A Pragmatic Justification for the Use of Bayesian Methods in the Social Sciences

Andrew D. Martin
Washington University

October 7, 2005
University of South Carolina
Political Science Research Workshop

Outline

- I. The Bayesian Approach
- II. Model Fitting via Simulation
- III. Pragmatic Justification
- IV. Practical Considerations and Software
- V. Conclusion

I. The Bayesian Approach

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

- Set up the full probability model
- Posit prior beliefs
- Calculate the posterior distribution
- Summarize the posterior distribution
- Check model adequacy

Notation

- Data: $\mathbf{y}$
- Covariates: $\mathbf{x}$
- Parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$
- Probability Model: $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$ $f(\mathbf{y}|\boldsymbol{\theta})$
- Priors: $f(\boldsymbol{\theta})$

Bayes Theorem

- Posterior: $$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})} \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$
- Prior predictive distribution: $$f(\mathbf{y}) = \int_{\boldsymbol{\theta}} f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$$
- Posterior predictive distribution: $$f(\mathbf{y}_{new}|\mathbf{y}) = \int_{\boldsymbol{\theta}} f(\mathbf{y}_{new}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

Example: Linear Regression

- Observations: $i = 1, \ldots, n$
- DV: $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$
- IV: $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- Parameters: $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$
- Probability Model: $y_i | \boldsymbol{\beta}, \sigma^2, \mathbf{x}_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, \sigma^2)$

Likelihood

$$f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) = \prod_{i=1}^{n} \phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) =$$
$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right]$$

- Estimate using OLS or ML and rely on asymptotic theory to get standard errors, perform hypothesis tests, etc.
- Regularity conditions
- Sometimes "integrate out" incidental parameters

Bayesian Inference for Linear Regression

- Priors:

$$\boldsymbol{\beta} \sim \mathcal{N}_K(\mathbf{b}_0, \mathbf{B}_0^{-1})$$

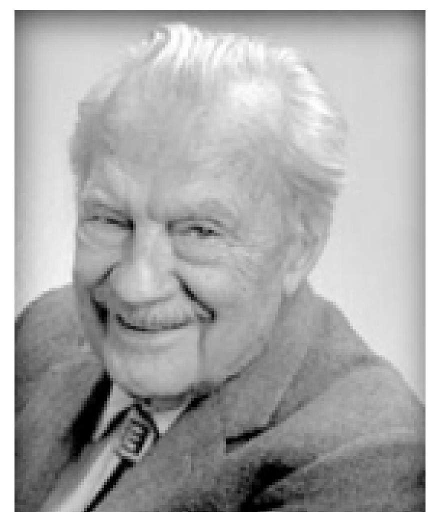
$$\sigma^{-2} \sim \mathcal{G}amma(c_0/2, d_0/2)$$

- Posterior:

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right] \times f(\boldsymbol{\beta}) f(\sigma^2)$$

