
p=.1	0.028	119	3	0.0005
p=.2	0.111	3288	365	0.0567
p=.3	0.278	9902	2751	0.4269
p=.4	0.278	10000	2778	0.4312
p=.5	0.111	4457	495	0.0769
p=.6	0.056	878	49	0.0076
p=.7	0.028	61	2	0.0003
p=.8	0.028	1	0	0
p=.9	0.028	0	0	0
p=1	0.028	0	0	0
		Successes	Failures	
	DATA	7	13	

UPDATE CLEAR

Math-Stat Bayes (with Monika Hu)



- Target audience is undergraduates with a calculus background
- Can be used in MATH 4410-4420 *Probability and Mathematical Statistics*

Learning Outcomes in Math/Stat Bayes

- How to construct priors
- How are the prior and data information combined
- Applications of prediction
- Simulation-based inference
- Bayes in popular methods (regression and multilevel modeling)

Question: How to Compute?

- Which Bayesian computational method should be recommended?
- Which method will help in achieving the Bayesian learning goals?
- Different methods are available.
- Is a "black-box" Bayesian tool desirable?

Bayesian Computation Challenge

- Bayesian model assumes $y \sim f(y|\theta)$ and the vector θ has a prior $g(\theta)$
- By Bayes' rule, the posterior of θ is

$$g(\theta|y) \propto f(y|\theta)g(\theta)$$

- Challenge: How to summarize this multivariate posterior probability distribution?
- A big numerical integration problem

Computational Methods

- Grid (discrete) approach
- Normal approximation
- Conjugate Priors
- MCMC - Metropolis Sampling / Gibbs Sampling
- MCMC - Hamiltonian Sampling
- New Methods - Approximate Bayesian Computation (ABC)

Example: A Bayesian Multilevel Model

Data: Collect number of hits (y) and number of at-bats (n) for a group of N baseball players

- $y_1, \ldots, y_N, y_i \sim Binomial(n_i, p_i)$
- $p_1, \ldots, p_N \sim Beta(K\eta, K(1-\eta))$
- $\eta \sim Beta(a, b)$, $\log K \sim Logistic(logn, 1)$

Focus on Second-Stage Parameters

- Have $N+2$ parameters $p_1, \ldots, p_N, K, \eta$
- Interested in marginal posterior of (η, K) :

$$g(\eta, K|y) \propto g(\eta, K) \prod_{j=1}^{N} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j)}{B(K\eta, K(1-\eta))}$$

- Need some computational method to summarize this posterior.

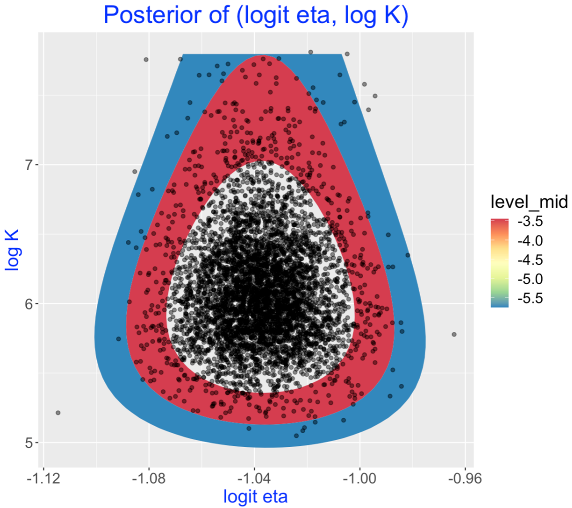
Grid Computation

- Set up a grid of values for each parameter
- Can use quadrature rules to choose grid efficiently
- Effort in 1980's to write adaptive quadrature algorithms for arbitrary Bayesian models (Adrian Smith and Bayes 4)
- Curse of dimensionality – number of posterior calculations increases exponentially

Grid Computation for Example

- By trial and error, choose a 50 by 50 grid that covers posterior
- Graph posterior by contour plot
- Can simulate values of parameters from grid

Grid Computation & Simulation



Grid Computation - Pros and Cons

- Easy to implement and visualize
- What if parameters are correlated?
- Only works for problems with a small number of parameters

1960's: Conjugate Priors

- Suppose have a sample from exponential family (normal, binomial, Poisson, etc)
- For each distribution, there exists a "conjugate" prior so that both prior and posterior have same functional form
- Posterior and predictive distributions are available

