

Bayesian Inference and Computation in Political Science

Andrew D. Martin

Department of Political Science

Washington University

St. Louis, Missouri, USA

<http://adm.wustl.edu>

<http://mcmcpack.wustl.edu>

March 9, 2004

Nuffield College, Oxford University

Talk Outline

- Principles of Bayesian Inference
- The Practice of Bayesian Inference
- Linear Regression, Simulation, and MCMCpack
- Application: Ideal Point Estimation
- What's to Come: Modeling Elections in MCMCpack

The Logic of Bayesian Inference

- Bayesian inference is a means of making rational probability statements about quantities of interest (observables, model parameters, functions of model parameters). The central feature of Bayesian inference [is] the direct **quantification of uncertainty** (*Gelman, et. al. 1996, p. 4*).

The Logic of Bayesian Inference

- Bayesian inference is a means of making rational probability statements about quantities of interest (observables, model parameters, functions of model parameters). The central feature of Bayesian inference [is] the direct **quantification of uncertainty** (*Gelman, et. al. 1996, p. 4*).
- Inferences are to be made by combining the information provided by **prior probabilities** with that given by the sample **data**; this combination is achieved by 'the repeated use of **Bayes' theorem**' (Lindley, 1965, p. xi), and the final inferences are expressed solely by the **posterior probabilities** (*Barnett 1999, p. 208*).

The Process of Bayesian Data Analysis

- Setting up a full probability model

The Process of Bayesian Data Analysis

- Setting up a full probability model
- Positing prior beliefs

The Process of Bayesian Data Analysis

- Setting up a full probability model
- Positing prior beliefs
- Calculating and interpreting the posterior distribution

The Process of Bayesian Data Analysis

- Setting up a full probability model
- Positing prior beliefs
- Calculating and interpreting the posterior distribution
- Evaluating model adequacy

Advantages of the Bayesian Approach

- Answers the questions that researchers are really interested in; e.g., “What is the probability that”

Advantages of the Bayesian Approach

- Answers the questions that researchers are really interested in; e.g., “What is the probability that”
- Natural way to combine information from multiple studies

Advantages of the Bayesian Approach

- Answers the questions that researchers are really interested in; e.g., “What is the probability that”
- Natural way to combine information from multiple studies
- Provides a formal method for combining prior qualitative information with observed quantitative information

Advantages of the Bayesian Approach

- Answers the questions that researchers are really interested in; e.g., “What is the probability that”
- Natural way to combine information from multiple studies
- Provides a formal method for combining prior qualitative information with observed quantitative information
- More general way to deal with issues of model identification

Advantages of the Bayesian Approach

- Answers the questions that researchers are really interested in; e.g., “What is the probability that”
- Natural way to combine information from multiple studies
- Provides a formal method for combining prior qualitative information with observed quantitative information
- More general way to deal with issues of model identification
- Principled means to compare non-nested models (Bayes factors)

Advantages of the Bayesian Approach

- Answers the questions that researchers are really interested in; e.g., “What is the probability that”
- Natural way to combine information from multiple studies
- Provides a formal method for combining prior qualitative information with observed quantitative information
- More general way to deal with issues of model identification
- Principled means to compare non-nested models (Bayes factors)
- Allows one to fit very realistic (complicated) models

Disadvantages of the Bayesian Approach

- Often computationally more demanding than classical inference

Disadvantages of the Bayesian Approach

- Often computationally more demanding than classical inference
- At the moment no general-purpose software packages, but see MCMCpack and WinBUGS

Disadvantages of the Bayesian Approach

- Often computationally more demanding than classical inference
- At the moment no general-purpose software packages, but see MCMCpack and WinBUGS
- Requires either:
 1. An elicitation and defense of real subjective prior probability distributions, or
 2. Sensitivity analysis to show that the choice of subjective beliefs is not determining one's inferences

Disadvantages of the Bayesian Approach

- Often computationally more demanding than classical inference
- At the moment no general-purpose software packages, but see MCMCpack and WinBUGS
- Requires either:
 1. An elicitation and defense of real subjective prior probability distributions, or
 2. Sensitivity analysis to show that the choice of subjective beliefs is not determining one's inferences
- Allows one to fit *overly* complicated (realistic) models

Criticisms of Conventional Frequentist Approaches

- The interpretation of confidence intervals

Criticisms of Conventional Frequentist Approaches

- The interpretation of **confidence intervals**
- An implicit reliance on the logic of **repeated sampling**

Criticisms of Conventional Frequentist Approaches

- The interpretation of **confidence intervals**
- An implicit reliance on the logic of **repeated sampling**
- The subjective nature of model specification

Criticisms of Conventional Frequentist Approaches

- The interpretation of **confidence intervals**
- An implicit reliance on the logic of **repeated sampling**
- The subjective nature of model specification
- Null **hypothesis testing**

Criticisms of Conventional Frequentist Approaches

- The interpretation of **confidence intervals**
- An implicit reliance on the logic of **repeated sampling**
- The subjective nature of model specification
- Null **hypothesis testing**
- Numerical and asymptotic properties of **maximum likelihood estimation**

Probability Models

- Consider an observed sample of data y

Probability Models

- Consider an observed sample of data y
- A *probability model* for y consists of 2 things:

Probability Models

- Consider an observed sample of data y
- A *probability model* for y consists of 2 things:
 1. An assumption about the probability distribution with density $p(y|\theta)$ that generated y

Probability Models

- Consider an observed sample of data y
- A *probability model* for y consists of 2 things:
 1. An assumption about the probability distribution with density $p(y|\theta)$ that generated y
 2. The set Θ of possible values of the model parameters θ

Probability Models

- Consider an observed sample of data y
- A *probability model* for y consists of 2 things:
 1. An assumption about the probability distribution with density $p(y|\theta)$ that generated y
 2. The set Θ of possible values of the model parameters θ
- $p(y|\theta)$ is called the *sampling density* and is the joint density of all the observed y s

Probability Models

- Consider an observed sample of data y
- A *probability model* for y consists of 2 things:
 1. An assumption about the probability distribution with density $p(y|\theta)$ that generated y
 2. The set Θ of possible values of the model parameters θ
- $p(y|\theta)$ is called the *sampling density* and is the joint density of all the observed y s
- When $p(y|\theta)$ is viewed as a function of θ for fixed y it is referred to as the *likelihood function* and is written $L(\theta|y)$

Example: Understanding Voter Turnout

- Consider a random sample of size n from the population of registered US voters.

