Florida State University Bayesian Workshop

Applied Bayesian Analysis for the Social Sciences

Day 1: Theory and Foundations

Ryan Bakker
University of Georgia

- **Purpose:**

  To introduce traditional Bayesian methods, Bayesian statistical theory, modern Bayesian computing (MCMC), and applications. Topics include:

  - History, Background and Notation
  - Required Probability Theory
  - Likelihood Theory and Estimation
  - Bayesian Model Theory
  - All Kinds of Priors
  - Assessing Model Quality
  - Bayesian Hypothesis Testing
  - Posterior Simulation, MCMC
  - Bayesian Hierarchical Models
  - More MCMC

Early Challenges:

- Matrix algebra.
- Simple calculus.
- Maximum Likelihood.
- Probability basics.
- `R`.

▶ Campaign Promises:

1. You will understand what the Bayesian approach is all about.
2. You will have a better understanding of conditional probability and mathematical statistics in general.
3. You will gain experience with state-of-the-art tools used by research statisticians.
4. You will understand the basics of MCMC and how to produce results using `WinBUGS`.
5. You will work hard/have fun.

So What's All This *&#$@*$% Bayesian Stuff Anyway?

- Overt and clear model assumptions.
- A rigorous way to make *probability* statements about the real quantities of theoretical interest.
- An ability update these statements (i.e. learn) as new information is received.
- Systematic incorporation of *qualitative* knowledge on the subject.
- Recognition that population quantities are changing over time rather than fixed immemorial.
- The ability to model a wide class of data types.
- Straightforward assessment of both model quality and sensitivity to assumptions.
- Freedom from the flawed NHST paradigm.

Problems with Traditional Social Science Statistics

- Hypothesis Testing/Stars.

- “Confidence.”

- Contrived ignorance.

- Buried assumptions.

Some Problems with Bayesian Statistics

- A very parametric approach (in the basic setup).

- Assigning prior distributions adds more mathematical formalism.

- Interpretation of low-n results.

- Challenge in communicating results with substantive scholars.

The History of Bayesian Statistics–Milestones

- Reverend Thomas Bayes.
- Pierre Simon Laplace.
- Pearson (Karl), Fisher, Neyman and Pearson.
- Jeffreys, de Finetti, Good, Savage, Lindley.
- “Holy wars.”
- The revolution: Gelfand and Smith (1990).
- Today.

Two Primary Principles of Bayesian Inference

1. Explicit and direct use of probability for describing uncertainty:
   - ▷ probability models (likelihood fn.) for data given parameters,
   - ▷ probability distributions (PDF,PMF) for parameters.
2. Inference for unknown values conditioned on observed data:
   - ▷ use of inverse probability,
   - ▷ Bayes theorem,
   - ▷ description of full posterior.

Three General Steps

- Specify a probability model for unknown parameter values that includes some prior knowledge about the parameters if available.

- Update knowledge about the unknown parameters by conditioning this probability model on observed data.

- Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

Simple Mechanics

$$\pi(\theta|\mathbf{x}) = \frac{p(\theta)L(\theta|\mathbf{x})}{\int_\Theta p(\theta)L(\theta|\mathbf{x})d\theta}$$

$$\propto p(\theta)L(\theta|\mathbf{x})$$

Posterior Probability $\propto$ Prior Probability × Likelihood Function

A Note on the Denominator

- The “integrated likelihood” is the denominator of Bayes law calculated here by:

$$p(\mathbf{x}) = \int \underbrace{L(\theta|\mathbf{x})p(\theta)}_{\text{likelihood}\times\text{prior}} \; d\theta$$

- This is also called the “marginal likelihood,” the “marginal probability of the data,” or the “predictive probability of the data”.

- Why do we treat this as a constant?

- This quantity is often ignored since it can be recovered later, but it is important in Bayesian model comparison.

Example: the Beta-Binomial

- $X_1, X_2, \dots, X_n$ iid Bernoulli, $p \sim \text{beta}(A, B)$ prior.

- Standard trick: $Y = \sum_{i=1}^{n} X_i \sim \text{binomial}(n, p)$.

- Joint Distribution:

$$\begin{aligned} f(y,p) &= f(y|p)f(p) \\ &= \left[\binom{n}{y} p^y (1-p)^{n-y}\right] \quad \times \quad \left[\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} p^{A-1}(1-p)^{B-1}\right] \\ &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} p^{y+A-1}(1-p)^{n-y+B-1} \end{aligned}$$

- Marginal Distribution for y:

$$\begin{aligned} f(y) &= \int_0^1 \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} p^{y+A-1}(1-p)^{n-y+B-1} dp \\ &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} \frac{\Gamma(y+A)\Gamma(n-y+B)}{\Gamma(n+A+B)} \end{aligned}$$

Example: the Beta-Binomial, Cont.