II. Model Fitting Using Simulation

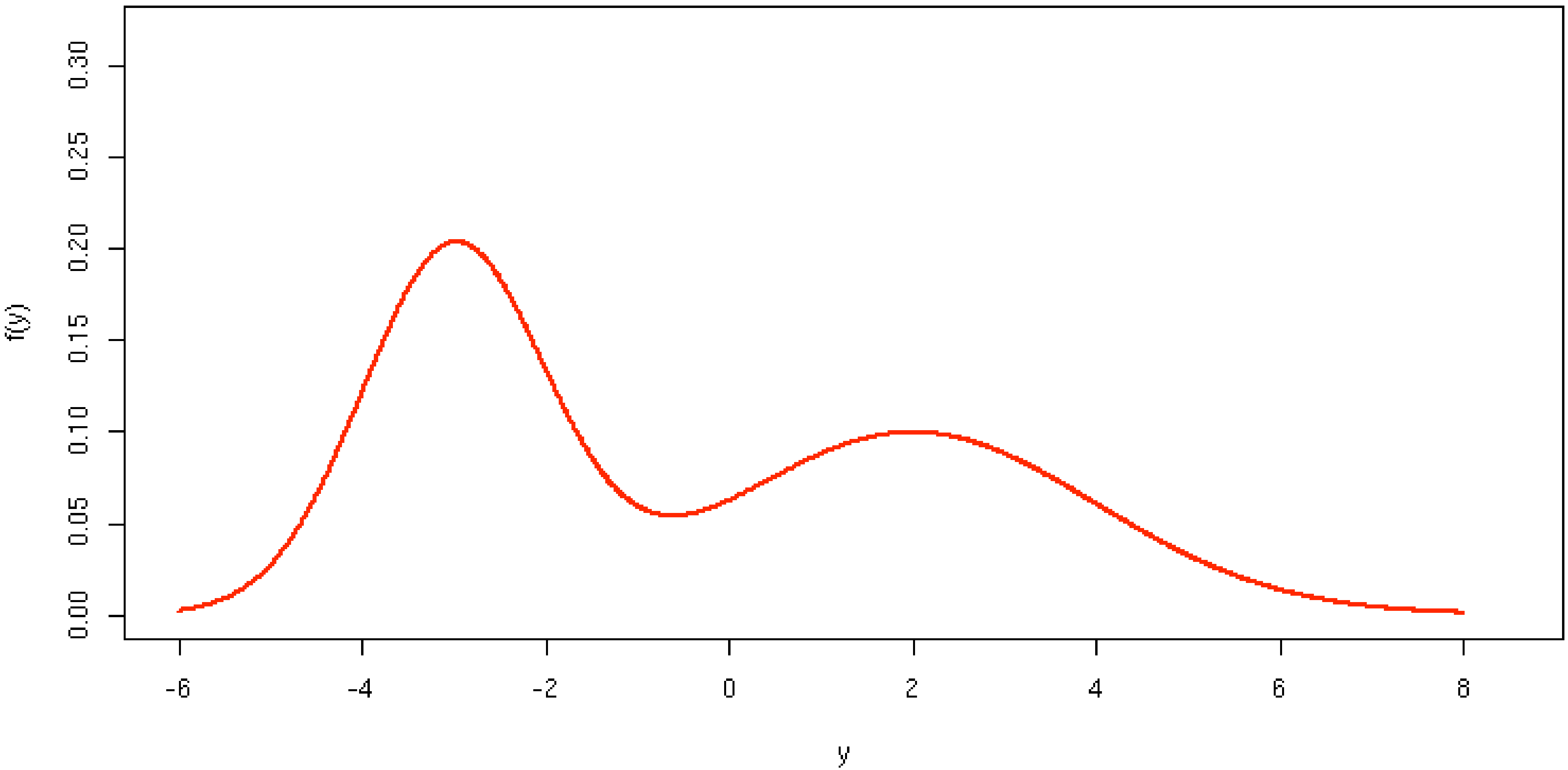
- Monte Carlo Method -- learn anything we want about a random variable if we can randomly sample from its density
- We want to learn about the posterior density



- Markov chain Monte Carlo
- CLARIFY

Learning About A Mixture of Normals

$$f(y) = \frac{1}{2}\phi\left(\frac{y-2}{2}\right) + \frac{1}{2}\phi\left(\frac{y+3}{1}\right)$$



Bayesian Inference Using Simulation

- Posterior distribution is "target"
- We generate sequence of draws from the target $f_g^*(\boldsymbol{\theta}|\mathbf{y})$
- Compute quantities of interest, such as the posterior mean, standard deviation, and credible intervals
- For example:

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

Gibbs Sampling

- For many posteriors (especially with semi-conjugate priors) the conditionals are known distributions

