Bayesian Inference and Markov Chain Monte Carlo

Bob Carpenter

Columbia University

Machine Learing Summer School 2015, Sydney

Warmup Exercise I

Sample Variation

Repeated i.i.d. Trials

- Suppose we repeatedly generate a random outcome from among several potential outcomes
- Suppose the outcome chances are the same each time
  - i.e., outcomes are independent and identically distributed (i.i.d.)
- For example, spin a spinner, such as one from *Family Cricket*.

Image source: http://replaycricket.com/2010/10/29/family-cricket/

Repeated i.i.d. Binary Trials

- Suppose the outcome is binary and assigned to 0 or 1; e.g.,
  - 20% chance of outcome 1: *ball in play*
  - 80% chance of outcome 0: *ball* not *in play*
- Consider different numbers of bowls delivered.
- How will proportion of successes in sample differ?

Simulating i.i.d. Binary Trials

- R Code: `rbinom(10, N, 0.2) / N`
  - **10 bowls** (10% to 50% success rate)

    2 3 5 2 4 1 2 2 1 1
  - **100 bowls** (16% to 26% success rate)

    26 18 23 17 21 16 21 15 21 26
  - **1000 bowls** (18% to 22% success rate)

    181 212 175 213 216 179 223 198 188 194
  - **10,000 bowls** (19.3% to 20.3% success rate)

    2029 1955 1981 1980 2001 2014 1931 1982 1989 2020

Simple Point Estimation

- Estimate chance of success θ by proportion of successes:

$$\theta^* = \frac{\text{successes}}{\text{attempts}}$$

- Simulation shows accuracy depends on the amount of data.

Confidence Intervals via Sim

- $P\%$ **confidence interval:** interval in which $P\%$ of the estimates are expected to fall.
- Simulation computes intervals to any accuracy.
- Just simulate, sort, and inspect the central empirical interval.

Example Interval Calculation

- To calculate 95% confidence interval of estimate based on 1000 samples:

```
> sims <- rbinom(10000, 1000, 0.2) / 1000
> sorted_sims <- sort(sims)
> sorted_sims[c(250, 9750)]
[1] 0.176 0.225
```

- The 95% confidence interval is thus $(0.175, 0.225)$

Estimator Bias

- **Bias:** expected difference of estimate from true value
- Continuing previous example

```
> sims <- rbinom(10000, 1000, 0.2) / 1000
> sorted_sims <- sort(sims)
```

- Take central point to get expected estimate from estimator

```
> sort(sims)[5000]
[1] 0.2
```

- Central value of 0.2 shows this estimator is *unbiased*

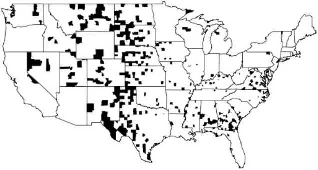
Simple Point Estimation (cont.)

- **Central Limit Theorem:** *expected* error in θ^* goes down as
$$\frac{1}{\sqrt{N}}$$
- A decimal place of accuracy requires $100\times$ more samples.
- The width of confidence intervals shrinks at the same rate.
- Can also use theory to show this estimator is unbiased.

Pop Quiz! Cancer Clusters

- Why do lowest and highest cancer clusters look so similar?

Lowest kidney cancer death rates



Highest kidney cancer death rates

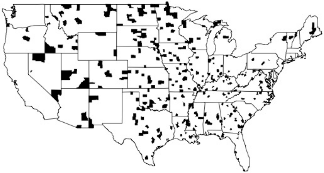


Image from Gelman et al., *Bayesian Data Analysis, 3rd Edition* (2013)

Pop Quiz Answer

- Hint: mix earlier simulations of repeated i.i.d. trials with 20% success and sort:

1/10	1/10	1/10	15/100	16/100
17/100	175/1000	179/1000	18/100	181/1000
188/1000	194/1000	198/1000	2/10	2/10
2/10	2/10	21/100	21/100	21/100
212/1000	213/1000	216/1000	223/1000	23/100
26/100	26/100	3/10	4/10	5/10

- More variation in observed rates with smaller sample sizes

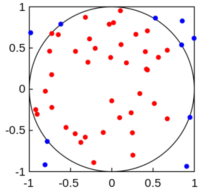
- *Answer*: High cancer and low cancer counties are small populations

Warmup Exercise II

Monte Carlo Integration

Monte Carlo Calculation of π

- Computing $\pi = 3.14\ldots$ via simulation is *the* textbook application of Monte Carlo methods.
- Generate points uniformly at random within the square
- Calculate proportion within circle ($x^2 + y^2 < 1$) and multiply by square's area (4) to produce the area of the circle.
- This area is π (radius is 1, so area is $\pi r^2 = \pi$)



Plot by Mysid Yoderj courtesy of Wikipedia.