- Posterior Distribution for p:

$$f(p|y) = \frac{f(y,p)}{f(y)}$$

$$= \frac{\frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} p^{y+A-1}(1-p)^{n-y+B-1}}{\frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} \frac{\Gamma(y+A)\Gamma(n-y+B)}{\Gamma(n+A+B)}}$$

$$= \frac{\Gamma(n+A+B)}{\Gamma(y+A)\Gamma(n-y+B)} p^{(y+A)-1}(1-p)^{(n-y+B)-1}$$

$$p|y \sim \text{beta}(y+A, n-y+B)$$

- An implication:

$$\hat{p} = \frac{(y+A)}{(y+A)+(n-y+B)} = \left[\frac{n}{A+B+n}\right]\left(\frac{y}{n}\right) + \left[\frac{A+B}{A+B+n}\right]\left(\frac{A}{A+B}\right)$$

Example: the Beta-Binomial, Cont.

- The Data (Romney 1999):

Response:	1	1	1	1	0	1	1	0	1	0	1	1
	1	0	1	1	1	1	1	1	0	0	0	1

- Two Priors: $\mathcal{BE}(p|15, 2)$, $\mathcal{BE}(p|1, 1)$

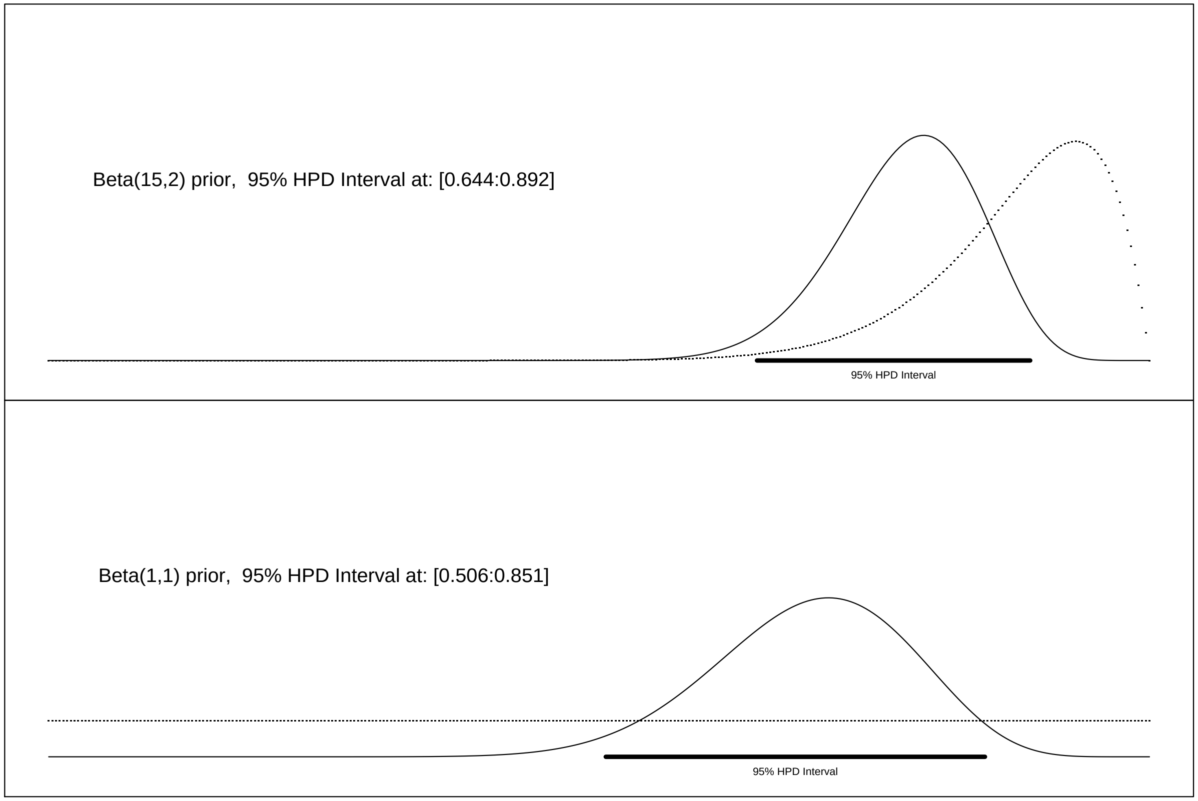
- Resulting Posteriors:

$$\mathcal{BE}\left(\sum x_i + 15, n - \sum x_i + 2\right) = \mathcal{BE}(32, 9),$$

and $$\mathcal{BE}\left(\sum x_i + 1, n - \sum x_i + 1\right) = \mathcal{BE}(18, 8)$$

Example: the Beta-Binomial, Cont.

