

# Markov chain Monte Carlo I

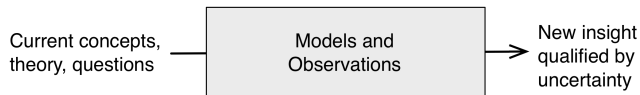
## Models for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

December 12, 2020

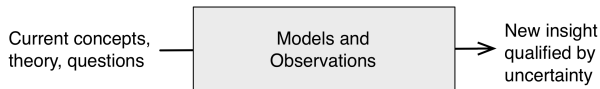


# What is this course about?

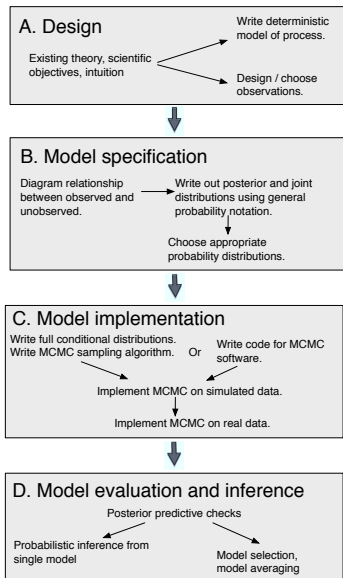


# You can understand it.

- ▶ Rules of probability
  - ▶ Conditioning and independence
  - ▶ Law of total probability
  - ▶ Factoring joint probabilities
- ▶ Distribution theory
- ▶ Markov chain Monte Carlo



# The Bayesian method



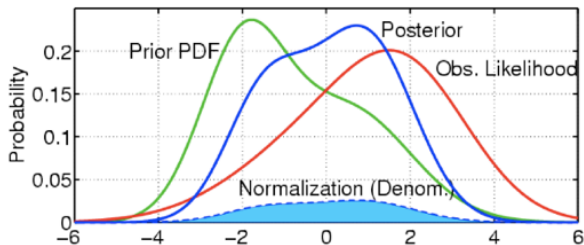
# The MCMC algorithm

- ▶ Why MCMC?
- ▶ Some intuition about how it works for a single parameter model
- ▶ MCMC for multiple parameter models
  - ▶ Full-conditional distributions (today)
  - ▶ Gibbs sampling (today)
  - ▶ Metropolis-Hastings algorithm (optional lecture notes and reading)
  - ▶ MCMC software (JAGS, today and tomorrow)

# MCMC learning outcomes

1. Develop a big picture understanding of how MCMC allows us to approximate the marginal posterior distributions of parameters and latent quantities.
2. Understand and be able to code a simple MCMC algorithm.
3. Appreciate the different methods that can be used within MCMC algorithms to make draws from the posterior distribution.
  - 3.1 Metropolis
  - 3.2 Metropolis-Hastings
  - 3.3 Gibbs
4. Understand concepts of burn-in and convergence.
5. Understand and be able to write full-conditional distributions.

# Remember the marginal distribution of the data



We have simple solutions for the posterior for simple models:

$$[\phi|y] = \text{beta} \left( \underbrace{\overbrace{\alpha}^{\text{The prior } \alpha}}_{\text{The new } \alpha} + y, \underbrace{\overbrace{\beta}^{\text{The prior } \beta}}_{\text{The new } \beta} + n - y \right)$$



# Problems of high dimension do not have simple solutions:

$$[\theta_1, \theta_2, \theta_3, \theta_4, \mathbf{z} \mid \mathbf{y}, \mathbf{u}] = \frac{\prod_{i=1}^n [y_i \mid \theta_1 z_i] [u_i \mid \theta_2, z_i] [z_i \mid \theta_3, \theta_4] [\theta_1] [\theta_2] [\theta_3] [\theta_4]}{\int \dots \int \prod_{i=1}^n [y_i \mid \theta_1 z_i] [u_i \mid \theta_2, z_i] [z_i \mid \theta_3, \theta_4] [\theta_1] [\theta_2] [\theta_3] [\theta_4] dz_i d\theta_1 d\theta_2 d\theta_3 d\theta_4}$$

# What we are doing in MCMC?

Recall that the posterior distribution is proportional to the joint: because the marginal distribution of the data  $\int [y|\theta][\theta]d\theta$  is a constant after the data have been observed.

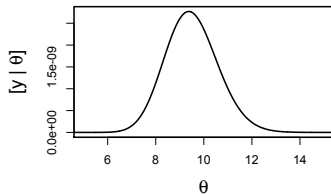
$$\overbrace{[\theta|y]}^{\text{Posterior}} \propto \overbrace{[y, \theta]}^{\text{Joint}} \quad (1)$$

$$\overbrace{[\theta|y]}^{\text{Posterior}} \propto \overbrace{[\theta|y]}^{\text{likelihood}} \overbrace{[\theta]}^{\text{prior}} \quad (2)$$

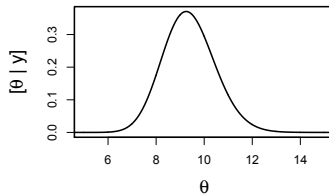
Factoring the joint distribution into a product of probability distributions using the chain rule of probability is where we start all Bayesian modeling. The factored joint distribution provides the basis for MCMC.