Monte Carlo Calculation of π (cont.)

- R code to calcuate π with Monte Carlo simulation:

```
> x <- runif(1e6,-1,1)
> y <- runif(1e6,-1,1)

> prop_in_circle <- sum(x^2 + y^2 < 1) / 1e6

> 4 * prop_in_circle
[1] 3.144032
```

Accuracy of Monte Carlo

- Monte Carlo is *not* an approximation!
- It can be made exact to within any ϵ
- Monte Carlo draws are i.i.d. by definition
- Central limit theorem: expected error decreases at rate of

$$\frac{1}{\sqrt{N}}$$

- 3 decimal places of accuracy with sample size 1e6
- Need $100\times$ larger sample for each digit of accuracy

General Monte Carlo Integration

- MC can calculate arbitrary definite integrals,
$$\int_a^b f(x)\,dx$$
- Let d upper bound $f(x)$ in (a,b); tightness determines computational efficiency
- Then generate random points uniformly in the rectangle bounded by (a,b) and $(0,d)$
- Multiply proportion of draws (x,y) where $y < f(x)$ by area of rectangle, $d \times (b-a)$.
- Can be generalized to multiple dimensions in obvious way

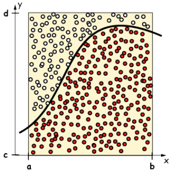


Image courtesy of Jeff Cruzan, http://www.drcruzan.com/NumericalIntegration.html

Warmup Exercise II

Maximum Likelihood Estimation

Observations, Counterfactuals, and Random Variables

- Assume we observe data $y = y_1, \ldots, y_N$
- Statistical modeling assumes even though y is observed, the values could have been different *ceteris paribus*
- John Stuart Mill first characterized this **counterfactual** nature of statistical modeling in:

  A System of Logic, Ratiocinative and Inductive (1843)
- In modern (Kolmogorov-ian) language, we say y is a **random variable**

Likelihood Functions

- A **likelihood function** is a probability function (density, mass, or mixed)

$$p(y|\theta, x),$$

- where θ is a vector of **parameters**,
- x is some fixed data (e.g., regression predictors or "features"),
- considered as a function $\mathcal{L}(\theta)$ of θ for fixed x and y.

Maximum Likelihood Estimation

- The statistical inference problem is to estimate parameters θ given observations y.

- Maximum likelihood estimation (MLE) chooses the estimate θ^* that maximizes the likelihood function, i.e.,

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta p(y|\theta, x)$$

- This function of $\mathcal{L}$ and y (x) is called an *estimator*

Example of MLE

- The frequency-based estimate

$$\theta^* = \frac{1}{N}\sum_{n=1}^{N} y_n,$$

is the observed rate of “success” (outcome 1) observations.

- This is the MLE for the model

$$p(y|\theta) \;=\; \prod_{n=1}^{N} p(y_n|\theta) \;=\; \prod_{n=1}^{N} \text{Bernoulli}(y_n|\theta)$$

where for $u \in \{0, 1\}$,

$$\text{Bernoulli}(u|\theta) = \begin{cases} \theta & \text{if } u = 1 \\ 1 - \theta & \text{if } u = 0 \end{cases}$$

Example of MLE (cont.)

- The first step $p(y|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$ is the i.i.d. (or exchangeability) modeling *assumption*.
- The second step $p(y_n|\theta) = \text{Bernoulli}(y_n|\theta)$ is a modeling *assumption*.
- The frequency estimate is the MLE, because
  - derivative is zero (indicating min or max),

$$\mathcal{L}'_y(\theta^*) = 0,$$

  - and second derivative is negative (indicating max),

$$\mathcal{L}''_y(\theta^*) < 0.$$

MLEs can be Dangerous!

- Recall the cancer cluster example
- Accuracy is low with small counts
- What we need are hierarchical models (stay tuned)

Part I
Bayesian Data Analysis

Bayesian Data Analysis

- "By Bayesian data analysis, we mean practical methods for making inferences from data using probability models for quantities we observe and about which we wish to learn."
- "The essential characteristic of Bayesian methods is their **explict use of probability for quantifying uncertainty** in inferences based on statistical analysis."

Gelman et al., *Bayesian Data Analysis*, 3rd edition, 2013

Bayesian Mechanics

1. Set up full probability model
    - for all observable & unobservable quantities
    - consistent w. problem knowledge & data collection
2. Condition on observed data
    - caclulate posterior probability of unobserved quantities conditional on observed quantities
3. Evaluate
    - model fit
    - implications of posterior

Ibid.