Figure 1: The Effects of Different Beta Priors in the Beta-Binomial Model



Example: the Beta-Binomial, Cont.

```
par(oma=c(1,1,1,1),mar=c(0,0,0,0),mfrow=c(2,1))
x <- c(1,1,1,1,0,1,1,0,1,0,1,1,1,0,1,1,1,1,1,1,0,0,0,1)
ruler <- seq(0,1,length=300)

A <- 15; B <- 2
beta.prior <- dbeta(ruler,A,B)
beta.posterior <- dbeta(ruler,sum(x)+A,length(x)-sum(x)+B)
plot(ruler,beta.prior, ylim=c(-0.7,9.5),xaxt="n", yaxt="n", xlab="", ylab="", pch=".")
lines(ruler,beta.posterior)
hpd.95 <- qbeta(c(0.025,0.975),sum(x)+A,length(x)-sum(x)+B)
segments(hpd.95[1],0,hpd.95[2],0,lwd=4)
text(mean(hpd.95),-0.4,"95% HPD Interval",cex=0.6)
text(0.25,5,paste("Beta(",A,",",B,") prior, 95% HPD Interval at: [",
  round(hpd.95[1],3),":",round(hpd.95[2],3),"]",sep=""),cex=1.1)
```

```
A <- 1; B <- 1
beta.prior <- dbeta(ruler,A,B)
beta.posterior <- dbeta(ruler,sum(x)+A,length(x)-sum(x)+B)
plot(ruler,beta.prior, ylim=c(-0.7,9.5),xaxt="n", yaxt="n", xlab="", ylab="", pch=".")
lines(ruler,beta.posterior)
hpd.95 <- qbeta(c(0.025,0.975),sum(x)+A,length(x)-sum(x)+B)
segments(hpd.95[1],0,hpd.95[2],0,lwd=4)
text(mean(hpd.95),-0.4,"95% HPD Interval",cex=0.6)
text(0.25,5,paste("Beta(",A,",",B,") prior, 95% HPD Interval at: [",
  round(hpd.95[1],3),":",round(hpd.95[2],3),"]",sep=""),cex=1.1)
```

Bureaucratic Politics Example

- Contains *every* federal political appointee to full-time positions requiring Senate confirmation from November, 1964 through December, 1984 (collected by Mackenzie and Light, ICPSR Study Number 8458, Spring 1987).

- The survey queries various aspects of the Senate confirmation process, acclamation to running an agency or program, and relationships with other functions of government.

Bureaucratic Politics Example

- **Outcome Variable:** `stress` as a surrogate measure for self-perceived effectiveness and job-satisfaction, measured as a five-point scale from “not stressful at all” to “very stressful.”

- **Explanatory Variables:**
  - `Government Experience`,
  - `Ideology`,
  - `Committee Relationship`,
  - `Career.Exec-Compet`,
  - `Career.Exec-Liaison/Bur`,
  - `Career.Exec-Liaison/Cong`,
  - `Career.Exec-Day2day`,
  - `Career.Exec-Diff`,
  - `Confirmation Preparation`,
  - `Hours/Week`,
  - `President Orientation`.

The Model

- A Bayesian random effects specification for ordered survey outcomes, so latent variable thresholds for $\mathbf{Y}$ are assumed on the ordering:

$$\mathbf{U}_i:\ \theta_0 \underset{c=1}{\Longleftrightarrow} \theta_1 \underset{c=2}{\Longleftrightarrow} \theta_2 \underset{c=3}{\Longleftrightarrow} \theta_3 \ldots \theta_{C-1} \underset{c=C}{\Longleftrightarrow} \theta_C$$

- The vector of (unseen) utilities across individuals in the sample, $\mathbf{U}$, is determined by a linear additive specification of explanatory variables: $\mathbf{U} = -\mathbf{X}'\gamma + \boldsymbol{\eta}$, where $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_p]$ does not depend on the θ_j, and $\boldsymbol{\eta} \sim F_{\boldsymbol{\eta}}$.

- This means that the probability that individual i in the sample is observed to be in category r or lower is:

$$P(\mathbf{Y}_i \leq r|\mathbf{X}_i) = P(\mathbf{U}_i \leq \theta_r) = P(\boldsymbol{\eta} \leq \theta_r + \mathbf{X}_i'\gamma) = F_{\boldsymbol{\eta}_i}(\theta_r + \mathbf{X}_i'\gamma).$$

- Specifying a logistic distributional assumption on the errors and adding the random effect term produces this logistic cumulative specification for the whole sample:

$$F_{\boldsymbol{\eta}}(\theta_r + \mathbf{X}'\gamma + \mathbf{b}) = P(\mathbf{Y} \leq r|\mathbf{X}) = [1 + \exp(-\theta_r - \mathbf{X}'\gamma + \mathbf{b})]^{-1}$$

The Model (cont.)

- Prior distributions are either semi-informed or skeptical:

$$
\begin{aligned}
p(\gamma_k) &\sim \mathcal{N}(\mu_{\gamma_k}, \sigma^2_\gamma),\ \ k = 1, \ldots, p \text{ for each of the } p \text{ explanatory variables,} \\
p(\theta_j) &\sim \mathcal{N}(0, \sigma^2_\theta),\ \ j = 1, \ldots, C - 1 \text{ for the four latent variable thresholds,} \\
b_i &\sim \mathcal{N}(0, \tau) \text{ for the random effects term,} \\
\tau &\sim \mathcal{IG}(\delta_1, \delta_2) \text{ for the random effects hyperprior,}
\end{aligned}
$$

The Model (cont.)