Example: Understanding Voter Turnout

- Consider a random sample of size n from the population of registered US voters.
- We observe that y of the n citizens voted in the 2000 presidential election, the remainder abstained.

Example: Understanding Voter Turnout

- Consider a random sample of size n from the population of registered US voters.
- We observe that y of the n citizens voted in the 2000 presidential election, the remainder abstained.
- We assume that each citizen's decision to vote or abstain follows the Bernoulli distribution, the probability of each citizen voting is equal, and the decisions to vote or abstain are independent.

Example: Understanding Voter Turnout

- Consider a random sample of size n from the population of registered US voters.
- We observe that y of the n citizens voted in the 2000 presidential election, the remainder abstained.
- We assume that each citizen's decision to vote or abstain follows the Bernoulli distribution, the probability of each citizen voting is equal, and the decisions to vote or abstain are independent.
- This implies that y follows the binomial distribution with sample size n and probability of success π .

Example: Understanding Voter Turnout

- Consider a random sample of size n from the population of registered US voters.
- We observe that y of the n citizens voted in the 2000 presidential election, the remainder abstained.
- We assume that each citizen's decision to vote or abstain follows the Bernoulli distribution, the probability of each citizen voting is equal, and the decisions to vote or abstain are independent.
- This implies that y follows the binomial distribution with sample size n and probability of success π .
 - ★ this is the distributional assumption

- The binomial sample size is fixed at n and the parameter π can be any number in the $[0,1]$ interval.

- The binomial sample size is fixed at n and the parameter π can be any number in the $[0,1]$ interval.
 - ★ this determines the parameter space

- The binomial sample size is fixed at n and the parameter π can be any number in the $[0,1]$ interval.
 - ★ this determines the parameter space
- Thus our probability model is:

$$p(y|n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}, \quad \pi \in [0, 1]$$

- The binomial sample size is fixed at n and the parameter π can be any number in the $[0,1]$ interval.

★ this determines the parameter space

- Thus our probability model is:

$$p(y|n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}, \quad \pi \in [0, 1]$$

- The maximum likelihood estimate is:

$$\hat{\pi}_{ML} = \frac{y}{n}$$

Bayesian Inference

- Goal of Bayesian inference is to make probability statements about model parameters θ , and/or functions of model parameters $g(\theta)$, given a probability model and observed data

Bayesian Inference

- Goal of Bayesian inference is to make probability statements about model parameters θ , and/or functions of model parameters $g(\theta)$, given a probability model and observed data
- In other words, we want to know $p(\theta|y)$

Bayesian Inference

- Goal of Bayesian inference is to make probability statements about model parameters θ , and/or functions of model parameters $g(\theta)$, given a probability model and observed data
- In other words, we want to know $p(\theta|y)$
- Note that our probability model is defined in terms of $p(y|\theta)$ which is not quite what we want

Bayesian Inference

- Goal of Bayesian inference is to make probability statements about model parameters θ , and/or functions of model parameters $g(\theta)$, given a probability model and observed data
- In other words, we want to know $p(\theta|y)$
- Note that our probability model is defined in terms of $p(y|\theta)$ which is not quite what we want
- How do we get from $p(y|\theta)$ to $p(\theta|y)$?

- Recall that:

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

- Recall that:

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

- Recall that:

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

- This identity is known as *Bayes' rule*

- Bayes' Rule gives us a formula for calculating the *posterior density* of θ given y [denoted $p(\theta|y)$] from knowledge of the *sampling density* of y [denoted $p(y|\theta)$] and the *prior density* of θ [denoted $p(\theta)$]

- Bayes' Rule gives us a formula for calculating the *posterior density* of θ given y [denoted $p(\theta|y)$] from knowledge of the *sampling density* of y [denoted $p(y|\theta)$] and the *prior density* of θ [denoted $p(\theta)$]
- The function $p(\theta)$ plays a crucial role in transforming our observed data and probability model into probability statements about θ

- Bayes' Rule gives us a formula for calculating the *posterior density* of θ given y [denoted $p(\theta|y)$] from knowledge of the *sampling density* of y [denoted $p(y|\theta)$] and the *prior density* of θ [denoted $p(\theta)$]
- The function $p(\theta)$ plays a crucial role in transforming our observed data and probability model into probability statements about θ
- Since $p(\theta)$ doesn't depend on the observed data y it represents the researcher's subjective *a priori* beliefs about the likely values of θ

- Bayes' Rule gives us a formula for calculating the *posterior density* of θ given y [denoted $p(\theta|y)$] from knowledge of the *sampling density* of y [denoted $p(y|\theta)$] and the *prior density* of θ [denoted $p(\theta)$]
- The function $p(\theta)$ plays a crucial role in transforming our observed data and probability model into probability statements about θ
- Since $p(\theta)$ doesn't depend on the observed data y it represents the researcher's subjective *a priori* beliefs about the likely values of θ
- The fact that $p(\theta)$ is a subjective probability implies that $p(\theta|y)$ is a subjective probability

- Note that since $p(y)$ is a constant for fixed y we can write:

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

- Note that since $p(y)$ is a constant for fixed y we can write:

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

In words, the *posterior density* is proportional to the *sampling density* times the *prior density*.