Notation for Basic Quantities

- Basic Quantities
  - y: observed data
  - θ: parameters (and other unobserved quantities)
  - x: constants, predictors for conditional (aka "discriminative") models
- Basic Predictive Quantities
  - $\tilde{y}$: unknown, potentially observable quantities
  - $\tilde{x}$: predictors for unknown quantities

Naming Conventions

- **Joint**: $p(y, \theta)$
- **Sampling / Likelihood**: $p(y|\theta)$
  - Sampling is function of y with θ fixed (prob function)
  - Likelihood is function of θ with y fixed (*not* prob function)
- **Prior**: $p(\theta)$
- **Posterior**: $p(\theta|y)$
- **Data Marginal (Evidence)**: $p(y)$
- **Posterior Predictive**: $p(\tilde{y}|y)$

Bayes's Rule for Posterior

$$\begin{aligned}
p(\theta|y) &= \frac{p(y,\theta)}{p(y)} && \text{[def of conditional]} \\
&= \frac{p(y|\theta)\,p(\theta)}{p(y)} && \text{[chain rule]} \\
&= \frac{p(y|\theta)\,p(\theta)}{\int_\Theta p(y,\theta')\,d\theta'} && \text{[law of total prob]} \\
&= \frac{p(y|\theta)\,p(\theta)}{\int_\Theta p(y|\theta')\,p(\theta')\,d\theta'} && \text{[chain rule]}
\end{aligned}$$

- *Inversion:* Final result depends only on sampling distribution (likelihood) $p(y|\theta)$ and prior $p(\theta)$

Bayes's Rule up to Proportion

- If data y is fixed, then

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)\,p(\theta)}{p(y)} \\ &\propto p(y|\theta)\,p(\theta) \\ &= p(y,\theta) \end{aligned}$$

- Posterior proportional to likelihood times prior
- Equivalently, posterior proportional to joint

Posterior Predictive Distribution

- Predict new data $\tilde{y}$ based on observed data y
- Marginalize out parameter from posterior

$$p(\tilde{y}|y) \; = \; \int_{\Theta} p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta.$$

  averaging predictions $p(\tilde{y}|\theta)$, weighted by posterior $p(\theta|y)$

- $\Theta = \{\theta \mid p(\theta|y) > 0\}$ is the support of $p(\theta|y)$
- For discrete parameters θ,

$$p(\tilde{y}|y) \; = \; \sum_{\theta \in \Theta} p(\tilde{y}|\theta)\, p(\theta|y).$$

- Can mix continuous and discrete (integral as shorthand)

Event Probabilities

- Recall that an event A is a collection of outcomes
- Suppose event A is determined by indicator on parameters

$$f(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A \end{cases}$$

- e.g., $f(\theta) = \theta_1 > \theta_2$ for $\mathrm{Pr}[\theta_1 > \theta_2 \mid y]$
- Bayesian event probabilities calculate posterior mass

$$\mathrm{Pr}[A] = \int_{\Theta} f(\theta)\, p(\theta|y)\, d\theta.$$

- Not frequentist, because involves parameter probabilities

Example I

Male Birth Ratio

Laplace Turns the Crank

- Laplace's data on live births in Paris from 1745–1770:

sex	*live births*
female	241 945
male	251 527

- Question 1 (Event Probability)
  Is a boy more likely to be born than a girl?
- Question 2 (Estimate)
  What is the birth rate of boys vs. girls?
- Bayes formulated the basic binomial model
- Laplace solved the integral (Bayes couldn't)

Binomial Distribution

- Binomial distribution is number of successes y in N i.i.d. Bernoulli trials with chance of success θ
- If $y_1, \ldots, y_N \sim \text{Bernoulli}(\theta)$,

  then $(y_1 + \cdots + y_N) \sim \text{Binomial}(N, \theta)$
- The analytic form is

$$\text{Binomial}(y|N, \theta) = \binom{N}{y} \theta^y (1 - \theta)^{N-y}$$

where the binomial coefficient normalizes for permutations of which $y_n = 1$,

$$\binom{N}{y} = \frac{N!}{y!\,(N-y)!}$$

Bayes's Binomial Model

- Data
  - y: total number of male live births (241,945)
  - N : total number of live births (493,472)
- Parameter
  - $\theta \in (0, 1)$: proportion of male live births
- Likelihood

$$p(y|N, \theta) \;=\; \text{Binomial}(y|N, \theta) \;=\; \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- Prior

$$p(\theta) \;=\; \text{Uniform}(\theta|0, 1) \;=\; 1$$

Beta Distribution

- Required for analytic posterior of Bayes's model
- For parameters $\alpha, \beta > 0$ and $\theta \in (0,1)$,

$$\mathsf{Beta}(\theta|\alpha,\beta) \;=\; \frac{1}{\mathsf{B}(\alpha,\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