- All this produces a posterior distribution according to:

$$\begin{aligned}\pi(\gamma,\boldsymbol{\theta}|\mathbf{X},\mathbf{Y}) &\propto L(\gamma,\boldsymbol{\theta}|\mathbf{X},\mathbf{Y})p(\boldsymbol{\theta})p(\gamma)p(b|\tau)p(\tau)\\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{C-1}\prod_{k=1}^{p}\left[\Lambda(\theta_j-\mathbf{X}_i'\gamma+\mathbf{b}_i)-\Lambda(\theta_{j-1}-\mathbf{X}_i'\gamma+\mathbf{b}_i)\right]^{z_{ij}}\\ &\qquad\times\exp\left(-\frac{(\gamma_k-\mu_{\gamma_k})^2}{2\sigma_\gamma^2}-\frac{\theta_j^2}{2\sigma_\theta^2}-\frac{b_i^2}{2\tau^2}-\frac{\delta_2}{\tau}\right)\tau^{-(\delta_1+1)}\end{aligned}$$

  which is kind of ugly (and hard marginalize).

- Solution: Gibbs sampling.

Who is Markov, and What is He Doing with Chains?

- A major historical problem with the Bayesian approach: sometimes realistic models led to posterior calculations that were difficult or impossible to perform analytically.

- Suppose that instead of performing difficult analytical calculations, one could produce a large number of simulations from the posterior and describe statistics of interest *empirically*.

- Bayesian hierarchical models with realistically specified priors often lead to difficult posterior calculations, but they also setup MCMC procedures in a very straightforward manner.

What is a Markov Chain?

- A type of stochastic process that will help us estimate posterior quantities.

- A *stochastic process* is a **consecutive** set of random quantities defined on some known state space, Θ, indexed so that the order is known: $\{\theta^{[t]} \colon t \in T\}$.

- Frequently, but not necessarily, T is the set of positive integers implying consecutive, even-spaced time intervals:
$$\{\theta^{[t=0]}, \theta^{[t=1]}, \theta^{[t=2]}, \ldots\}.$$

- A stochastic process must also be defined with respect to a *state space*, Θ, which identifies the range of possible values of θ. This state space is either discrete or continuous depending on how the variable of interest is measured.

What is a Markov Chain? (cont.)

- A *Markov chain* is a stochastic process with the property that any specified state in the series, $\theta^{[t]}$, is dependent only on the previous value of the chain, $\theta^{[t-1]}$.

- Therefore values are *conditionally* independent of all other previous values: $\theta^{[0]}, \theta^{[1]}, \ldots, \theta^{[t-2]}$.

- Formally:
$$P(\theta^{[t]} \in A|\theta^{[0]}, \theta^{[1]}, \ldots, \theta^{[t-2]}, \theta^{[t-1]}) = P(\theta^{[t]} \in A|\theta^{[t-1]}),$$
where A is any identified set (an event or range of events) on the complete state space. (We will use this A notation extensively.)

- Colloquially:

> "A Markov chain wanders around the state space remembering only where it has been in the last period."

What is a Markov Chain? (cont.)

- This "short-term" memory property is very useful because when the chain eventually finds the region of the state space with highest density, it will wander around there producing a sample that is only modestly nonindependent.

- If this is the posterior region, then we can use these "empirical" values as legitimate posterior sample values.

- Thus difficult posterior calculations can be done with MCMC by letting the chain wander around "sufficiently long", thus producing summary statistics from recorded values.

What is a Markov Chain? (cont.)

- How does the Markov chain decide to move?

- Define the *transition kernel*, K, as a general mechanism for describing the probability of moving to some other specified state based on the current chain status.

- $K(\theta, A)$ is actually a probability measure for all θ points in the state space to the set $A \in \Theta$.

- So $K(\theta, A)$ maps potential transition events to their probability of occurrence.

What is a Markov Chain? (cont.)

- When the state space is discrete, K is a matrix mapping, $k \times k$ for k discrete elements in A, where each cell defines the probability of a state transition from the first term to all possible states:

$$P_A = \begin{bmatrix} p(\theta_1, \theta_1) & \ldots & p(\theta_1, \theta_k) \\ \vdots & & \vdots \\ p(\theta_k, \theta_1) & \ldots & p(\theta_k, \theta_k) \end{bmatrix},$$

  where the row indicates where the chain is at this period and the column indicates where the chain is going in the next period.

- Each matrix element is a well-behaved probability, $p(\theta_i, \theta_j) \geq 0, \ \forall i, j \in A$.

- Rows of P_A sum to one and define a conditional PMF since they are all specified for the same starting value and cover each possible destination in the state space: for row i: $\sum_{j=1}^{k} p(\theta_i, \theta_j)$.

What is a Markov Chain? (cont.)

- When the state space is continuous, then each row of K is a conditional PDF: $f(\theta|\theta_i)$.

- This *conditional* stipulation means: $f(\theta|\theta_i)$ is a properly defined probability statement for all $\theta \in A$, given some given current state θ_i.

- Continuous state space Markov chains have more involved theory; so it's often convenient to think about discrete Markov chains at first.

A Simple Markov Chain

- A two-dimensional state space: a discrete vote choice between two political parties, a commercial purchase decision between two brands, etc.

- Voters/consumers who normally select θ_1 have an 80% chance of continuing to do so, and voters/consumers who normally select θ_2 have only a 40% chance of continuing to do so.