- Once a probability model is formed and a prior specified Bayesian inference proceeds by summarizing $p(\theta|y)$.

- Once a probability model is formed and a prior specified Bayesian inference proceeds by summarizing $p(\theta|y)$.
- Interesting quantities include (but are not limited to)
 - ★ The posterior mean of θ :

$$\mathbb{E}[\theta|y] = \int_{\Theta} \theta p(\theta|y) d\theta$$

- Once a probability model is formed and a prior specified Bayesian inference proceeds by summarizing $p(\theta|y)$.
- Interesting quantities include (but are not limited to)
 - ★ The posterior mean of θ :

$$\mathbb{E}[\theta|y] = \int_{\Theta} \theta p(\theta|y) d\theta$$

- ★ The posterior variance of θ :

$$\text{Var}[\theta|y] = \int_{\Theta} (\theta - \mathbb{E}[\theta|y])^2 p(\theta|y) d\theta$$

- Once a probability model is formed and a prior specified Bayesian inference proceeds by summarizing $p(\theta|y)$.
- Interesting quantities include (but are not limited to)
 - ★ The posterior mean of θ :

$$\mathbb{E}[\theta|y] = \int_{\Theta} \theta p(\theta|y) d\theta$$

- ★ The posterior variance of θ :

$$\text{Var}[\theta|y] = \int_{\Theta} (\theta - \mathbb{E}[\theta|y])^2 p(\theta|y) d\theta$$

- ★ A $100 \times (1 - \alpha)\%$ credible set $C \subset \Theta$ where C is chosen to satisfy:

$$1 - \alpha = \int_C p(\theta|y) d\theta$$

The Binomial Model, continued

- We need to decide on a *prior density* $p(\pi)$

The Binomial Model, continued

- We need to decide on a *prior density* $p(\pi)$
- Suppose that our prior beliefs about π can be represented by a beta density with parameters a and b

The Binomial Model, continued

- We need to decide on a *prior density* $p(\pi)$
- Suppose that our prior beliefs about π can be represented by a beta density with parameters a and b
- In other words,

$$p(\pi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}$$

The Binomial Model, continued

- We need to decide on a *prior density* $p(\pi)$
- Suppose that our prior beliefs about π can be represented by a beta density with parameters a and b

- In other words,

$$p(\pi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}$$

- Ignoring the normalizing constant we can write:

$$p(\pi) \propto \pi^{a-1} (1-\pi)^{b-1}$$

- Working with the proportional form of Bayes' rule we can write:

$$p(\pi|y) \propto p(y|\pi)p(\pi)$$

- Working with the proportional form of Bayes' rule we can write:

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &\propto \pi^y(1-\pi)^{(n-y)}\pi^{a-1}(1-\pi)^{b-1} \end{aligned}$$

- Working with the proportional form of Bayes' rule we can write:

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &\propto \pi^y(1-\pi)^{(n-y)}\pi^{a-1}(1-\pi)^{b-1} \\ &\propto \pi^{y+a-1}(1-\pi)^{n-y+b-1} \end{aligned}$$

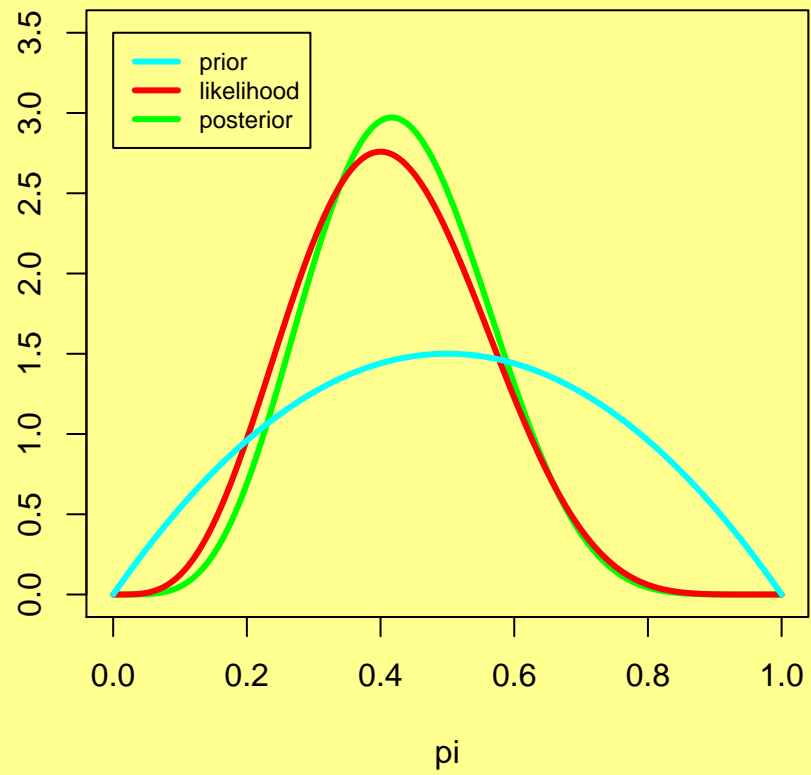
which is proportional to a $\text{Beta}(y + a, n - y + b)$ density

- Working with the proportional form of Bayes' rule we can write:

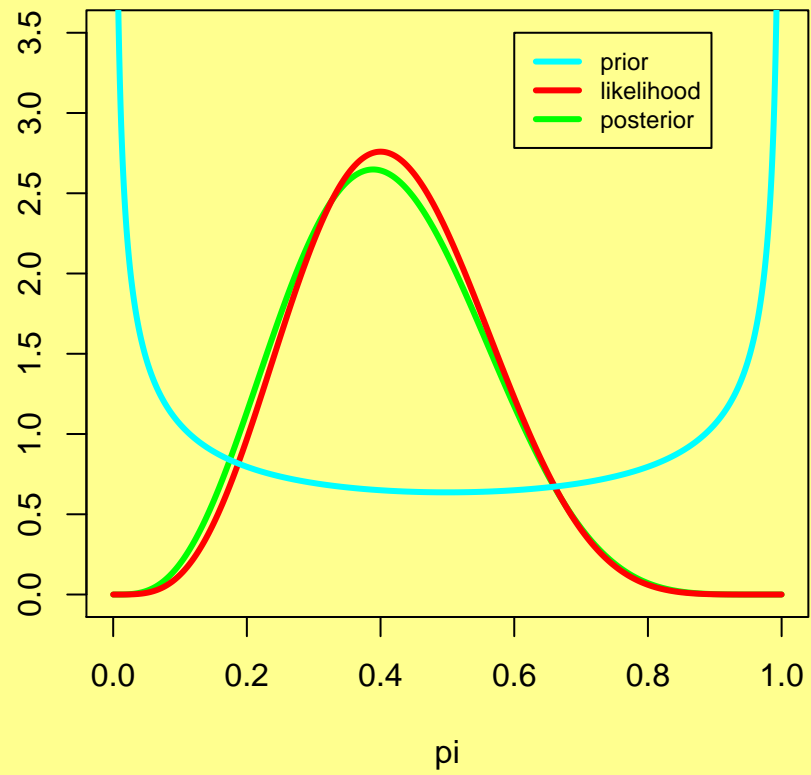
$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &\propto \pi^y(1-\pi)^{(n-y)}\pi^{a-1}(1-\pi)^{b-1} \\ &\propto \pi^{y+a-1}(1-\pi)^{n-y+b-1} \end{aligned}$$

which is proportional to a $\text{Beta}(y + a, n - y + b)$ density

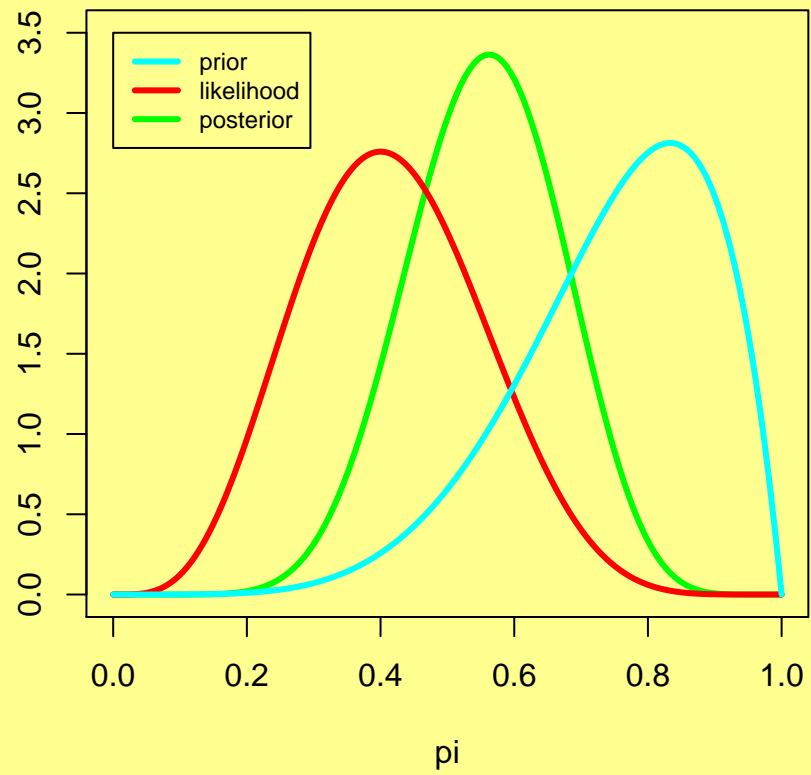
- Interesting: binomial likelihood \times beta prior = beta posterior



$$n = 10, y = 4, a = 2, b = 2$$



$$n = 10, y = 4, a = \frac{1}{2}, b = \frac{1}{2}$$



$$n = 10, y = 4, a = 6, b = 2$$

Example: The Linear Regression Model

- Consider the following simple linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where \mathbf{y} is a $n \times 1$ vector of responses, \mathbf{X} is a $n \times k$ matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of disturbances.

Example: The Linear Regression Model

- Consider the following simple linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where \mathbf{y} is a $n \times 1$ vector of responses, \mathbf{X} is a $n \times k$ matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of disturbances.

- We assume

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Example: The Linear Regression Model

- Consider the following simple linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where \mathbf{y} is a $n \times 1$ vector of responses, \mathbf{X} is a $n \times k$ matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of disturbances.

- We assume

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

- Given the assumption of Gaussian disturbances, we can write the sampling

density as:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right]$$

density as:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right]$$

- Before moving on the Bayesian treatment of this model recall that the OLS and maximum likelihood estimators for $\boldsymbol{\beta}$ are both given by:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

density as:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right]$$

- Before moving on the Bayesian treatment of this model recall that the OLS and maximum likelihood estimators for $\boldsymbol{\beta}$ are both given by:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The classical unbiased estimator of σ^2 is:

$$\hat{\sigma}_{\text{Unbiased}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n - k}$$

- And the maximum likelihood estimator of σ^2 is:

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

- And the maximum likelihood estimator of σ^2 is:

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

- To perform Bayesian inference, the goal is to summarize:

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) (\boldsymbol{\beta}, \sigma^2)$$

- And the maximum likelihood estimator of σ^2 is:

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

- To perform Bayesian inference, the goal is to summarize:

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) (\boldsymbol{\beta}, \sigma^2)$$

- How should we capture prior beliefs?