- Euler's beta function is used to normalize,

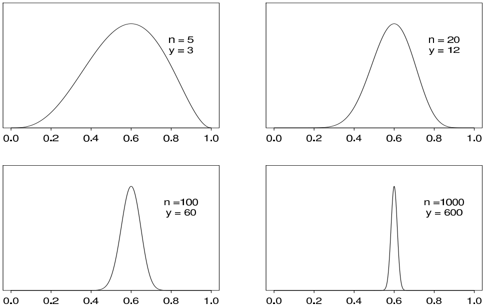
$$\mathsf{B}(\alpha,\beta) \;=\; \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du \;=\; \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

  where $\Gamma()$ is continuous generalization of factorial

- Note: $\mathsf{Beta}(\theta|1,1) = \mathsf{Uniform}(\theta|0,1)$

Beta Distribution — Examples

- Unnormalized posterior density assuming uniform prior and y successes out of n trials (all with mean 0.6).



Gelman et al. (2013) *Bayesian Data Analysis*, 3rd Edition.

Laplace Turns the Crank

- Given Bayes's general formula for the posterior

$$p(\theta|y,N) \;=\; \frac{\text{Binomial}(y|N,\theta)\,\text{Uniform}(\theta|0,1)}{\int_{\Theta}\text{Binomial}(y|N,\theta')\,p(\theta')d\theta'}$$

- Laplace used Euler's beta function (B) to solve the integral required for normalizationm,

$$p(\theta|y,N) \;=\; \text{Beta}(\theta|y+1,N-y+1)$$

Estimation

- Posterior is $distroBeta(\theta|1 + 241\,945,\ 1 + 251\,527)$
- Posterior mean:

$$\frac{1 + 241\,945}{1 + 241\,945 + 1 + 251\,527} \approx 0.4902913$$

- Maximum likelihood estimate same as posterior mode (because of uniform prior)

$$\frac{241\,945}{241\,945 + 251\,527} \approx 0.4902912$$

- As number of observations $\to \infty$, MLE approaches posterior mean

Event Probability Inference

- What is probability that a male live birth is more likely than a female live birth?

$$
\begin{aligned}
\mathsf{Pr}[\theta > 0.5] &= \int_{\Theta} \mathsf{I}[\theta > 0.5]\, p(\theta|y,N)\, d\theta \\
&= \int_{0.5}^{1} p(\theta|y,N)\, d\theta \\
&= 1 - F_{\theta|y,N}(0.5) \\
&\approx 1^{-42}
\end{aligned}
$$

- $F_{\theta|y,N}$ is posterior cumulative distribution function (cdf),
- $\mathsf{I}[\phi] = 1$ if condition ϕ is true and 0 otherwise

Bayesian Fisher Exact Test

- Suppose we have the following data on handedness

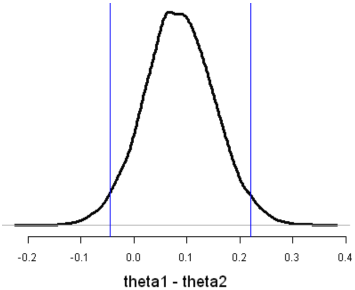
	sinister	*dexter*	TOTAL
male	9 (y_1)	43	52 (N_1)
female	4 (y_2)	44	48 (N_2)

- Assume likelihoods Binomial$(y_k | n_k, \theta_k)$ and uniform priors
- Are men more likely to be lefthanded?

$$\begin{aligned} \text{Pr}[\theta_1 > \theta_2 \mid y, N] &= \int_{\Theta \times \Theta} p(\theta_1 | y_1, N_1)\, p(\theta_2 | y_2, N_2)\, d\theta_1\, d\theta_2 \\ &\approx 0.91 \end{aligned}$$

Visualizing Posterior Difference

- Plot of posterior difference, $p(\theta_1 - \theta_2 \mid y, N)$



- Vertical bars: central 95% posterior interval $(-0.05, 0.22)$

Part III
Conjugate Priors

Conjugate Priors

- Before MCMC techniques became practical, Bayesian analysis mostly involved conjugate priors
- Still widely used because analytic solutions are more efficient than MCMC
- Family $\mathcal{F}$ is a conjugate prior for family $\mathcal{G}$ if
  - prior in $\mathcal{F}$ and
  - likelihood in $\mathcal{G}$,
  - entails posterior in $\mathcal{F}$