- The transition matrix P:

$$\begin{array}{cc} & \overbrace{\begin{array}{cc}\theta_1 & \theta_2\end{array}}^{\text{next period}} \\ \text{current period}\left\{\begin{array}{c}\theta_1 \\ \theta_2\end{array}\right. & \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} \end{array}.$$

A Simple Markov Chain (cont.)

- Assign a starting point:

$$S_0 = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$

- To get to the first state, we simply multiply the initial state by the transition matrix:

$$S_1 = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \end{bmatrix}.$$

- This process continues multiplicatively as long as we like:

$$\text{Second state:} \quad S_2 = \begin{bmatrix} 0.7 & 0.3 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.74 & 0.26 \end{bmatrix}$$

$$\text{Third state:} \quad S_3 = \begin{bmatrix} 0.74 & 0.26 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.748 & 0.252 \end{bmatrix}$$

$$\text{Fourth state:} \quad S_4 = \begin{bmatrix} 0.748 & 0.252 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.7496 & 0.2504 \end{bmatrix}.$$

A Simple Markov Chain (cont.)

- Actually, for this simple example we could solve directly for the steady state $S = [s_1, s_2]$ by stipulating:

$$\begin{bmatrix} s_1 & s_2 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \end{bmatrix},$$

  and solving the resulting two equations for the two unknowns.

- This operation of running a Markov chain until it reaches its stationary distribution is exactly the process employed in MCMC.

- The **Big Picture**: Imagine that this stationary distribution was the articulation of some posterior distribution that we could not analytically describe but would like to. If we could run *the right kind* of Markov chain sufficiently long we would eventually get the stationary distribution that describes this posterior empirically with the simulated values.

The Gibbs Sampler

- The Gibbs sampler is a transition kernel created by a series of full conditional distributions.

- It is a Markovian updating scheme based on conditional probability statements.

- If the limiting distribution of interest is $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is an k length vector of coefficients to estimate, then the objective is to produce a Markov chain that cycles through these conditional statements moving toward and then around this distribution.

- The set of full conditional distributions for $\boldsymbol{\theta}$ are denoted $\boldsymbol{\Theta}$ and defined by $\pi(\boldsymbol{\Theta}) = \pi(\theta_i|\boldsymbol{\theta}_{-i})$ for $i = 1, \ldots, k$, where the notation $\boldsymbol{\theta}_{-i}$ indicates a specific parametric form from $\boldsymbol{\Theta}$ without the θ_i coefficient.

Gibbs Sampler Mechanics

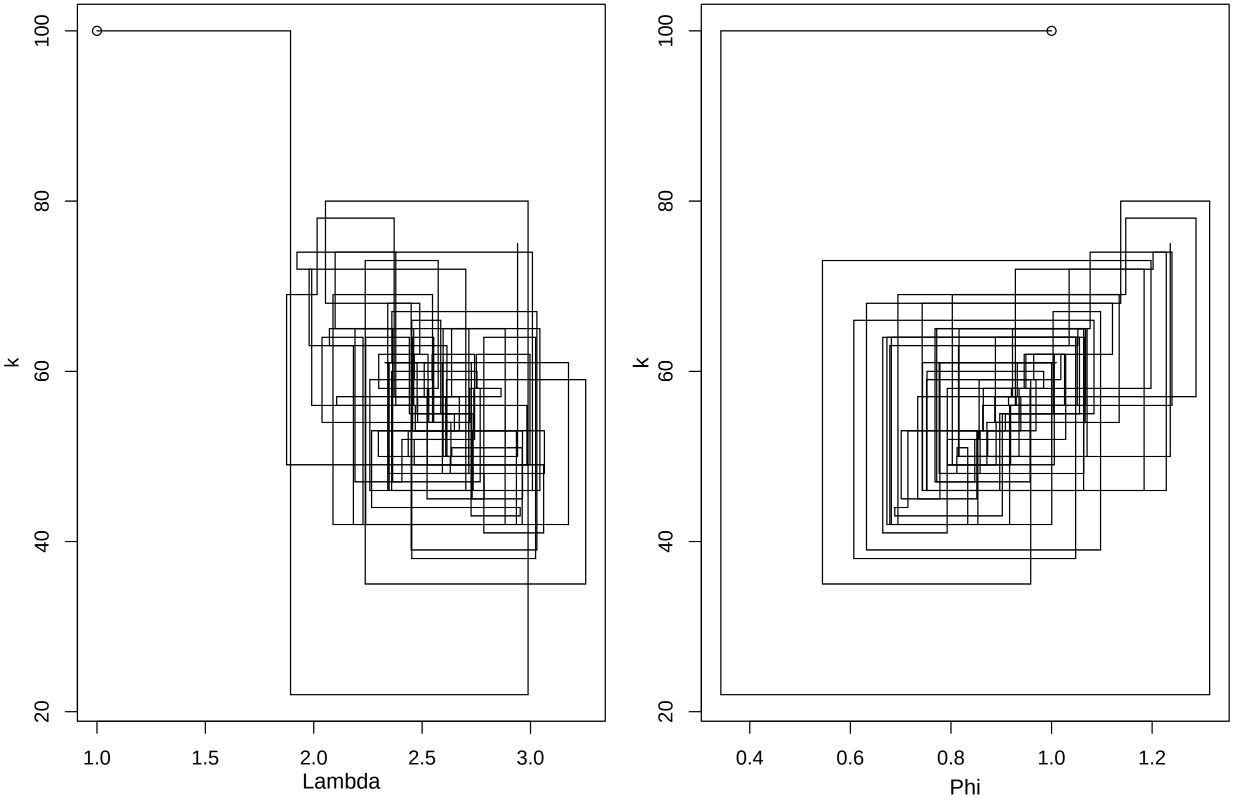
- Steps:

  1. Choose starting values: $\boldsymbol{\theta}^{[0]} = [\theta_1^{[0]}, \theta_2^{[0]}, \ldots, \theta_k^{[0]}]$
  2. At the j^{th} step starting at $j = 1$, drawing values from the k distributions given by:

$$\begin{aligned}
&\theta_1^{[j]} \sim \pi(\theta_1 | \theta_2^{[j-1]}, \theta_3^{[j-1]}, \ldots, \theta_{k-1}^{[j-1]}, \theta_k^{[j-1]}) \\
&\theta_2^{[j]} \sim \pi(\theta_2 | \theta_1^{[j]}, \theta_3^{[j-1]}, \ldots, \theta_{k-1}^{[j-1]}, \theta_k^{[j-1]}) \\
&\theta_3^{[j]} \sim \pi(\theta_3 | \theta_1^{[j]}, \theta_2^{[j]}, \ldots, \theta_{k-1}^{[j-1]}, \theta_k^{[j-1]}) \\
&\vdots \\
&\theta_{k-1}^{[j]} \sim \pi(\theta_{k-1} | \theta_1^{[j]}, \theta_2^{[j]}, \theta_3^{[j]} \ldots, \theta_k^{[j-1]}) \\
&\theta_k^{[j]} \sim \pi(\theta_k | \theta_1^{(j)}, \theta_2^{[j]}, \theta_3^{[j]} \ldots, \theta_{k-1}^{[j]})
\end{aligned}$$

  3. Increment j and repeat until convergence.

Simple Gibbs Sampler Example



WinBUGS Code

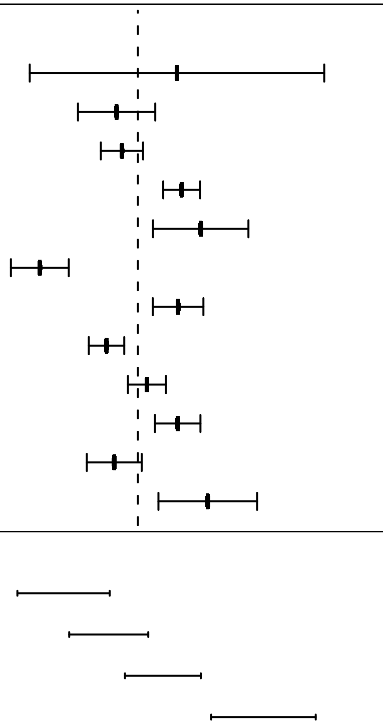
```
      for (i in 1 : N) {
                b[i] ~ dnorm(0.0, tau)
                mu[i] <- theta[1]
                  + theta[2]*previous.empl[i]        + theta[3]*ideology[i]
                  + theta[4]*senate.relat[i]         + theta[5]*confirm.prep[i]
                  + theta[6]*hours.week[i]          + theta[7]*career.exec.compet[i]
                  + theta[8]*career.exec.liason.bur[i]  + theta[9]*career.exec.liason.cong[i]
                  + theta[10]*career.exec.day2day[i]   + theta[11]*career.exec.diff[i]
                  + theta[12]*president.orient[i]

                for (j in 1 : Ncut) {
                           # cumulative probability of lower response than j
                           logit(Q[i, j]) <-  -(k[j] + mu[i] + b[i])
                }
                # probability that response = j
                p[i, 1] <- max( min(1 - Q[i, 1], 1), 0)
                for (j in 2 : Ncut) { p[i,j] <- max( min(Q[i, j - 1] - Q[i, j],1), 0) }
                p[i, (Ncut+1)] <- max( min(Q[i, Ncut], 1), 0)
                stress[i] ~ dcat(p[i, ])
     }
}
list(k=c(-4,-3,-2,-1),tau=2)
```

Posterior Summary

Table 1: Posterior Summary, Model for Survey of Political Executives

	Mean	Std.Err.	95% HPD Intervals
Explanatory Variables:			
`Constant Term`	1.20215	2.24169	
`Government Experience`	-0.65500	0.60664	
`Ideology`	-0.49140	0.32964	
`Committee Relationship`	1.14550	0.39131	
`Career.Exec-Compet`	1.73300	0.85088	
`Career.Exec-Liaison/Bur`	-2.81800	0.55620	
`Career.Exec-Liaison/Cong`	1.03675	0.49193	
`Career.Exec-Day2day`	-0.76595	0.38019	
`Career.Exec-Diff`	0.27780	0.29838	
`Confirmation Preparation`	1.02440	0.45753	
`Hours/Week`	-0.72720	0.42732	
`President Orientation`	1.94950	0.86787	
Threshold Intercepts:			
`None|Little`	-5.93500	1.85782	
`Little|Some`	-2.69250	1.59126	
`Some|Significant`	1.19300	1.52653	
`Significant|Extreme`	8.40450	2.10379	



Posterior $Var(\mathbf{b}) = 6.04350$ (1.20325). Dashed vertical line at zero.