Semi-Conjugate Priors for the Linear Regression Model

- Let us assume that our prior information about our two parameters are independent:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}) \times p(\sigma^2)$$

Semi-Conjugate Priors for the Linear Regression Model

- Let us assume that our prior information about our two parameters are independent:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}) \times p(\sigma^2)$$

- Further, assume that:

$$p(\boldsymbol{\beta}) \propto \exp \left[-\frac{(\boldsymbol{\beta} - \mathbf{m})' \mathbf{V}^{-1} (\boldsymbol{\beta} - \mathbf{m})}{2} \right]$$

In words, our prior belief is that $\boldsymbol{\beta}$ follows a normal distribution with mean \mathbf{m} and variance-covariance matrix \mathbf{V} .

Semi-Conjugate Priors for the Linear Regression Model

- Let us assume that our prior information about our two parameters are independent:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}) \times p(\sigma^2)$$

- Further, assume that:

$$p(\boldsymbol{\beta}) \propto \exp \left[-\frac{(\boldsymbol{\beta} - \mathbf{m})' \mathbf{V}^{-1} (\boldsymbol{\beta} - \mathbf{m})}{2} \right]$$

In words, our prior belief is that $\boldsymbol{\beta}$ follows a normal distribution with mean \mathbf{m} and variance-covariance matrix \mathbf{V} .

- The semi-conjugate prior distribution for σ^2 is an inverse gamma distribution. We will assume that $\sigma^2 \sim \mathcal{IG}(\nu/2, \delta/2)$.

- With these priors we can (ignoring constants of proportionality) write the posterior density of $\boldsymbol{\beta}$ and σ^2 as:

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} \right] \\ \times \exp \left[-\frac{(\boldsymbol{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m})}{2} \right] (\sigma^2)^{-(\nu/2+1)} \exp \left[-\frac{\delta/2}{\sigma^2} \right] \quad (1)$$

- With these priors we can (ignoring constants of proportionality) write the posterior density of $\boldsymbol{\beta}$ and σ^2 as:

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} \right] \\ \times \exp \left[-\frac{(\boldsymbol{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m})}{2} \right] (\sigma^2)^{-(\nu/2+1)} \exp \left[-\frac{\delta/2}{\sigma^2} \right] \quad (1)$$

- How do we summarize this posterior density?

Estimation via Simulation: Markov chain Monte Carlo (MCMC)

- Gibbs Sampling

Estimation via Simulation: Markov chain Monte Carlo (MCMC)

- Gibbs Sampling
- Metropolis-Hastings

Estimation via Simulation: Markov chain Monte Carlo (MCMC)

- Gibbs Sampling
- Metropolis-Hastings
- We can characterize a posterior density like this:

$$E(\theta|y) = \int_{\Theta} \theta f(\theta|y) d\theta$$

Estimation via Simulation: Markov chain Monte Carlo (MCMC)

- Gibbs Sampling
- Metropolis-Hastings
- We can characterize a posterior density like this:

$$\begin{aligned} E(\theta|y) &= \int_{\Theta} \theta f(\theta|y) d\theta \\ &\approx \frac{1}{G} \sum_{g=1}^G f_g^*(\theta|y) \end{aligned}$$

Software for Bayesian Inference

- WinBUGS

Software for Bayesian Inference

- WinBUGS
- MCMCpack

Software for Bayesian Inference

- WinBUGS
- MCMCpack
- Example: state murder rate regressed on the unemployment rate

```
library(MCMCpack)
m <- matrix(c(0,3),2,1)
V <- matrix(c(1,0,0,1),2,2)
post2 <- MCMCregress(murder ~ unemp, b0=m, B0=V, data=murder)
print(summary(post2))
```

Results from Ordinary Least Squares

```
adm@ichiro R> summary(lm(murder ~ unemp, data=murder))
```

Residuals:

Min	1Q	Median	3Q	Max
-5.5386	-2.6677	-0.6324	2.3935	8.9443

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.5509	2.4810	-0.222	0.82521
unemp	1.4204	0.4535	3.132	0.00296 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.599 on 48 degrees of freedom

Multiple R-Squared: 0.1697, Adjusted R-squared: 0.1524

F-statistic: 9.809 on 1 and 48 DF, p-value: 0.002957

Results from Bayesian Analysis

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
(Intercept)	-0.3296	0.9045	0.02023	0.021013
unemp	1.3965	0.1874	0.00419	0.004327
sigma2	13.2130	2.8381	0.06346	0.061413

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
(Intercept)	-2.029	-0.9553	-0.3485	0.2896	1.424
unemp	1.025	1.2721	1.4022	1.5214	1.750
sigma2	8.881	11.1496	12.9348	14.7436	19.970

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points
- Number of parameters is increasing in the number of votes or voters, which causes problems with frequentist inference

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points
- Number of parameters is increasing in the number of votes or voters, which causes problems with frequentist inference
- A Bayesian approach affords the ability to:

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points
- Number of parameters is increasing in the number of votes or voters, which causes problems with frequentist inference
- A Bayesian approach affords the ability to:
 - ★ Identify the model

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points
- Number of parameters is increasing in the number of votes or voters, which causes problems with frequentist inference
- A Bayesian approach affords the ability to:
 - ★ Identify the model
 - ★ Compute probabilities of quantities of interest

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points
- Number of parameters is increasing in the number of votes or voters, which causes problems with frequentist inference
- A Bayesian approach affords the ability to:
 - ★ Identify the model
 - ★ Compute probabilities of quantities of interest
 - ★ Incorporate real prior beliefs