Beta is Conjugate to Binomial

- Prior: $p(\theta|\alpha,\beta) = \text{Beta}(\theta|\alpha,\beta)$
- Likelihood: $p(y|N,\theta) = \text{Binomial}(y|N,\theta)$
- Posterior:

$$
\begin{aligned}
p(\theta|y,N,\alpha,\beta) &\propto p(\theta|\alpha,\beta)\, p(y|N,\theta) \\
&= \text{Beta}(\theta|\alpha,\beta)\, \text{Binomial}(y|N,\theta) \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)} \theta^{\alpha-1}\,(1-\theta)^{\beta-1} \binom{N}{y} \theta^{y}\,(1-\theta)^{N-y} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)} \theta^{y+\alpha-1}\,(1-\theta)^{N-y+\beta-1} \\
&\propto \text{Beta}(\theta|\alpha+y,\ \beta+N-y)
\end{aligned}
$$

Chaining Updates

- Start with prior Beta$(\theta|\alpha, \beta)$
- Receive binomial data in K statges $(y_1, N_1), \ldots, (y_K, N_K)$
- After (y_1, N_1), posterior is Beta$(\theta|\alpha + y_1,\ \beta + N_1 - y_1)$
- Use as prior for (y_2, N_2), with posterior

  Beta$(\theta|\alpha + y_1 + y_2,\ \ \beta + (N_1 - y_1) + (N_2 - y_2))$
- Lather, rinse, repeat, until final posterior

  Beta$(\theta|\alpha + y_1 + \cdots + y_K,\ \ \beta + (N_1 + \cdots + N_K) - (y_1 + \cdots + y_K))$
- Same result as if we'd updated with combined data

  $(y_1 + \cdots + y_K,\ \ N_1 + \cdots + N_K)$

Part II
(Un-)Bayesian Point Estimation

MAP Estimator

- For a Bayesian model $p(y, \theta) = p(y|\theta)\, p(\theta)$, the max a posteriori (MAP) estimate maximizes the posterior,

$$\begin{aligned}
\theta^* &= \arg\max_\theta \, p(\theta|y) \\
&= \arg\max_\theta \, \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= \arg\max_\theta \, p(y|\theta)p(\theta). \\
&= \arg\max_\theta \log p(y|\theta) + \log p(\theta).
\end{aligned}$$

- *not* Bayesian because it doesn't integrate over uncertainty
- *not* frequentist because of distributions over parameters

MAP and the MLE

- MAP estimate reduces to the MLE if the prior is uniform, i.e.,

$$p(\theta) = c$$

because

$$\begin{aligned}
\theta^* &= \arg\max_\theta \, p(y|\theta)\, p(\theta) \\
&= \arg\max_\theta \, p(y|\theta)\, c \\
&= \arg\max_\theta \, p(y|\theta).
\end{aligned}$$

Penalized Maximum Likelihood

- The MAP estimate can be made palatable to frequentists via philosophical sleight of hand
- Treat the negative log prior $-\log p(\theta)$ as a "penalty"
- e.g., a Normal$(\theta|\mu,\sigma)$ prior becomes a penalty function

$$\log\sigma + \frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2$$

- Maximize sum of log likelihood and negative penalty
- Called a "penalized maximum likelihood estimate," but quantitatively equal to MAP

Proper Bayesian Point Estimates

- Choose estimate to minimize some loss function
- To minimize expected squared error (L2 loss), $\mathbb{E}[(\theta - \theta')^2]$, use the posterior mean

$$\hat{\theta} \;=\; \arg\min_{\theta'} \mathbb{E}[(\theta - \theta')^2] \;=\; \int_{\Theta} \theta \times p(\theta|y)\, d\theta.$$

- To minimize expected absolute error (L1 loss), $\mathbb{E}[|\theta - \theta'|]$, use the posterior median.
- Other loss (utility) functions possible, the study of which falls under decision theory

Point Estimates for Inference?

- Common in machine learning to generate a point estimate θ^* then use it for inference, $p(\tilde{y}|\theta^*)$
- This is *defective* because it

Underestimates Uncertainty

- To properly estimate uncertainty, apply full Bayes
- If you don't, Dutch book can be made against you (i.e., if you use your strategy to bet, I can beat you in the long run using full Bayes)

Part III
Philosophical Interlude

Exchangeability

- Roughly, an exchangeable probability function is such that for a sequence of random variables $y = y_1, \ldots, y_N$,

$$p(y) = p(\pi(y))$$

  for every N-permutation π (i.e, a one-to-one mapping of $\{1, \ldots, N\}$)

- i.i.d. implies exchangeability, but not vice-versa

Exchangeability Assumptions

- Models almost always make some kind of exchangeability assumption
- Typically when other knowledge is not available
  - e.g., treat voters as conditionally i.i.d. given their age, sex, income, education level, religous affiliation, and state of residence
  - But voters have many more properties (hair color, height, profession, employment status, marital status, car ownership, gun ownership, etc.)
  - Missing predictors introduce additional error (on top of measurement error)

de Finetti's Theorem

- de Finetti's Theorem: Given some background variables, every exchangeable sequence is conditionally i.i.d.