Model Selection with the Bayes Factor

- 2 competing models, M_1: $f_1(\mathbf{x}|\boldsymbol{\theta}_1)$, M_2: $f_2(\mathbf{x}|\boldsymbol{\theta}_2)$.

- Parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ from $\boldsymbol{\Theta}$ or $\boldsymbol{\Theta_1}$ and $\boldsymbol{\Theta_2}$.

- Specify parameter priors: $p_1(\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{\theta}_2)$ and model priors: $p(M_1)$ and $p(M_2)$.

- The "integrated likelihood" is:

$$p(\mathbf{x}|M_k) = \int \underbrace{\ell(\boldsymbol{\theta}_k|M_k,\mathbf{x})}_{\text{likelihood}}\ \underbrace{p(\boldsymbol{\theta}_k)}_{\text{prior}}\, d\boldsymbol{\theta}_k$$

(also called the "marginal likelihood," the "marginal probability of the data," or the "predictive probability of the data").

Model Selection with the Bayes Factor

- Using Bayes Law, we can get the posterior for M_1:

$$p(M_1|\mathbf{x}) = \frac{p(\mathbf{x}|M_1)p(M_1)}{p(\mathbf{x}|M_1)p(M_1) + p(\mathbf{x}|M_2)p(M_2)}$$

which assumes that only two credible models exist.

- What about greater numbers (K) of identified models?

$$p(M_1|\mathbf{x}) = \frac{p(\mathbf{x}|M_1)p(M_1)}{\sum_{k=1}^{K} p(\mathbf{x}|M_k)p(M_k)}$$

Model Selection with the Bayes Factor

- This allows us to calculate:

$$\underbrace{\frac{p(M_1|\mathbf{x})}{p(M_2|\mathbf{x})}}_{\text{posterior odds}} = \underbrace{\frac{p(M_1)/p(\mathbf{x})}{p(M_2)/p(\mathbf{x})}}_{\text{prior odds/data}} \times \underbrace{\frac{\int \ell(\boldsymbol{\theta}_1|M_1,\mathbf{x})p(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int \ell(\boldsymbol{\theta}_k|M_2,\mathbf{x})p(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2}}_{\text{Bayes factor}}.$$

- Which we can rearrange to:

$$B(\mathbf{x}) = \frac{p(M_1|\mathbf{x})/p(M_1)}{p(M_2|\mathbf{x})/p(M_2)}.$$

Bayes Factor Criteria

- Jeffrey's typology:

$$\begin{array}{rl} B(\mathbf{x}) \geq 1 & \text{model 1 supported} \\ 1 > B(\mathbf{x}) \geq 10^{-\frac{1}{2}} & \text{minimal evidence against model 1} \\ 10^{-\frac{1}{2}} > B(\mathbf{x}) \geq 10^{-1} & \text{substantial evidence against model 1} \\ 10^{-1} > B(\mathbf{x}) \geq 10^{-2} & \text{strong evidence against model 1} \\ 10^{-2} > B(\mathbf{x}) & \text{decisive evidence against model 1} \end{array}$$

- Also, if uniform priors are specified this reduces to a likelihood ratio test.

- Bayes Factor for dropping: `Career.Exec-Diff`, `Hours/Week`, adding: `frustration.cong party.id` equals: 0.7837.

- Important point: Bayes factor comparisons don't need to be nested.

AIC

- Akaike, 1973, 1974, 1976.

- Select a model that minimizes the negative likelihood penalized by the number of parameters:

  $$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}|\mathbf{y}) + 2p,$$

  where $\ell(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant).

- Has a strong bias toward models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly

BIC Approximation

- Schwartz 1978, Kass 1993, Kass and Raftery 1995.

- Comparision with the null (constant only) model given by:

$$\text{BIC} = -2\ell(\hat{\theta}|\mathbf{x}) + p\text{log}(n)$$

- For comparing two not-necessarily-nested models, M_j and M_k:

$$B(\mathbf{x}_{jk}) = \frac{p(M_j|\mathbf{x})/p(M_j)}{p(M_k|\mathbf{x})/p(M_k)} = \frac{p(\mathbf{x}|M_j)}{p(\mathbf{x}|M_k)} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_k)} \bigg/ \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_j)} = B(\mathbf{x}_{0k})/B(\mathbf{x}_{0j}).$$

- So:

$$2\log B(\mathbf{x}_{jk}) = 2\log B(\mathbf{x}_{0k}) - 2B(\mathbf{x}_{0j}) \approx BIC_k - BIC_j$$

Schwartz Criterion

- The last result can be restated:

$$S = \text{BIC}_{\text{model k}} - \text{BIC}_{\text{model j}} = \ell(\hat{\theta}_{\mathbf{k}}|\mathbf{x}) - \ell(\hat{\theta}_{\mathbf{j}}|\mathbf{x}) - \frac{1}{2}(p_k - p_j)\log(n)$$

- with the asymptotic property:

$$\frac{S - \log(B(\mathbf{x}_{jk}))}{\log(B(\mathbf{x}_{jk}))} \xrightarrow[n\to\infty]{} 0.$$

- but the bad news:

$$\frac{\exp[S]}{B(\mathbf{x}_{jk})} \nrightarrow 1 \quad \text{as} \quad n \rightarrow \infty$$

Bayesian Model Selection

- Suppose we are choosing between: $M_1, M_2, \ldots, M_K$.