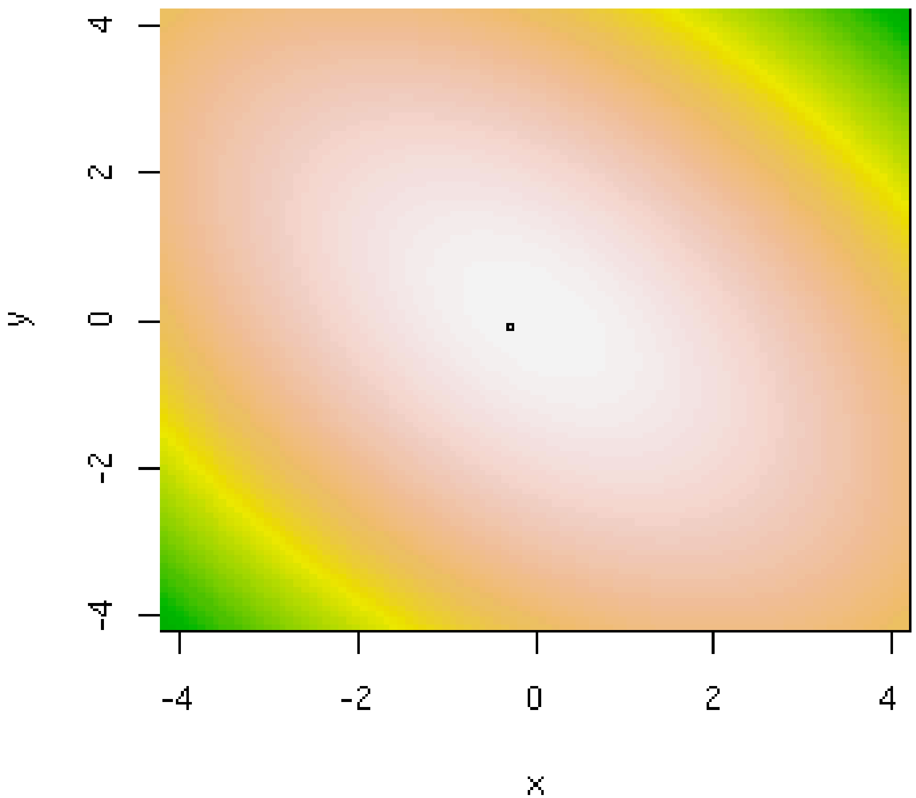
- Target: $f(\theta_1, \theta_2, \theta_3 | \mathbf{y})$

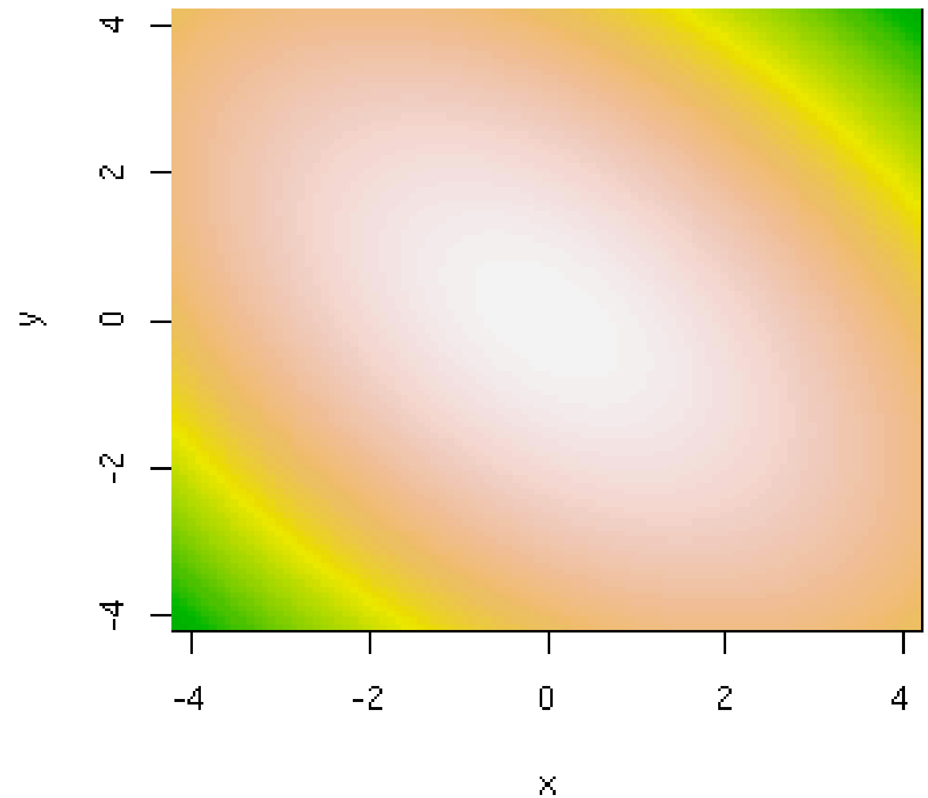
- Starting Values: $\theta_2^{(0)}$ and $\theta_3^{(0)}$

