Introduction to Bayesian Inference

Robert W Rankin (Murdoch University, PhD Candidate)

November 18, 2015

Outline

Outline

Philosophical differences

- Frequentists vs. Bayesian

Priors

- densities, impacts

Practise: JAGS

Computation (e.g., Gibbs sampling, MCMC; if time)

- simulation-based approximatation of the Posteriors

Advantages of Bayesian Inference

- inference statesments: easy to understand (only Bayesians can make probabilistic statements about θ)
- small sample sizes: exact inference
- missing data: very easy to impute
- integrate other information, or calculate 'derived parameters'

Hierarchical Bayesian

- model dependences (space & time)
- "random-effects" models
- "model-selection" / "model-multi inference" (AIC, Lasso, etc., are just types of Bayesian models)
- shrinkage: deflate influence of outlier values

What is "Bayesian" inference?

what comes to mind when you think about "Bayesian"

- ???

What is "Bayesian" inference?

what comes to mind when you think about "Bayesian"

- Priors: most common ecologist's answer (not necessarily true)
- Posterior density $p(\theta|Y)$

– "(posterior) probability density of θ *given* the observed data Y".

- inference on θ *given* data
- θ has a **distribution** of values

What is "Bayesian" inference

compare to the Likelihood

- Priors
- Posterior density $p(\theta|Y)$

– "(posterior) probability density of θ *given* the observed data Y".
– *Only* Bayesian's have access to the Posterior

- Likelihood: $p(Y;\theta)$

– "the joint probability of a realization of the data *given* a particular value of θ".

Maximum-Likelihood

- basis most Frequentist analysis
- Often (but not always) the MLE is the best estimator according to Frequentist's values (consistency, unbiased, etc)

The likelihood & Frequentism

- before we can talk about the posterior... what is the likelihood?

$$p(Y|\theta)$$

data is what is random; θ is given?

- find the value of θ that *maximizes* the probability of having observed the data
- Frequentist emphasize *repeated use*:

if repeat the experiment -> observed slightly different data. Want estimates that are optimal over all theorectical samples of data.

A little demotivation

Most Biologists are "Agnostic Bayesians"

- **Frequentist** vs. **Bayesian**: point estimates nearly identical (under certain conditions)

Example data:

- men's height, n=20 observations
- first run the Frequentist's glm($y \sim 1$) function

Frequentist Example

Example data:

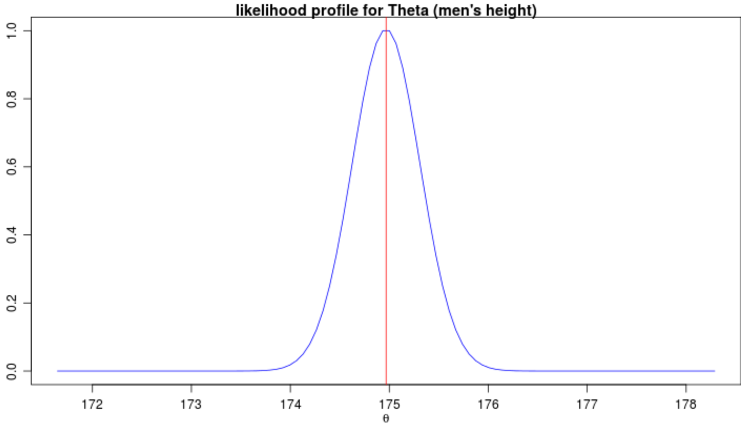
- men's height, n=20 observations
- first run the Frequentist's glm($y \sim 1$) function

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 174.9676   1.530233 114.3405 1.965619e-28
```

- Frequentist: start with a point-estimate, then estimate:

```
[1] "Frequentist:"
       MLE          se      lo95CI      hi95CI
174.967636    1.530233 171.968435 177.966837
```

Likelihood



- "It would be very (un)likely to have seen the data that I saw, if the value of θ were X"
- Choose θ: that which maximize's the likelihood seeing Y
- θ_{MLE} is NOT the "most probabilty value of θ

Bayesians: The Posterior

- Frequentist: start with a point-estimate (MLE), then estimate S.E., 95% Confidence interval, etc