- So if our i.i.d. assumptions are invalid, condition on more predictors

Bayesian vs. Frequentist

- *Everyone*: Model data y as “random”
- *Everyone*: Parameters have single, true (but unknown) value
- *Everyone*: Admit Bayes’s rule of probability
- *Bayesians Only*: Model parameters θ as “random”
- *Frequentists Only*: Probabilities are long-run frequencies of observables, which excludes parameters (unobservable)
- *Bayesians Only*: Allow probabilities conditioned on parameters

Random Parameters: Doxastic or Epistemic?

- Bayesians treat distributions over parameters as epistemic (i.e., about knowledge)
- They do *not* treat them as being doxastic (i.e., about beliefs)
- Priors encode our knowledge before seeing the data
- Posteriors encode our knowledge after seeing the data
- Bayes's rule provides the way to update our knowledge
- People like to pretend models are ontological (i.e., about what exists)

Arbitrariness: Priors vs. Likelihood

- Bayesian analyses often criticized subjective (arbitrary)
- Choosing priors is no more arbitrary than choosing a likelihood function (or an exchangeability assumption)
- As George Box famously wrote (1987),

All models are wrong, but some are useful.

- This is true for frequentists as well as Bayesians

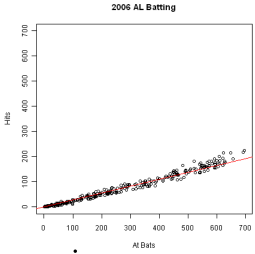
Part IV
Hierarchical Models

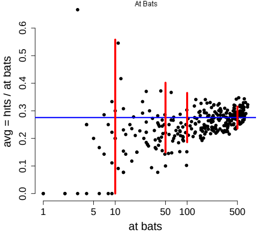
Baseball At-Bats

- For example, consider baseball batting ability.
  - Baseball is sort of like cricket, but with round bats, a one-way field, stationary “bowlers”, four bases, short games, and no tied games.
- Batters have a number of “at-bats” in a season, out of which they get a number of “hits” (hits are a good thing)
- Nobody with higher than 40% success rate since 1950s.
- No player (excluding “bowlers”) bats much less than 20%.
- Same approach applies to hospital pediatric surgery complications (a BUGS example), reviews on Yelp, test scores in multiple classrooms, . . .

Baseball Data

- Hits versus at bats for the 2006 American League season
- Not much variation in ability!
- Ignore fact that players with more at-bats tend to be better (need GLM for that...)





Pooling Data

- How do we estimate the ability of a player who we observe getting 6 hits in 10 at-bats? Or 0 hits in 5 at-bats? Estimates of 60% or 0% are absurd!
- Same logic applies to players with 152 hits in 537 at bats.
- No pooling: estimate each player separately
- Complete pooling: estimate all players together (assume no difference in abilities)
- Partial pooling: somewhere in the middle

Hierarchical Models

- Hierarchical models are principled way of determining how much pooling to apply.
- Assume population ability levels and pull estimates toward the population mean based on how much variation in the population
  - low variance population: more pooling
  - high variance population: less pooling
- In limit
  - as variance goes to 0, get complete pooling
  - as variance goes to ∞, get no pooling

Hierarchical Batting Ability

- Instead of fixed priors, estimate priors along with other parameters
- Still only uses data once for a single model fit
- Data: y_n, B_n: hits, at-bats for player n
- Parameters: θ_n: ability for player n
- Hyperparameters: α, β: population mean and variance
- Hyperpriors: fixed priors on α and β (hardcoded)

Hierarchical Batting Model (cont.)

$$
\begin{aligned}
y_n &\sim \text{Binomial}(B_n, \theta_n) \\
\theta_n &\sim \text{Beta}(\alpha, \beta) \\
\frac{\alpha}{\alpha + \beta} &\sim \text{Uniform}(0, 1) \\
(\alpha + \beta) &\sim \text{Pareto}(1.5)
\end{aligned}
$$

- Sampling notation syntactic sugar for:

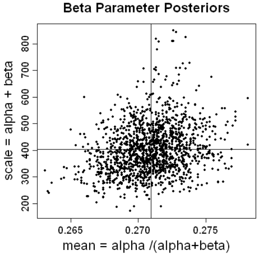
  $p(y, \theta, \alpha, \beta) \;=\; \text{Pareto}(\alpha+\beta|1.5) \prod_{n=1}^{N} \left(\text{Binomial}(y_n|B_n, \theta_n)\ \text{Beta}(\theta_n|\alpha, \beta)\right)$

- Pareto provides power law: $\text{Pareto}(u|\alpha) \propto \frac{\alpha}{u^{\alpha+1}}$

- Should use more informative hyperpriors!