- Repeat: $g = 1, ..., G$

$$\text{Draw } \theta_1^g \text{ from } f(\theta_1 | \theta_2, \theta_3, \mathbf{y})$$
$$\text{Draw } \theta_2^g \text{ from } f(\theta_2 | \theta_1, \theta_3, \mathbf{y})$$
$$\text{Draw } \theta_3^g \text{ from } f(\theta_3 | \theta_1, \theta_2, \mathbf{y})$$

Gibbs Sampling from a Bivariate Normal Distribution

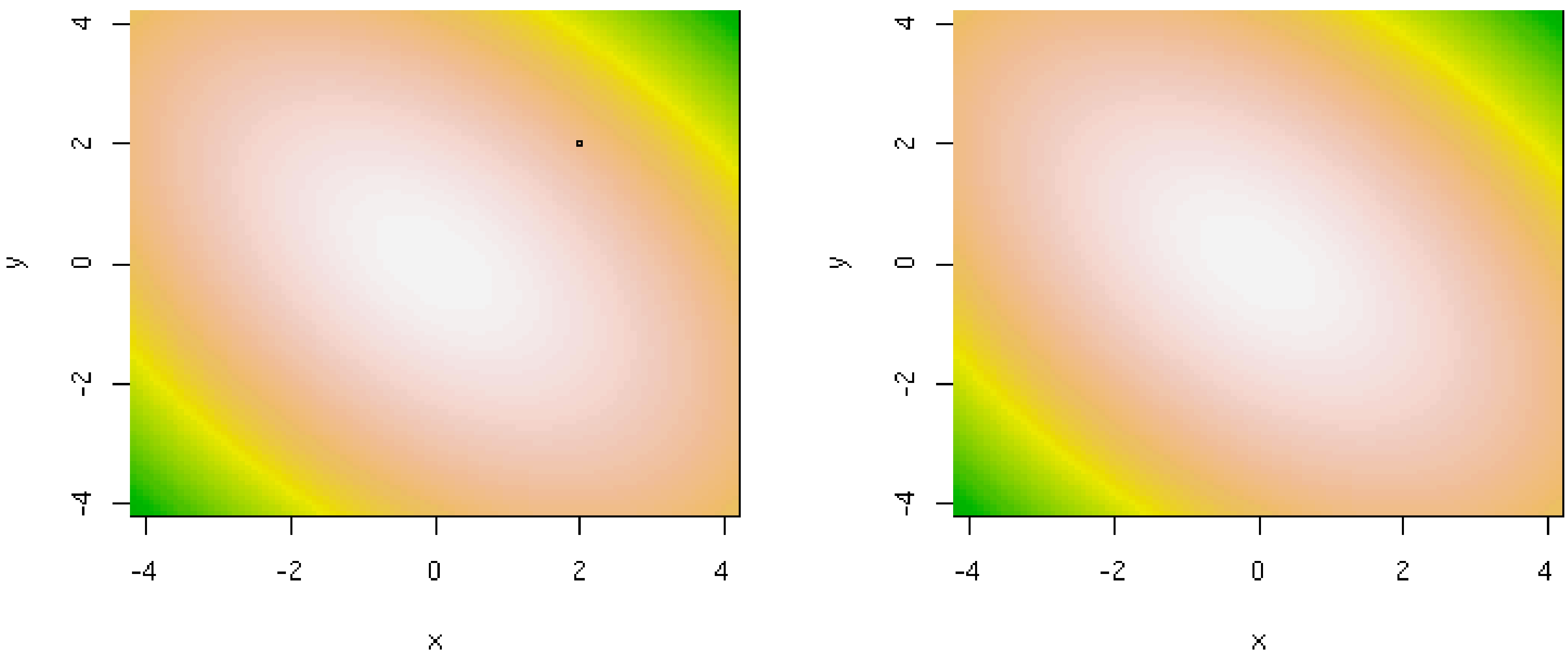




Metropolis-Hastings

- Target: $f(\boldsymbol{\theta}|\mathbf{y})$
- Starting Value: $\boldsymbol{\theta}^{(0)} \qquad g = 1, ..., G$
- Draw proposal: $\boldsymbol{\theta}^* \text{ from } p_g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(g-1)})$
- Set: $$\boldsymbol{\theta}^{(g)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \alpha^* \\ \boldsymbol{\theta}^{(g-1)} & \text{with probability } (1-\alpha^*) \end{cases}$$
- With: $$\alpha^* = \min\left\{\frac{f(\boldsymbol{\theta}^*|\mathbf{y})}{f(\boldsymbol{\theta}^{(g-1)}|\mathbf{y})}\frac{p_g(\boldsymbol{\theta}^{(g-1)}|\boldsymbol{\theta}^*)}{p_g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(g-1)})}, 1\right\}$$

Sampling from a Bivariate Normal Distribution using M-H



III. Pragmatic Justification (Or Six Reasons To Use Bayesian Methods)

A. Intuitive Interpretation of Results

- Confidence Interval vs. Credible Interval
- Hypothesis Testing
- Predicted Values and Probabilities
- No Asymptotic Theory Necessary

B. Quantities of Interest

- Using MCMC it is easy to directly estimate quantities of interest *and* assign probabilities to them
- This is not unlike CLARIFY for predicted values, but is more powerful