Application: Ideal Point Estimation

- Long tradition of scaling votes in committees to uncover ideal points
- Number of parameters is increasing in the number of votes or voters, which causes problems with frequentist inference
- A Bayesian approach affords the ability to:
 - ★ Identify the model
 - ★ Compute probabilities of quantities of interest
 - ★ Incorporate real prior beliefs
- Easy to extend the model to include auxiliary information about votes or voters, model the agenda process, and model dynamics

Application: October 2000 Term of the U.S. Supreme Court

- K cases decided by J justices

Application: October 2000 Term of the U.S. Supreme Court

- K cases decided by J justices
- $X \subset \mathbb{R}^D$ the policy space

Application: October 2000 Term of the U.S. Supreme Court

- K cases decided by J justices
- $X \subset \mathbb{R}^D$ the policy space
- $\theta_j \in X$ justice j 's ideal point

Application: October 2000 Term of the U.S. Supreme Court

- K cases decided by J justices
- $X \subset \mathbb{R}^D$ the policy space
- $\theta_j \in X$ justice j 's ideal point
- $\mathbf{x}_k^{(a)} \in X$ policy outcome if Court affirms on case k

Application: October 2000 Term of the U.S. Supreme Court

- K cases decided by J justices
- $X \subset \mathbb{R}^D$ the policy space
- $\theta_j \in X$ justice j 's ideal point
- $\mathbf{x}_k^{(a)} \in X$ policy outcome if Court affirms on case k
- $\mathbf{x}_k^{(r)} \in X$ policy outcome if Court reverses on case k

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to reverse:

$$u_{k,j}^{(r)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 + \delta_{k,j}^{(r)}$$

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to reverse:

$$u_{k,j}^{(r)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 + \delta_{k,j}^{(r)}$$

- Difference in utility:

$$z_{k,j} = u_{k,j}^{(a)} - u_{k,j}^{(r)}$$

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to reverse:

$$u_{k,j}^{(r)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 + \delta_{k,j}^{(r)}$$

- Difference in utility:

$$\begin{aligned} z_{k,j} &= u_{k,j}^{(a)} - u_{k,j}^{(r)} \\ &= -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)} + \|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 - \delta_{k,j}^{(r)} \end{aligned}$$

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to reverse:

$$u_{k,j}^{(r)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 + \delta_{k,j}^{(r)}$$

- Difference in utility:

$$\begin{aligned} z_{k,j} &= u_{k,j}^{(a)} - u_{k,j}^{(r)} \\ &= -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)} + \|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 - \delta_{k,j}^{(r)} \\ &= \left[\mathbf{x}_k^{(r)'} \mathbf{x}_k^{(r)} - \mathbf{x}_k^{(a)'} \mathbf{x}_k^{(a)} \right] + 2\boldsymbol{\theta}_j' \left[\mathbf{x}_k^{(a)} - \mathbf{x}_k^{(r)} \right] + \left[\delta_{k,j}^{(a)} - \delta_{k,j}^{(r)} \right] \end{aligned}$$

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to reverse:

$$u_{k,j}^{(r)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 + \delta_{k,j}^{(r)}$$

- Difference in utility:

$$\begin{aligned} z_{k,j} &= u_{k,j}^{(a)} - u_{k,j}^{(r)} \\ &= -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)} + \|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 - \delta_{k,j}^{(r)} \\ &= \left[\mathbf{x}_k^{(r)'} \mathbf{x}_k^{(r)} - \mathbf{x}_k^{(a)'} \mathbf{x}_k^{(a)} \right] + 2\boldsymbol{\theta}_j' \left[\mathbf{x}_k^{(a)} - \mathbf{x}_k^{(r)} \right] + \left[\delta_{k,j}^{(a)} - \delta_{k,j}^{(r)} \right] \\ &= \alpha_k + \boldsymbol{\beta}_k' \boldsymbol{\theta}_j + \varepsilon_{k,j} \quad \varepsilon_{k,j} \stackrel{iid}{\sim} \mathcal{N}(0, 1) \end{aligned}$$

- Observation equation:

$$y_{k,j} = \begin{cases} 1 \text{ (affirm)} & \text{if } z_{k,j} > 0 \\ 0 \text{ (reverse)} & \text{if } z_{k,j} \leq 0 \end{cases}$$

- Observation equation:

$$y_{k,j} = \begin{cases} 1 \text{ (affirm)} & \text{if } z_{k,j} > 0 \\ 0 \text{ (reverse)} & \text{if } z_{k,j} \leq 0 \end{cases}$$

- Distributional assumption:

$$y_{k,j} | \alpha_k, \beta_k, \theta_j \stackrel{iid}{\sim} \text{Bernoulli}(\pi_{k,j})$$

where

$$\pi_{k,j} = \Phi(\alpha_k + \beta_k' \theta_j)$$

- Observation equation:

$$y_{k,j} = \begin{cases} 1 \text{ (affirm)} & \text{if } z_{k,j} > 0 \\ 0 \text{ (reverse)} & \text{if } z_{k,j} \leq 0 \end{cases}$$

- Distributional assumption:

$$y_{k,j} | \alpha_k, \beta_k, \theta_j \stackrel{iid}{\sim} \text{Bernoulli}(\pi_{k,j})$$

where

$$\pi_{k,j} = \Phi(\alpha_k + \beta_k' \theta_j)$$

- Sampling density:

$$f(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{k \in K} \prod_{j \in J} \Phi(\alpha_k + \beta_k' \theta_j)^{y_{k,j}} \times [1 - \Phi(\alpha_k + \beta_k' \theta_j)]^{1 - y_{k,j}}$$

- Posterior density:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$$

- Posterior density:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$$

- Standard priors (rotational and scale invariance):