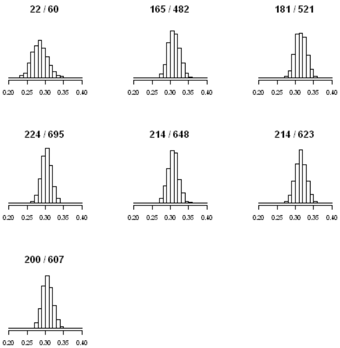
Hierarchical Prior Posterior

- Draws from posterior (crosshairs at posterior mean)
- Prior population mean: 0.271
- Prior population scale: 400
- Together yield prior std dev of 0.022
- Mean better estimated than scale (typical)



Posterior Ability (High Avg Players)

- Histogram of posterior draws for high-average players
- 22/60 = 0.367
- Note uncertainty grows with lower at-bats



Multiple Comparisons

- Who has the highest ability (based on this data)?
- Probabilty player n is best is

Average	*At-Bats*	Pr[best]
.347	521	0.12
.343	623	0.11
.342	482	0.08
.330	648	0.04
.330	607	0.04
.367	60	0.02
.322	695	0.02

- No clear winner—sample size matters.
- In last game (of 162), Mauer (Minnesota) edged out Jeter (NY)

"Naive Bayes" Four Ways

- Joint Distribution $p(\pi, \phi, z, w)$, defined by:
  - $\pi \sim \text{Dirichlet}(\alpha)$ (topic prevalence)
  - $\phi_k \sim \text{Dirichlet}(\beta)$ (word prevalence in topic k)
  - $z_d \sim \text{Categorical}(\pi)$ (topic for doc d)
  - $w_{d,n} \sim \text{Categorical}(\phi_{z_d})$ (word n in doc d)
- Inference Problems
  - fully supervised learning: $p(\pi, \phi \mid w, z)$
  - semi-supervised learning: $p(\pi, \phi \mid w, z, \tilde{w})$
  - unsupervised clustering: $p(\tilde{z} \mid \tilde{w})$
  - fully supervised prediction: $p(\tilde{z} \mid \tilde{w}, w, z)$
  - semi-supervised prediction: $p(\tilde{z} \mid \tilde{w}, w, z, \tilde{\tilde{w}})$

Part V

Markov Chain Monte Carlo

Markov Chain Monte Carlo

- Standard Monte Carlo draws i.i.d. samples

$$\theta^{(1)}, \dots, \theta^{(M)}$$

  according to a probability function $p(\theta)$

- Drawing i.i.d. samples is often impossible when dealing with complex densities like Bayesian posteriors $p(\theta|y)$

- So we use Markov chain Monte Carlo (MCMC) in these cases and draw $\theta^{(1)}, \dots, \theta^{(M)}$ from a Markov chain

Markov Chains

- A Markov Chain is a sequence of random variables

$$\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(M)}$$

  such that $\theta^{(m+1)}$ only depends on $\theta^{(m)}$, i.e.,

$$p(\theta^{(m+1)}|y, \theta^{(1)}, \dots, \theta^{(m)}) = p(\theta^{(m+1)}|y, \theta^{(m)})$$

- Drawing $\theta^{(1)}, \dots, \theta^{(M)}$ from a Markov chain according to $p(\theta^{(m+1)} \mid \theta^{(m)}, y)$ is more tractable
- Still require marginal $p(\theta^{(m)}|y)$ at each element of the chain $p(\theta^{(1)}, \dots, \theta^{(M)}|y)$ to be equal to true posterior

Random-Walk Metropolis

- Draw random initial parameter vector $\theta^{(1)}$ (in support)
- For $m \in 2{:}M$
  - Sample proposal from a (symmetric) jumping distribution, e.g.,

    $$\theta^* \sim \text{MultiNormal}(\theta^{(m-1)}, \ \sigma \mathbf{I})$$

    where $\mathbf{I}$ is the identity matrix
  - Draw $u^{(m)} \sim \text{Uniform}(0, 1)$ and set

    $$\theta^{(m)} = \begin{cases} \theta^* & \text{if } u^{(m)} < \dfrac{p(\theta^* | y)}{p(\theta^{(m)} | y)} \\ \theta^{(m-1)} & \text{otherwise} \end{cases}$$