- For each k, determine a model prior: $p(M_k)$.

- Recall that the "integrated likelihood" is:

$$p(\mathbf{x}|M_k) = \int \underbrace{\ell(\boldsymbol{\theta}_k|M_k, \mathbf{x})p(\boldsymbol{\theta}_k|M_k)}_{\text{likelihood}\times\text{prior}} d\boldsymbol{\theta}_k.$$

Bayesian Model Selection, Cont.

- This lets calculate the *posterior model probability* for each model, $\kappa = 1 \ldots K$:

$$p(M_\kappa|\mathbf{x}) = \frac{p(\mathbf{x}|M_\kappa)p(M_\kappa)}{\sum_{j=1}^{K} p(\mathbf{x}|M_j)p(M_j)}.$$

- Interestingly, if we set $p(M_1) = P(M_2) = \ldots = P(M_K) = \frac{1}{K}$, then

$$p(M_k|\mathbf{x}) \approx \frac{\exp[-\frac{1}{2}BIC_k]}{\sum_{j=1}^{K} \exp[-\frac{1}{2}BIC_j]}$$

Simple Bayesian Model Averaging

- For comparative purposes, use a simple averaging scheme from Raftery.

- Start with the posterior mean and variance for the j^{th} coefficient of the κ^{th} model, $\boldsymbol{\theta}_j(\kappa)$ and $Var_{\boldsymbol{\theta}_j}(\kappa)$.

- Consider:

$$P(\boldsymbol{\theta}_j \neq 0|\mathbf{x}) = \sum_{\boldsymbol{\theta}_j \in M_\kappa} p(M_\kappa|\mathbf{x})$$

which is just the "posterior probability that $\boldsymbol{\theta}_j$ is in the model," as well as:

$$E[\boldsymbol{\theta}_j|\mathbf{x}] \approx \sum_{i=1}^{K} \hat{\boldsymbol{\theta}}_j(\kappa)p(M_\kappa|\mathbf{x})$$

and:

$$Var[\boldsymbol{\theta}_j|\mathbf{x}] \approx \sum_{i=1}^{K} \left[(Var_{\boldsymbol{\theta}_j}(\kappa) + \hat{\boldsymbol{\theta}}_j(\kappa)^2)p(M_\kappa|\mathbf{x})) - E[\boldsymbol{\theta}_j|\mathbf{x}]^2\right]$$

Empirical Example

- Raftery reanalyzes Ehrlich's 1933-1969 state level crime data (a famous and highly criticized study of the economic motivations for crime).

- The first published instance of a multivariate linear model on deterrence.

- Outcome variable: crime rate.

- 15 possible explanatory variables: % young male, south, education, police 1960, police 1969, labor part., sex ratio, population, nonwhites, unemployment 14-24, unemployment 35-39, GDP, inequality, probability of prison, prison time.

- Number of possible models: 32,768, calculated from $Num = \sum_{r=0}^{K} \binom{K}{r}$

Empirical Example, Cont.

Table 2: Posterior Summary of Models

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$P(\boldsymbol{\theta}_j \neq 0\|\mathbf{x})$	$E[\boldsymbol{\theta}_j\|\mathbf{x}]$	$\sqrt{Var[T_j\|\mathbf{x}]}$
% young male	■	■	■	■	■	■	■		■	■		■	■	■	0.94	1.40	0.50
south																	
education	■	■	■	■	■	■	■	■	■	■	■	■	■	■	1.00	2.12	0.50
police 1960	■	■		■	■		■	■	■			■		■	0.76	0.95	0.20
police 1959			■			■				■	■		■		0.24	0.97	0.19
labor part.																	
sex ratio																	
population						■		■			■				0.12	-0.08	0.04
nonwhites	■	■	■	■		■		■	■	■	■	■			0.83	0.10	0.04
unemp. 14-24																	
unemp. 35-39	■	■	■		■		■						■		0.68	0.32	0.13
GDP																	
inequality	■	■	■	■	■	■	■	■	■	■	■	■	■	■	1.00	1.33	0.32
prob. prison	■	■	■	■	■	■	■	■	■	■	■	■		■	0.98	-0.24	0.10
prison time.	■			■						■					0.35	-0.30	0.15
R^2	0.84	0.83	0.82	0.82	0.80	0.82	0.80	0.80	0.80	0.81	0.79	0.79	0.78	0.78			
$p(M_k\|\mathbf{x})$	0.24	0.18	0.11	0.08	0.08	0.06	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.02			
BIC	-55.9	-55.4	-54.5	-53.8	-53.6	-53.1	-52.7	-52.4	-52.4	-51.5	-51.3	-51.2	-50.9	-50.9			