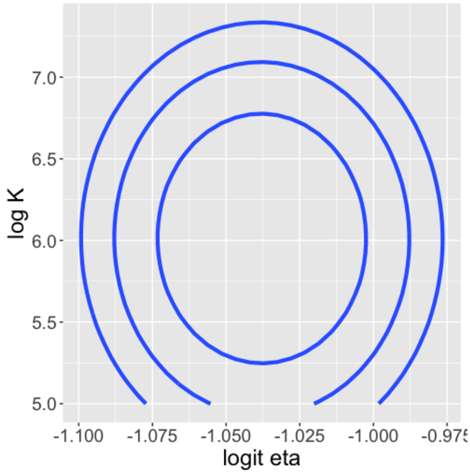
Nice Aspects of Conjugate Analyses

- Simple expressions for posterior mean and variance
- Easy to see how prior information and data get combined
- Conjugate analyses can be building blocks for multilevel models (such as our binomial/beta example)
- Can summarize posterior and predictive distributions by simulation

Normal Approximation

- Old idea – used by Laplace, but generalized in 1980's
- Expand logarithm of posterior in Taylor series about mode $\hat{\theta}$
- Approximate posterior by a $N(\hat{\theta}, V)$ distribution
- Implement approximation by Newton Raphson

Example: Normal Approximation



Nice Aspects of Normal Approximation

- General approach – can be used for arbitrary prior and sampling density
- Computationally quick
- Can use nice properties of multivariate normal
- Can use simulation methodology to do inference
- But ...

1990: MCMC

- Gelfand and Smith, JASA, 1990
- Idea: create a Markov Chain that will converge to posterior distribution
- Simulate from Markov Chain to get (approximate) posterior sample
- Gibbs sampling and Metropolis/Hastings were the early MCMC algorithms

Metropolis Algorithm

Random walk algorithm – suppose the current value is $\theta = \theta^c$. One step of algorithm:

1. Propose a value $\theta^p = \theta^c + scale \times Z$
2. Compute an acceptance probability P depending on the ratio $g(\theta^p)/g(\theta^c)$
3. With probability P move to proposal value θ^p, otherwise stay at current value θ^c

Nice Aspects of Metropolis Algorithm

- Simple algorithm
- Easy to program
- Motivates discussion of MCMC diagnostics such as acceptance rates, trace plots and autocorrelation plots

Bayesian MCMC Software

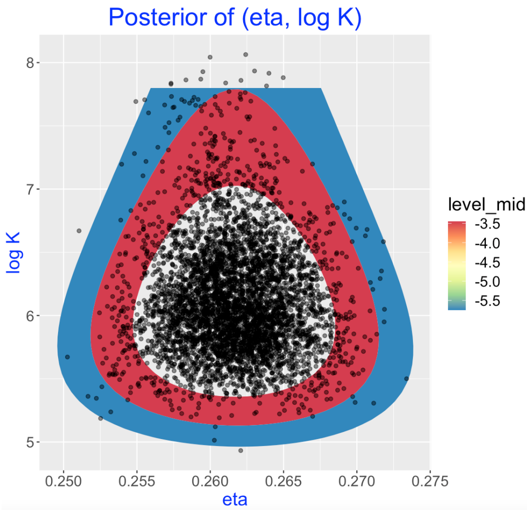
- Effort to create general-purpose software that incorporate conjugate analyses and basic MCMC algorithms (Gibbs sampling and Metropolis)
- BUGS (Bayesian Utilization of Gibbs Sampling) and JAGS (Just Another Gibbs Sampler)
- User writes a model script
- Single R function command does the sampling

Example: Metropolis - JAGS

Write a model script:

```
model {
      for (i in 1:N){
        y[i] ~ dbin(p[i], n[i])
      }
      for (i in 1:N){
        p[i] ~ dbeta(a, b)
      }
      a <- mu * eta
      b <- (1 - mu) * eta
      mu ~ dbeta(mua, mub)
      eta <- exp(logeta)
      logeta ~ dlogis(logn, 1)
    }
```

Metropolis Sampling with JAGS



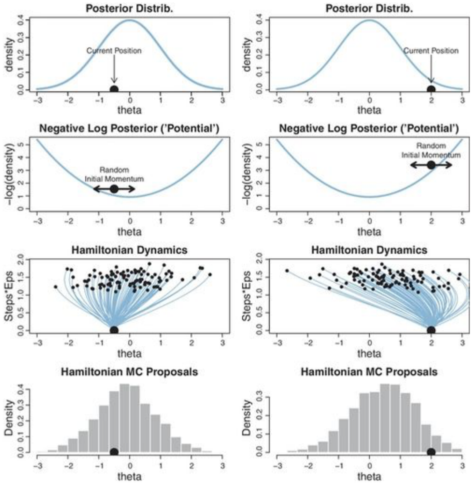
Limitations of Metropolis

- Efficient Metropolis may only accept 25% of the time.
- Can be slow in sampling of regions of high posterior content
- Metropolis doesn't work well for high-dimensional problems such as multilevel modeling
- Need a better (more efficient) method