```
[1] "Frequentist:"
       MLE          se      lo95CI      hi95CI
174.967636    1.530233 171.968435 177.966837
```

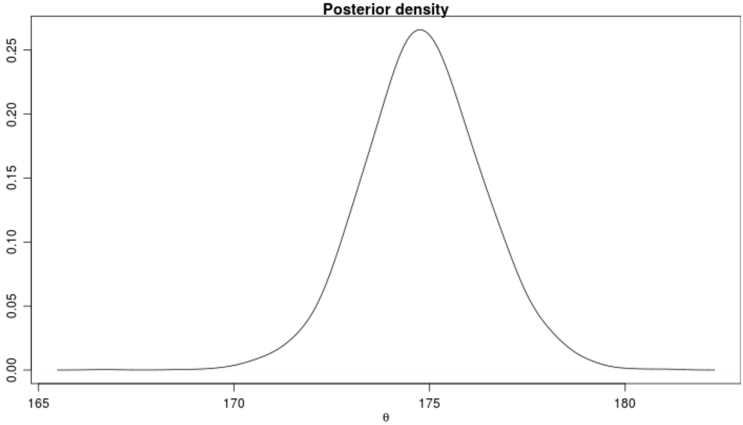
- Bayesian start with a distribution, and then summarize it with simple descriptive statistics

Mean, mode, S.E., 95% Credibility interval

```
[1] "Bayesian"
       mu         se     lo95CI     hi95CI
174.83200    1.54682 171.75725 177.91146
```

Posterior density

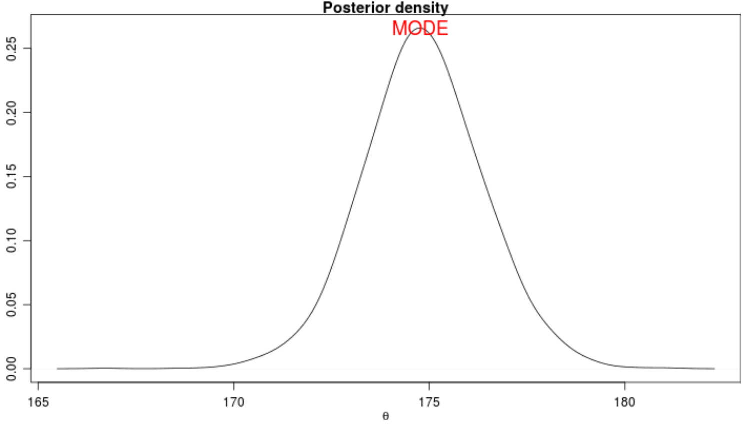
- IS a probability distribution



- easy to interpret

Posterior density

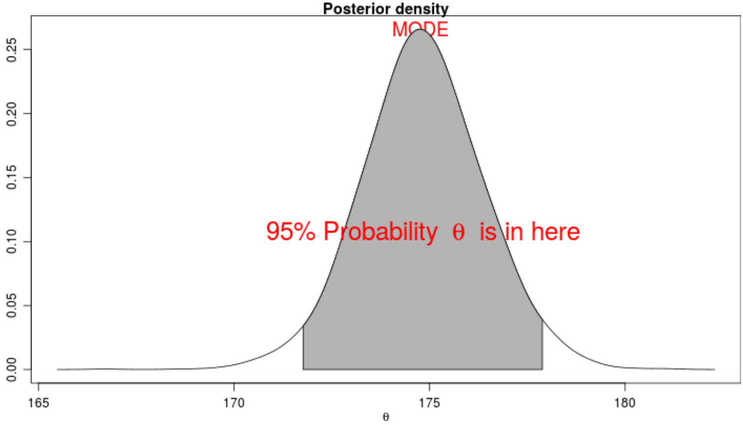
- IS a probability distribution



- Posterior mode: most probable value
- Posterior mean $\mathbb{E}[\theta] = \int p(\theta|Y)\theta d\theta$: expected value

Posterior density

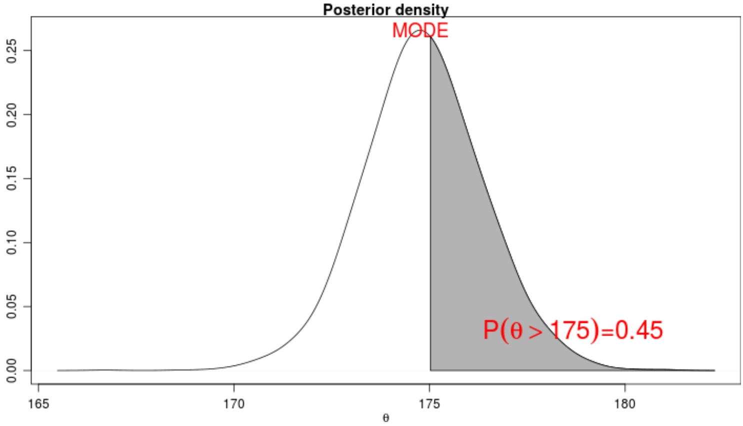
- IS a probability distribution



- Posterior mode: most probable value
- Posterior mean $\mathbb{E}[\theta] = \int p(\theta|Y)\theta d\theta$: expected value
- 95%CI of θ

Posterior density

- IS a probability distribution



- Posterior mode: most probable value
- Posterior mean $\mathbb{E}[\theta] = \int p(\theta|Y)\theta d\theta$: expected value
- What is the probability that $\theta > X$? Area of $p(\theta|Y) > X$

A little demotivation

Most Biologists are "Agnostic Bayesians"

- **Frequentist** vs. **Bayesian**: point estimates nearly identical (under certain conditions)

Example:

- men's height, n=20 observations S.E. and 95% CI