- Example -- Who is the median Senator? Is Senator Graham or DeMint more conservative? (Clinton, Jackman, and Rivers, *APSR*, 2004)

- Ideal point model with estimated ideal points: $$\theta_j \text{ for } j = 1, \ldots, J$$

- Draw from posterior: $$\{\theta_1^{(g)}, \theta_2^{(g)}, \ldots, \theta_J^{(g)}\}$$

- Estimate of median: $$\text{median}^{(g)} = \text{med}\left(\theta_1^{(g)}, \theta_2^{(g)}, \ldots, \theta_J^{(g)}\right)$$

- Probability DeMint to right of Graham: $$\theta_{DeMint}^{(g)} > \theta_{Graham}^{(g)}$$

C. Incorporate Prior Information

- Much of the time "uninformative" priors are used
- But other information, including that collected from qualitative sources, can *formally* be included in the analysis using priors
- Must translate this knowledge into statements about parameters; prior elicitation
- Example -- Priors on marginal effects of various covariates on attitudes toward the judiciary in Nicaragua Gill and Walker, *JOP,* 2005 (see also Western and Jackman, *APSR,* 1994)

D. Fit Otherwise Intractable Models

- Multinomial / Multivariate Probit Models (Quinn, Martin, and Whitford, *AJPS*, 1999)

$$\Pr(y_i = 3) = \int_{-\infty}^{0} \int_{0}^{\infty} \int_{-\infty}^{0} \phi_3(\mathbf{z}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) dz_1 dz_2 dz_3$$

- Measurement Models / Structural Models (Clinton, Jackman, and Rivers, *APSR*, 2004)
- Hierarchical / Multi-level Models (Western, *AJPS*, 1998, analyzing the Alvarez, Garrett, and Lange OECD data) / Complex Dependence
- Models with Discrete Parameters (change point models, mixture models with varying number of mixture components)