★ Ideal Points

$$\boldsymbol{\theta}_j \stackrel{iid}{\sim} \mathcal{N}(\mathbf{t}_0, \mathbf{T}_0) \quad \forall j \in \{1, \dots, J\}$$

- Posterior density:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$$

- Standard priors (rotational and scale invariance):

- ★ Ideal Points

$$\boldsymbol{\theta}_j \stackrel{iid}{\sim} \mathcal{N}(\mathbf{t}_0, \mathbf{T}_0) \quad \forall j \in \{1, \dots, J\}$$

- ★ Case Parameters

$$\begin{bmatrix} \alpha_k \\ \boldsymbol{\beta}_k \end{bmatrix} \stackrel{iid}{\sim} \mathcal{N}_{D+1}(\mathbf{b}_0, \mathbf{B}_0) \quad \forall k \in \{1, \dots, K\}$$

- Posterior density:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$$

- Standard priors (rotational and scale invariance):

- ★ Ideal Points

$$\boldsymbol{\theta}_j \stackrel{iid}{\sim} \mathcal{N}(\mathbf{t}_0, \mathbf{T}_0) \quad \forall j \in \{1, \dots, J\}$$

- ★ Case Parameters

$$\begin{bmatrix} \alpha_k \\ \boldsymbol{\beta}_k \end{bmatrix} \stackrel{iid}{\sim} \mathcal{N}_{D+1}(\mathbf{b}_0, \mathbf{B}_0) \quad \forall k \in \{1, \dots, K\}$$

- ★ Estimation via Gibbs sampling (using Data Augmentation)

The Data

- United States Supreme Court Database (Spaeth, 2004)
- Formally decided, precedent-setting cases
- $J = 9$
- $K = 43$

	Rehnquist	Stevens	OConner	Scalia	Kennedy	Souter	...
1	0	1	1	0	1	1	
2	0	1	0	0	0	1	
3	0	1	0	0	0	1	
4	0	0	0	0	0	0	
5	1	1	0	0	1	0	
6	0	1	0	0	0	0	...
7	0	1	1	0	0	1	
8	0	1	0	0	0	0	
9	0	1	0	0	1	1	
10	1	1	1	0	1	1	
11	0	1	1	0	1	1	

...

Estimation Using MCMCpack

```
library(MCMCpack)
data(SupremeCourt)
justices <- c("Rehnquist", "Stevens", "O'Connor", "Scalia",
             "Kennedy", "Souter", "Thomas", "Ginsburg", "Breyer")

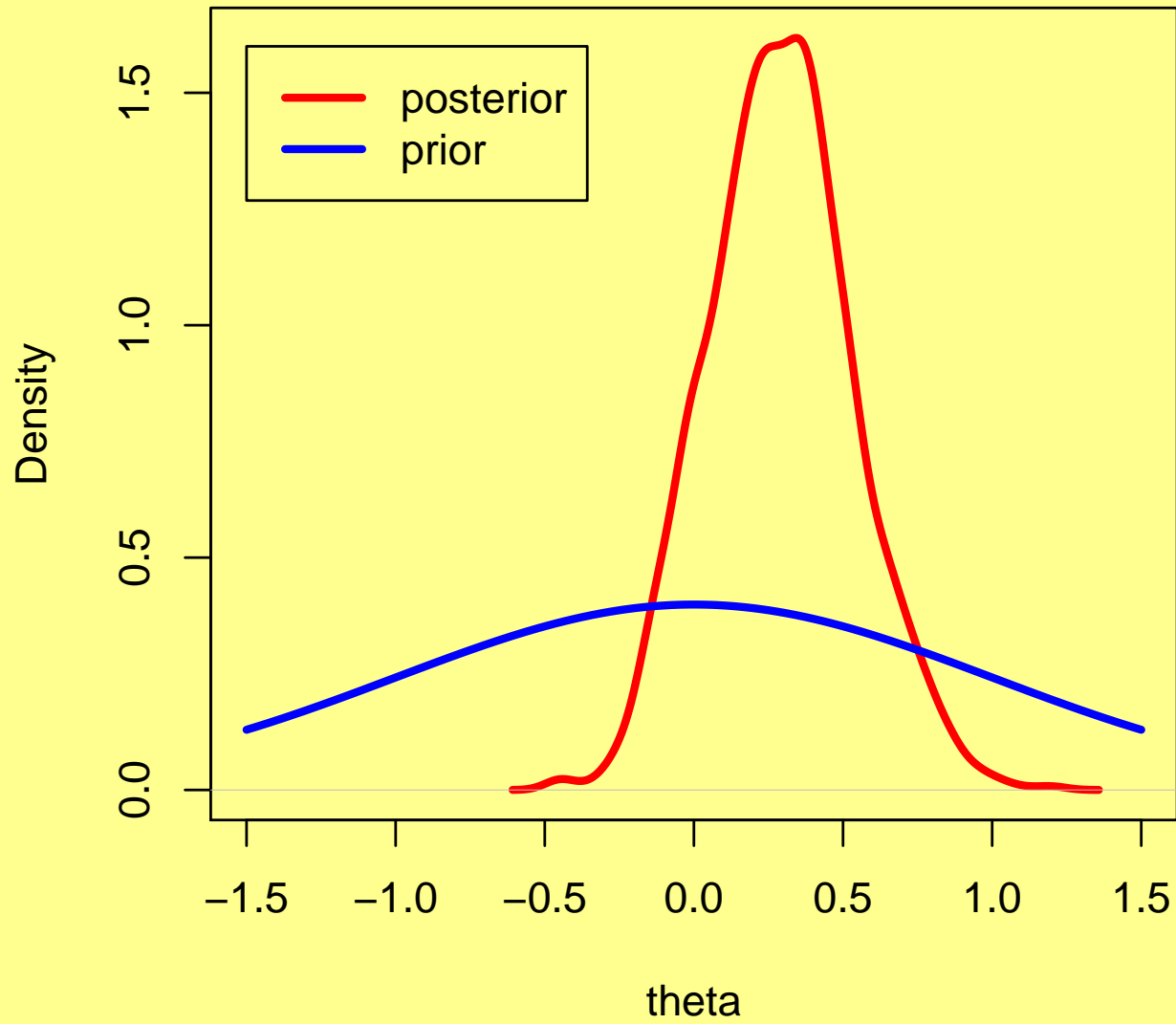
# simulate from the posterior density
posterior1 <- MCMCirt1d(SupremeCourt, theta.fixed=2, burnin=5000,
                       mcmc=10000, verbose=TRUE)
```