```
[1] "Frequentist:"
       MLE          se      lo95CI      hi95CI
174.967636   1.530233 171.968435 177.966837
[1] "Bayesian"
       mu         se     lo95CI    hi95CI
174.83200   1.54682 171.75725 177.91146
```

A little demotivation

Most Biologists are "Agnostic Bayesians"

- **Frequentist** vs. **Bayesian**: often nearly identical
- only true for: i) certain "priors", and ii) large-samples sizes
- key point: **Be a Master of Priors!**

Bayesians vs. Frequentism

Philosophy

- Bayesians: condition on the data, θ is random

think like a gambler

- Frequentism: data is random

think: had we repeated the experiment, we would get different data

Practical differences?

mostly, no. BUT, some important situations...

- priors!
- low-sample sizes, complex models
- missing data
- 'optional stopping'

Posterior Density

Posterior: the goal of Bayesian analysis...

- hard to evaluation
- Enter **Baye's Rule**!

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta)d\theta}$$

Posterior $\propto$ Likelihood x Prior

- *likelihood*: easy to evaluate
- *prior*: express as easy distribution (Norm, Gamma, Beta)

Priors

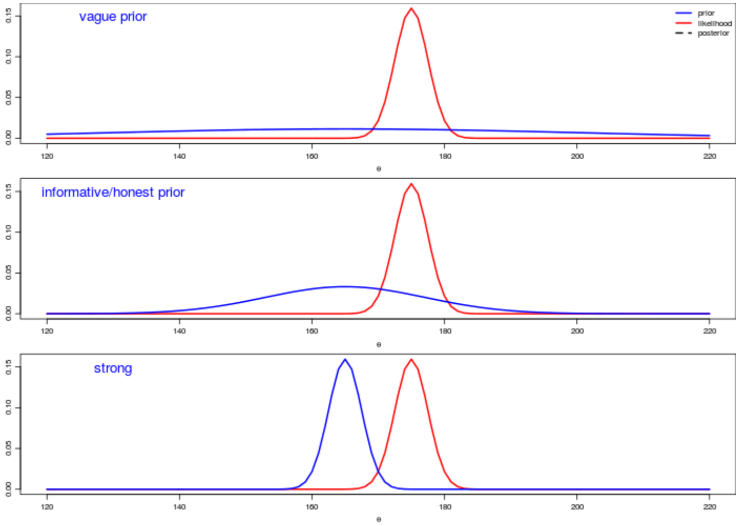
defn: "your belief about θ before observing data", or "a probability distribution about θ before observing data"

Posterior $\propto$ Likelihood x Prior

- The posterior: a mixture of "information in the data" (likelihood) and "information in the prior"
- **be a master or priors**

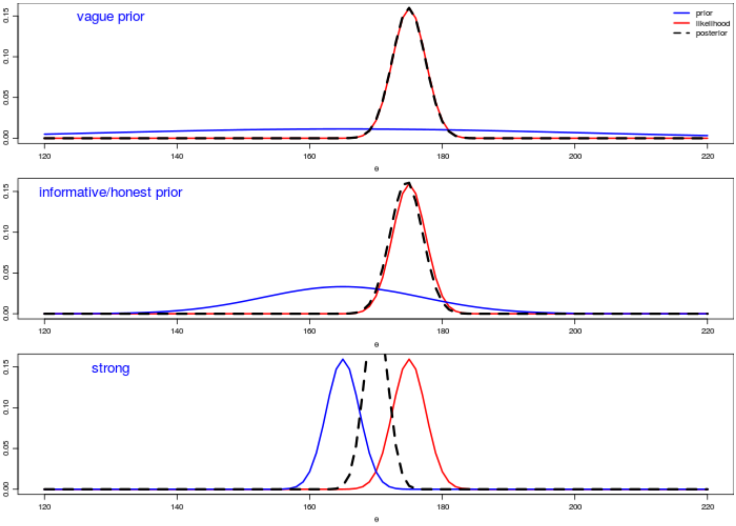
It is your responsibility to study and know how to express prior information in probabilitistic terms

Posterior $\propto$ Likelihood x Prior



Competing information: priors vs. likelihood

Posterior $\propto$ Likelihood x Prior



Competing information: priors vs. likelihood

Types of Priors

non-informative priors

- desire Posterior estimates similar to MLE
- deliberately ignore prior knowledge
- Jeffrey's priors

'Subjective Bayes'

- honest representation of your actual knowledge