E. Model Comparison

- Compare models with different blocks of covariates
- Compare models with different functional forms (e.g., logit vs. probit)
- Compare any number of models
- Bayes factors (which use the prior predictive distribution or the marginal likelihood)
- Example -- Multinomial Probit or Multinomial Logit: Quinn, Martin, and Whitford, AJPS, 1999 (see also Kass and Raftery, JASA, 1995)

F. Missing Data

- It is easy to deal with missing data in the Bayesian context
- Think of multiple imputation on the fly
- Data augmentation
- This is done automatically by certain software packages (BUGS)

Concerns

- This seems really complicated. What I currently do seems much easier.
- Priors seem subjective, and should not be part of a scientific analysis.
- Priors will dictate results.
- There is no way to know whether an MCMC algorithm has converged.
- There is not good software to do Bayesian inference.

IV. Practical Considerations and Software

- Roll your own
- The BUGS language
- WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs)
- JAGS (http://calvin.iarc.fr/~martyn/software/jags/)
- OpenBUGS (http://mathstat.helsinki.fi/openbugs/)
- R (http://www.r-project.org/)

BUGS Code to Fit Linear Regression

```
model {

   # likelihood
   for(i in 1:N) {
      Y[i] ~ dnorm(mu[i],tau)
      mu[i] <- inprod(X[i,], beta)
      }
   for(j in 1:M) {

   # priors
   for(k in 1:K) {
      beta[k] ~ dnorm(0,0.001)
      }
   tau ~ dgamma(0.001, 0.001)
   }
```