Hamiltonian Monte Carlo (HMC)

- Employs a guided proposal random walk
- Use gradient of log posterior to direct Markov chain towards regions of highest posterior density
- A well-tuned HMC chain will accept proposals at much higher rate
- Requires the log posterior and the gradient function

Nice Illustration of HMC from *Doing Bayesian Data Analysis*



Stan Software

- Stan is well-documented software for implementing a version of HMC for a wide variety of Bayesian models
- Stan interfaces with many programming languages (R, python, MATLAB, etc)
- There are R packages that provide high-level functions for popular Bayesian regression and multilevel models

Using Stan

The brms package will implement Stan for a variety of regression and multilevel models:

```
fit <- brm(data = DeathHeartAttackManhattan,
           family = binomial,
           Deaths | trials(Cases) ~ 1 + (1 | Hospital),
           refresh = 0)
```

21st Century: A Second Computational Revolution

- Original simulation methods are limited
- High dimensional problems where one has a large number of unknowns (parameters)
- Problems where one cannot express the sampling density in closed form
- Variety of new computing methods being developed
- "Approximate Bayes" Methods

Approximate Bayesian Calculation (ABC)

Want to approximate the posterior density $g(\theta|y)$ in situations where expression for sampling density is not available.

1. Simulate values of θ and data y – compute summary statistic $T(y)$.
2. Condition on values of $T(y)$ that are close to observed (T_{obs}).
3. Corresponding values of θ are from posterior density

Example: ABC

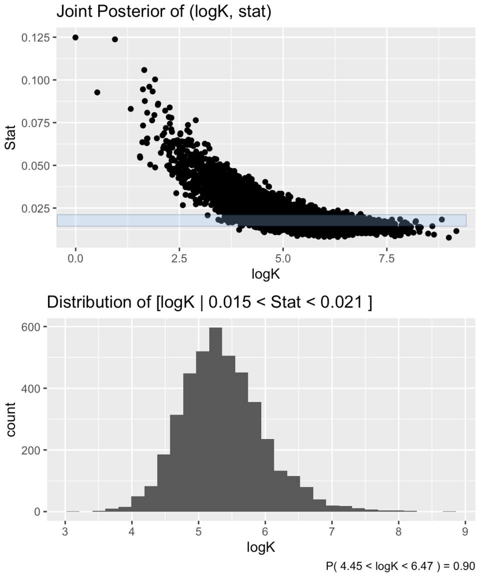
Simulate from Bayesian multilevel model:

- simulate η from a Beta(100, 275)
- simulate $\log K$ from a logistic(5, 1)
- simulate $p_1, ..., p_N$ where $p_j \sim Beta(K\eta, K(1-\eta))$
- simulate $y_1, ..., y_N$ where $y_j \sim binom(n_j, p_j)$
- compute $SD = sd(y/n)$

Example: ABC

- Repeat algorithm for 10,000 iterations – collect pairs $\{(\log K, SD)\}$
- Compute observed SD, SD_{obs}
- Interested in marginal posterior of $\log K$
- Collect values of $\log K$ where $|SD - SD_{obs}| < \epsilon$

Posterior of $\log K$ by ABC



What Methods are in *Probability and Bayesian Modeling*?

- Conjugate priors (proportion and mean)
- Gibbs sampling and Metropolis algorithms
- JAGS for regression and multilevel models
- Documentation if the instructor wishes to use Stan (HMC sampling)

It's a Great Time to Be a Bayesian

- Wide range of Bayesian computational methods available
- Use of Bayesian methods is spreading to many applied disciplines
- One of the best current Bayesian texts *Rethinking Statistics* is written by an anthropologist
- Eventually, Bayesian ideas will be taught to undergraduates

References

Albert and Hu (2020), "Bayesian Computing in the Undergraduate Statistics Curriculum," *Journal of Statistics and Data Science Education*.

This article is part of the Bayesian Cluster section of *Journal of Statistics and Data Science Education*, volume 28, issue 3 (2020)

References

- Martin, Frazier and Robert (2020), "Computing Bayes: Bayesian Computation from 1763 to the 21st Century" (nice survey of Bayesian computation methods)
- van de Shoot, et al (2020), "Bayesian Statistics and Modeling," *Nature Reviews* (nice overview of how one implements Bayesian modeling for applications)