- inference: how the data (via likelihood) updates Prior -> Posterior

Strong Priors

- computational reasons
- 'fixing' parameters to a certain value
- non-identifiability of parameter

Types of Priors

Know the distributions and their parameters

Name	Usage	Density	Lower	Upper
Beta	`dbeta(a,b)` $a > 0,\ b > 0$	$\frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)}$	0	1
Chi-square	`dchisqr(k)` $k > 0$	$\frac{x^{\frac{k}{2}-1}\exp(-x/2)}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}$	0	
Double exponential	`ddexp(mu,tau)` $\tau > 0$	$\tau \exp(-\tau\lvert x-\mu\rvert)/2$		
Exponential	`dexp(lambda)` $\lambda > 0$	$\lambda \exp(-\lambda x)$	0	
F	`df(n,m)` $n > 0,\ m > 0$	$\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}\left\{1+\frac{nx}{m}\right\}^{-\frac{(n+m)}{2}}$	0	
Gamma	`dgamma(r, lambda)` $\lambda > 0,\ r > 0$	$\frac{\lambda^r x^{r-1}\exp(-\lambda x)}{\Gamma(r)}$	0	
Generalized gamma	`dgen.gamma(r,lambda,b)` $\lambda > 0,\ b > 0,\ r > 0$	$\frac{b\lambda^{br}x^{br-1}\exp\{-(\lambda x)^b\}}{\Gamma(r)}$	0	
Logistic	`dlogis(mu, tau)` $\tau > 0$	$\frac{\tau\exp\{(x-\mu)\tau\}}{[1+\exp\{(x-\mu)\tau\}]^2}$		
Log-normal	`dlnorm(mu,tau)` $\tau > 0$	$\left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} x^{-1}\exp\left\{-\tau(\log(x)-\mu)^2/2\right\}$	0	
Noncentral Chi-squre	`dnchisqr(k, delta)` $k > 0, \delta \geq 0$	$\sum_{r=0}^{\infty}\frac{\exp(-\frac{\delta}{2})(\frac{\delta}{2})^r}{r!}\frac{x^{(k/2+r-1)}\exp(-\frac{x}{2})}{2^{(k/2+r)}\Gamma(\frac{k}{2}+r)}$	0	
Normal	`dnorm(mu,tau)` $\tau > 0$	$\left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}}\exp\{-\tau(x-\mu)^2/2\}$		
Pareto	`dpar(alpha, c)` $\alpha > 0,\ c > 0$	$\alpha c^{\alpha} x^{-(\alpha+1)}$	c	

Figure : rjags Plummer 2015

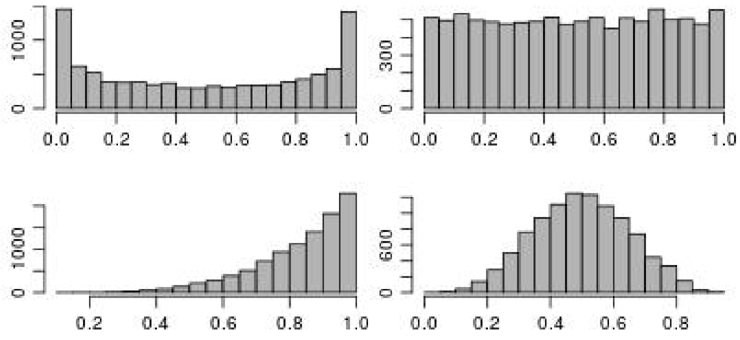
Types of Priors

Know the distributions and their parameters

Familiarize yourself with distributions: plot it, calculate some statistics

- Beta distribution example