Ideal Point Estimates

Posterior Means and Standard Deviations

	justice	post.means	post.sd
theta9	Breyer	-1.7958586	0.4633655
theta2	Stevens	-1.7126669	0.4578604
theta8	Ginsburg	-1.6554844	0.4557253
theta6	Souter	-1.2609081	0.3676670
theta3	O'Connor	0.2891483	0.2428497
theta5	Kennedy	0.6122917	0.2786350
theta1	Rehnquist	1.5873015	0.4068046
theta7	Thomas	2.2558902	0.5313497
theta4	Scalia	2.4241692	0.5701865

Prior and Posterior Densities for O'Connor



The Median Justice

Posterior Probabilities of Being Median Justice

	justices	probabilities
[1,]	"Rehnquist"	"5e-04"
[2,]	"Stevens"	"0"
[3,]	"O'Connor"	"0.841"
[4,]	"Scalia"	"0"
[5,]	"Kennedy"	"0.1585"
[6,]	"Souter"	"0"
[7,]	"Thomas"	"0"
[8,]	"Ginsburg"	"0"
[9,]	"Breyer"	"0"

Future Developments in MCMCpack

- MCMCpack contains code to estimate the following models: linear regression (with Gaussian errors), a general linear panel model, Wakefield's ecological inference model, Quinn's dynamic ecological inference model, Wakefield's hierarchical ecological inference model, a probit model, a logistic regression model, a one-dimensional item response theory model, a K-dimensional item response theory model, a Normal theory factor analysis model, a mixed response factor analysis model, an ordinal item response theory model, a Poisson regression, and an ordered probit model.

Future Developments in MCMCpack

- MCMCpack contains code to estimate the following models: linear regression (with Gaussian errors), a general linear panel model, Wakefield's ecological inference model, Quinn's dynamic ecological inference model, Wakefield's hierarchical ecological inference model, a probit model, a logistic regression model, a one-dimensional item response theory model, a K-dimensional item response theory model, a Normal theory factor analysis model, a mixed response factor analysis model, an ordinal item response theory model, a Poisson regression, and an ordered probit model.
- It is available via CRAN and from <http://mcmcpack.wustl.edu>.

Future Developments in MCMCpack

- MCMCpack contains code to estimate the following models: linear regression (with Gaussian errors), a general linear panel model, Wakefield's ecological inference model, Quinn's dynamic ecological inference model, Wakefield's hierarchical ecological inference model, a probit model, a logistic regression model, a one-dimensional item response theory model, a K-dimensional item response theory model, a Normal theory factor analysis model, a mixed response factor analysis model, an ordinal item response theory model, a Poisson regression, and an ordered probit model.
- It is available via CRAN and from <http://mcmcpack.wustl.edu>.
- The National Science Foundation has made a significant commitment to the project, and we will be extending it in a number of ways over the next three years.

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit
 - ★ Multinomial logit

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit
 - ★ Multinomial logit
 - ★ Models of compositional data

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit
 - ★ Multinomial logit
 - ★ Models of compositional data
 - ★ Mixed response models

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit
 - ★ Multinomial logit
 - ★ Models of compositional data
 - ★ Mixed response models

The first and last of these are essentially inestimable using frequentist methods.

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit
 - ★ Multinomial logit
 - ★ Models of compositional data
 - ★ Mixed response models

The first and last of these are essentially inestimable using frequentist methods.

- An important component of the project is to create an infrastructure others can use to distribute code of their models.

- One area of development are models for the study of (multiparty) elections:
 - ★ Multinomial probit
 - ★ Multinomial logit
 - ★ Models of compositional data
 - ★ Mixed response models

The first and last of these are essentially inestimable using frequentist methods.

- An important component of the project is to create an infrastructure others can use to distribute code of their models.
- We plan to develop in many other areas, and are always looking for feedback (and code!).

Conclusion

Bayesian inference is a powerful tool one can use to perform all types of analysis; a creative analyst can answer all sorts of interesting substantive questions. This takes place by: (a) developing reasonable probability models; (b) thinking carefully about priors; (c) estimating the models [typically using MCMC]; and (d) checking model fit and testing the robustness of conclusions. Moreover, there are many types of models **only** practically estimable using Bayesian methods.

References

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*, Second Edition. Boca Raton, FL: Chapman & Hall.

Jeff Gill. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: Chapman & Hall.

Simon Jackman. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *AJPS*. 44: 369-298.

Jim Albert and Siddhartha Chib. 1996. "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo." *Advances in Econometrics A*. 11: 3-24.

Siddhartha Chib and Edward Greenberg. 1996. "Markov Chain Monte Carlo Simulation Methods in Econometrics." *Econometric Theory*. 12:4090-431.

George Casella and Edward I. George. 1992. "Explaining the Gibbs Sampler." *The American Statistician*. 46: 167-174.

Siddhartha Chib and Edward Greenberg. 1995. "Understanding the Metropolis-Hastings Algorithm." *The American Statistician*. 49: 327-336.