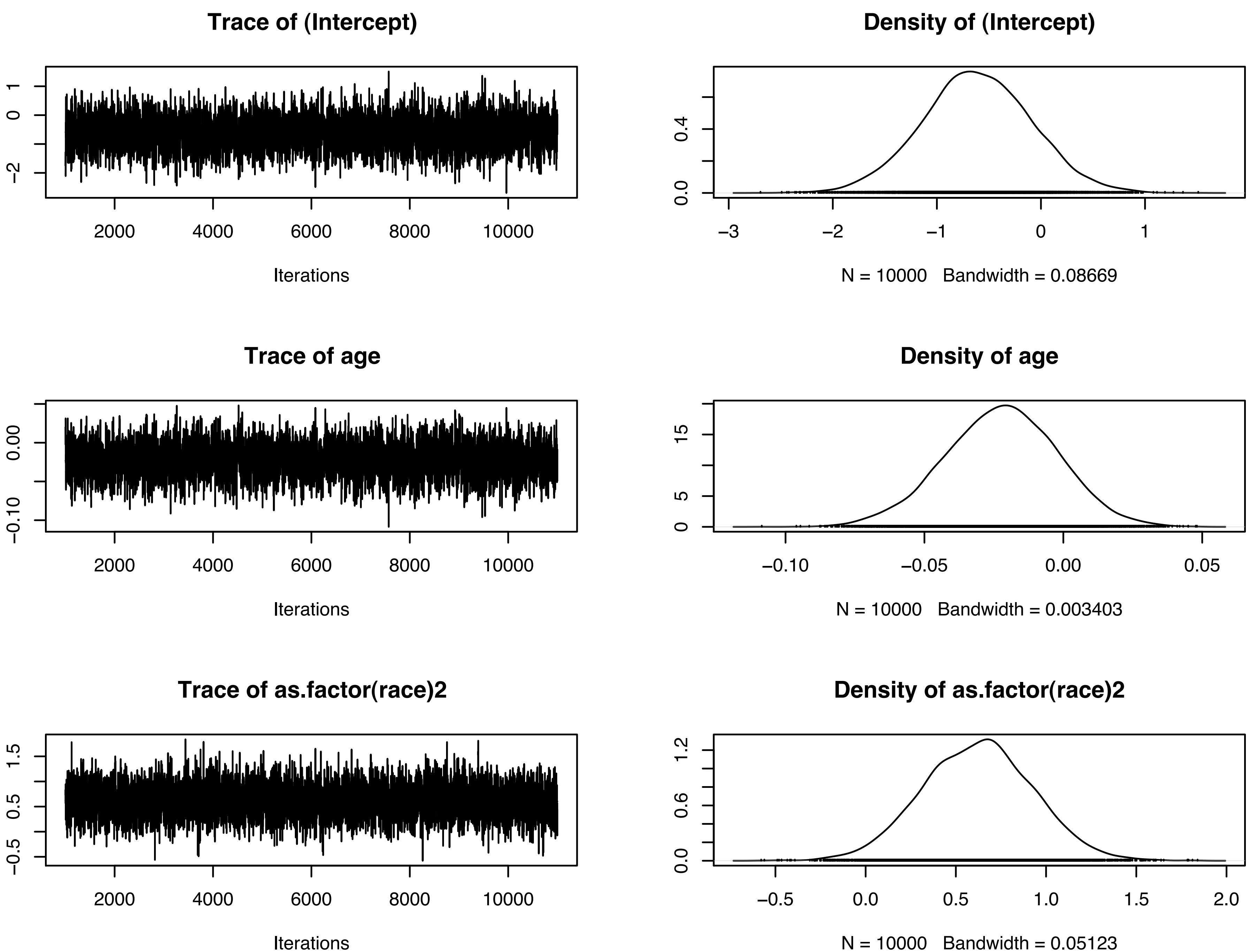
Fitting Models in R

- **MCMCpack** -- general model-fitting for linear regression, a general linear panel model, ecological inference models, probit model, logit, a item response theory models, factor analysis models, Poisson regression, tobit regression, multinomial logit, and an ordered probit model (http://mcmcpack.wustl.edu)
- **bayesm** -- more model fitting for a variety of models in econometrics
- **coda** -- post-estimation analysis for summarization and convergence diagnostics / mcmc objects
- See the Bayesian Task View on the Comprehensive R Archive Network (http://cran.r-project.org/)

MCMCpack Code to Fit a Probit Model

```
library(MCMCpack)
data(birthwt)
posterior <- MCMCprobit(low~age+
   as.factor(race) + smoke, data=birthwt)
summary(posterior)
plot(posterior)
```

```
adm@ichiro R> summary(posterior)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

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

                     Mean      SD  Naive SE Time-series SE
(Intercept)      -0.61453 0.52113 0.0052113      0.0089813
age              -0.02233 0.02026 0.0002026      0.0003908
as.factor(race)2  0.62144 0.30495 0.0030495      0.0051266
as.factor(race)3  0.65385 0.24595 0.0024595      0.0047243
smoke             0.69552 0.22075 0.0022075      0.0036388

2. Quantiles for each variable:

                     2.5%      25%      50%       75%   97.5%
(Intercept)      -1.64237 -0.95598 -0.62114 -0.264527 0.42319
age              -0.06312 -0.03586 -0.02195 -0.008426 0.01626
as.factor(race)2  0.02928  0.41279  0.62681  0.824368 1.21178
as.factor(race)3  0.17655  0.48533  0.65218  0.818054 1.13375
smoke             0.26961  0.54749  0.69319  0.845839 1.13360

adm@ichiro R>
```


Trace of (Intercept)
-2
0
1
2000
4000
6000
8000
10000
Iterations
Density of (Intercept)
0.0
0.4
-3
-2
-1
0
1
N = 10000 Bandwidth = 0.08669
Trace of age
-0.10
0.00
2000
4000
6000
8000
10000
Iterations
Density of age
0
5
15
-0.10
-0.05
0.00
0.05
N = 10000 Bandwidth = 0.003403
Trace of as.factor(race)2
-0.5
0.5
1.5
2000
4000
6000
8000
10000
Iterations
Density of as.factor(race)2
0.0
0.6
1.2
-0.5
0.0
0.5
1.0
1.5
2.0
N = 10000 Bandwidth = 0.05123

V. Conclusion

- Bayesian methods are important components of the political methodology toolkit
- The ability to creatively and flexibly model data and quantify all quantities of interest makes the approach very powerful
- Better software is needed before Bayesian methods become mainstream
- Perhaps it is time to re-think the political methodology canon as it is taught in nearly every Ph.D. program