```
r <- rbeta(10000, 0.5, 0.5)
```



Bayesian Analysis Example

Time to open up R and JAGS

'JAGS: Just Another Gibbs Sampler'

Uses BUGS-like syntax (similar to OpenBUGS, WinBUGS)

- rjags Package: R friendly JAGS interface
- easy easy easy Bayesian inference

Don't worry about 'samplers': JAGS does the hard work

- specify **likelihood** (how the data arose) and the **priors**

Bayesian Analysis Example

example model: height of 20 Australian

```
y <- c(183.46, 182.32, 178.31, 181.36, 165.12,
185.68, 170.47, 178.11, 174.86, 182.03, 180.09,
172.88, 177.94, 177.26, 182.58, 171, 173.74, 177.78,
180.02, 163.05)
```

- lets estimate the mean height (mu) and the dispersion (sigma)

JAGS we estimate the 'precision' (tau): $\tau = \frac{1}{\sigma^2}$

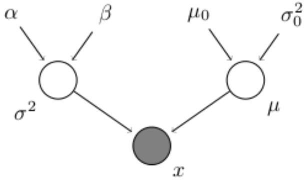


Figure : Prof Mike Jordan lecture notes

Bayesian Analysis Example 1

- open up R angs rjags
- download and open the R file:

Bayesian Analysis Example 1

Jags model syntax: specify priors and likelihood