Metropolis and Normalization

- Metropolis only uses posterior in a ratio:

$$\frac{p(\theta^* \mid y)}{p(\theta^{(m)} \mid y)}$$

- This allows the use of unnormalized densities
- Recall Bayes's rule:

$$p(\theta|y) \propto p(y|\theta)\, p(\theta)$$

- Thus we only need to evaluate sampling (likelihood) and prior
- It also sidesteps computing the normalizing term

$$\int_{\Theta} p(y|\theta)\, p(\theta) d\theta$$

Metropolis-Hastings

- Generalizes Metropolis to asymmetric proposals
- Acceptance ratio is

$$\frac{J(\theta^{(m)}|\theta^*) \times p(\theta^*|y)}{J(\theta^*|\theta^{(m)}) \times p(\theta^{(m)}|y)}$$

  where J is the proposal density

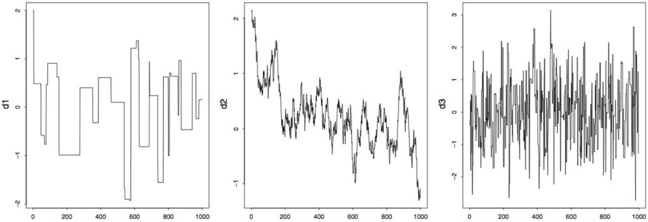
- i.e.,

$$\frac{\text{probabilty of being at } \theta^* \text{ and jumping to } \theta^{(m)}}{\text{probability of being at } \theta^{(m)} \text{ and jumping to } \theta^*}$$

- General form ensures equilibrium for any jumping distribution by maintaining *detailed balance*
- Like Metropolis, only requires ratios
- Many algorithms involve a Metropolis-Hastings “correction”

Optimal Proposal Scale?

- Proposal scale σ is a free; too low or high is inefficient



(a) Proposal variance too large (b) Proposal variance too small (c) Proposal variance approximately optimised

- *Traceplots* show parameter value on y axis, iterations on x
- Empirical tuning problem; theoretical optima exist for some cases

Roberts and Rosenthal (2001) Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*.

Correlations in Posterior Draws

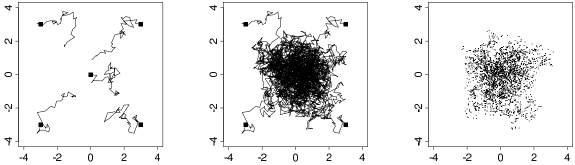
- Markov chains typically display autocorrelation in the series of draws $\theta^{(1)}, \ldots, \theta^{(m)}$
- Without i.i.d. draws, central limit theorem *does not apply*
- Effective sample size N_{eff} divides out autocorrelation
- N_{eff} must be estimated from sample
  - Use fast Fourier transform to efficiently compute correlations at all lags
- Estimation accuracy proportional to

$$\frac{1}{\sqrt{N_{\text{eff}}}}$$

- Compare previous plots; good choice of σ leads to high N_{eff}

Convergence

- Imagine releasing a hive of bees in a sealed house
  - they disperse, but eventually reach equilibrium where the same number of bees leave a room as enter it
- May take many iterations for Markov chain to reach equilibrium



- Four chains with different starting points
- *Left*) 50 iterations; *Center*) 1000 iterations; *Right*) Draws from second half of each chain

Gibbs Sampling

- Draw random initial parameter vector $\theta^{(1)}$ (in support)
- For $m \in 2{:}M$
  - For $n \in 1{:}N$:
    - draw $\theta_n^{(m)}$ according to conditional

$$p(\theta_n | \theta_1^{(m)}, \ldots, \theta_{n-1}^{(m)},\ \theta_{n+1}^{(m-1)}, \ldots, \theta_N^{(m-1)}, y).$$

- e.g, with $\theta = (\theta_1, \theta_2, \theta_3)$:
  - draw $\theta_1^{(m)}$ according to $p(\theta_1 | \theta_2^{(m-1)}, \theta_3^{(m-1)}, y)$
  - draw $\theta_2^{(m)}$ according to $p(\theta_2 | \theta_1^{(m)}, \theta_3^{(m-1)}, y)$
  - draw $\theta_3^{(m)}$ according to $p(\theta_3 | \theta_1^{(m)}, \theta_2^{(m)}, y)$

Generalized Gibbs

- "Proper" Gibbs requires the conditional Monte Carlo draws
  - typically works only for conjugate priors
- In general case, may need to use less efficient conditional draws
  - Slice sampling is a popular general technique that works for discrete or continuous θ_n
  - Adaptive rejection sampling is another alternative

Sampling Efficiency

- We care only about N_{eff} per second
- Decompose into
  1. Iterations per second
  2. Effective samples per iteration
- Gibbs and Metropolis have high iterations per second (especially Metropolis)
- But they have low effective samples per iteration (especially Metropolis)
- Both are particular weak when there is high correlation among the parameters in the posterior

Hamiltonian Monte Carlo & NUTS

- Slower iterations per second than Gibbs or Metropolis
- Much higher number of effective samples per iteration for complex posteriors (i.e., high curvature and correlation)

- Details in the next talk . . .
- Along with details of how Stan implements HMC and NUTS

The End