```
model.txt<-'model{
 # Normal priors on mean height
 mu0 <- 100
 sigma0 <- 35
 tau0 <- pow(sigma0,-2)
 mu ~ dnorm(mu0,tau0)
 # Gamma prior on precision
 alpha0 <- 0.1
 beta0 <- 0.1
 tau ~ dgamma(alpha0,beta0)
 # Likelihood: how the data arose
 for(i in 1:length(y)){
   y[i] ~ dnorm(mu,tau) T(0,) # truncated normal
 }
 sigma <- pow(tau,-0.5)
}'
```

Sample-based inference

Posteriors

often no 'analytical' solution to $P(\theta|Y)$

Solution: Sampling

- it is a Probability Distribution!!!
- find a way to "sample" from posterior.
- with enough samples: mean(samples) = Posterior Expectation

Sampling Algorithms

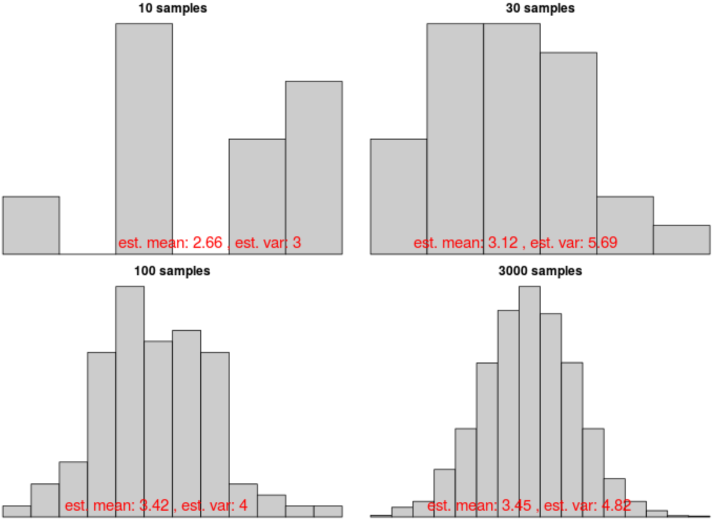
MCMC; Gibbs Sampling; Metropolis-Hastings; Slice-Sampling; Importance Sampling; "Conjugate Priors"; conditional probability

- all (sub)algorithms or concepts or techniques to help sample a posterior

Approximate the joint-posterior distribution"

example: estimate mean and variance of θ

$\theta_{\text{true}} = 3.44$; $\text{Var}(\theta)_{\text{true}} = 4.89$



Gibbs Sampling

break-down joint posterior into (simpler) conditional distributions

- difficult: sampling $P(\beta_0, \beta_1, \beta_2, \sigma^2 | Y)$
- easy: sampling $P(\beta_0, \beta_1, \beta_2, |\sigma^2, Y)$ then $P(\sigma^2|\beta_0, \beta_1, \beta_2, Y)$ then repeat

approximates the joint posterior

algorithm

- initialize: $\beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \sigma^{2(0)}$

$$\begin{aligned} \{\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}\} &\sim P(\beta|\sigma^{2(0)}, Y) \\ \sigma^{2(1)} &\sim P(\sigma^2|\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, Y) \\ \{\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}\} &\sim P(\beta|\sigma^{2(1)}, Y) \\ \sigma^{2(2)} &\sim P(\sigma^2|\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, Y) \end{aligned} \tag{1}$$

- repeat 1000's or 1000000 's times

BUGS to the rescue

Previously, Bayesian analysis demanded custom-coding MCMC algorithms

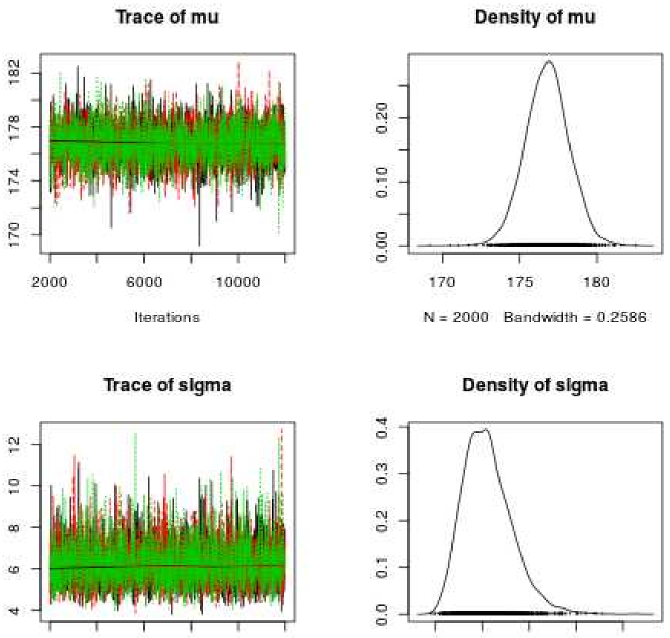
WinBUGS & OpenBUGS & JAGS

automatically use appropriate sampling techniques; so we don't have to worry

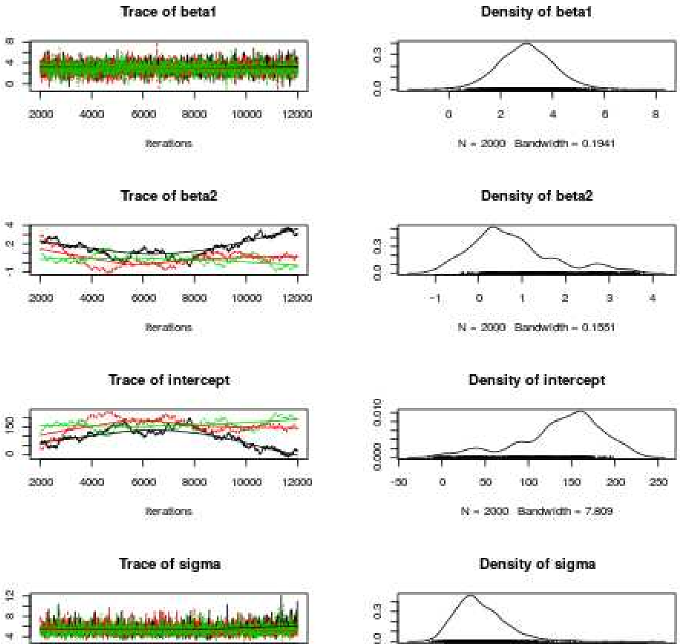
BUT you must: Monitor the MCMC!

- give reasonable **initial values**
- ensure **convergence**: no trend; independent chains give same answer
- ensure adequate **mixing**: independent samples

MCMC: Good mixing



MCMC: Poor convergence



MCMC

MCMC parameters in JAGS

- n.chains: num. of MCMC chains; more is better
- n.adapt: discard first samples; let algorithm 'adapt'
- n.burn: discard extra samples; allow algorithm to reach startionary distribution
- n.iter: total number of sample; more is better
- thin: take every k^{th} iteration for a sample; decorrelates one sample from the next; higher is better
- total samples: number of samples to approximate your Posterior; target at least 2000 to 5000

MCMC: what to do with bad mixing

- run longer chains
- ensure long enough adaption phase
- misspecified priors
- bad initial values?

Advantages of Bayesian Inference

- inference statesments: easy to understand (only Bayesians can make probabilistic statements about θ)
- small sample sizes: exact inference
- missing data: very easy to impute
- integrate other information, or calculate 'derived parameters'

Hierarchical Bayesian

- model dependences (space & time)
- "random-effects" models
- "model-selection" / "model-multi inference" (AIC, Lasso, etc., are just types of Bayesian models)
- shrinkage: deflate influence of outlier values

Where to go from here?

some Bayesian learning resources

- learn about prior distributions!
- great R package for learning the fundamentals of Gibbs sampling, MCMC, conditional probability, etc.

LearnBayes: Functions for Learning Bayesian Inference! See the Vignettes. https://cran.r-project.org/web/packages/LearnBayes/index.html

- OpenBUGS examples: read and run yourself

http://www.openbugs.net/w/Examples

- Textbook: WinBUGs for Ecologists, Marc Kery
- Blog: Andrew Gelman: http://andrewgelman.com/

Frequentism

- Excellent and accessible video lecture by Michael Jordan

http://videolectures.net/mlss09uk_jordan_bfway/