# What we are doing in MCMC?

**Likelihood**

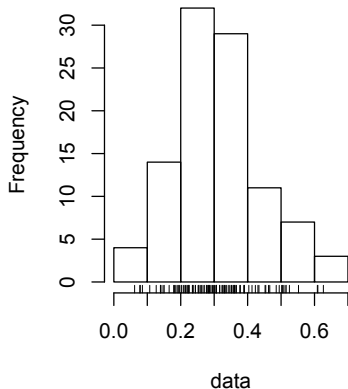


**Posterior**

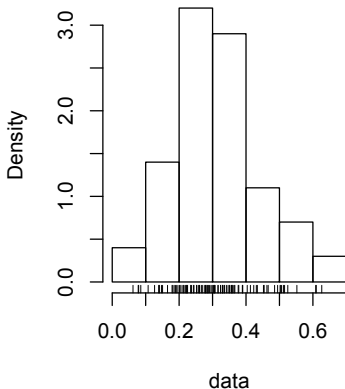


# What we are doing in MCMC?

**n=100, not normalized**



**n=100, normalized**



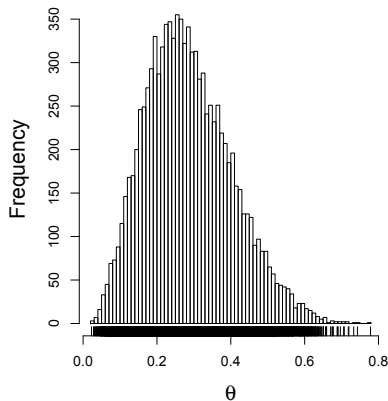
# What are we doing in MCMC?

- ▶ The posterior distribution is unknown, but the likelihood is known as a likelihood profile and we know the priors.
- ▶ We want to accumulate many, many values that represent random samples proportionate to their density in the *marginal* posterior distribution.
- ▶ MCMC generates these samples using the likelihood and the priors to decide which samples to keep and which to throw away.
- ▶ We can then use these samples to calculate statistics describing the distribution: means, medians, variances, credible intervals etc.

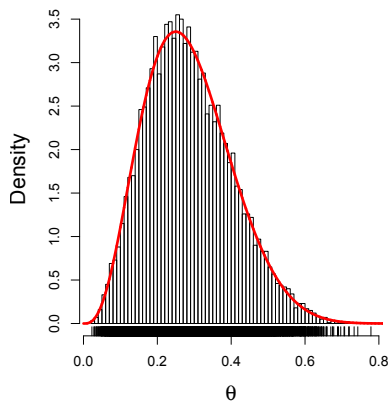
# What are we doing in MCMC?

The marginal posterior distribution of each unobserved quantity is approximated by samples accumulated in the chain.

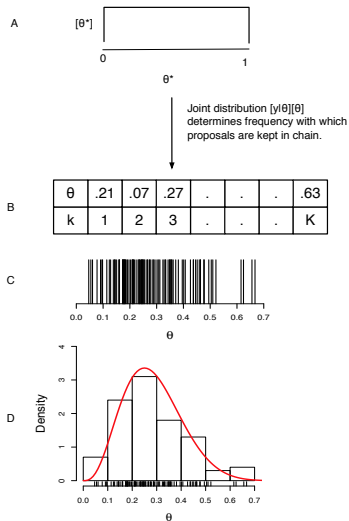
**n=100000, not normalized**



**n=100000, normalized**



# What are we doing in MCMC?



# Metropolis updates

We keep the more probable members of the posterior distribution by comparing a proposal with the current value in the chain.

$$\begin{array}{ccc}
 k & 1 & 2 \\
 \text{Proposal } \theta^{*k+1} & & \theta^{*2} \\
 \text{Test} & & P(\theta^{*2}) > P(\theta^1) \\
 \text{Chain}(\theta^k) & \theta^1 & \theta^2 = \theta^{*2}
 \end{array}$$



# Metropolis updates

We keep the more probable members of the posterior distribution by comparing a proposal with the current value in the chain.

	$k$	1	2	3
Proposal	$\theta^{*k+1}$		$\theta^{*2}$	$\theta^{*3}$
Test			$P(\theta^{*2}) > P(\theta^1)$	$P(\theta^2) > P(\theta^{*3})$
Chain( $\theta^k$ )		$\theta^1$	$\theta^2 = \theta^{*2}$	$\theta^3 = \theta^2$

# Metropolis updates

We keep the more probable members of the posterior distribution by comparing a proposal with the current value in the chain.

$k$	1	2	3	4				$K$
Proposal $\theta^{*k+1}$		$\theta^{*2}$	$\theta^{*3}$	$\theta^{*4}$	.	.	.	.
Test		$P(\theta^{*2}) > P(\theta^1)$	$P(\theta^2) > P(\theta^{*3})$	$P(\theta^3) > P(\theta^{*4})$	.	.	.	.
Chain $(\theta^k)$	$\theta^1$	$\theta^2 = \theta^{*2}$	$\theta^3 = \theta^2$	$\theta_4 = \theta_3$				

# Metropolis updates

$$\begin{aligned}
 [\theta^{*k+1}|y] &= \frac{\overbrace{[y|\theta^{*k+1}]}^{\text{likelihood}} \overbrace{[\theta^{*k+1}]}^{\text{prior}}}{\int \underbrace{[y|\theta]}_{\text{likelihood}} \underbrace{[\theta]}_{\text{prior}} d\theta} \\
 [\theta^k|y] &= \frac{\overbrace{[y|\theta^k]}^{\text{likelihood}} \overbrace{[\theta^k]}^{\text{prior}}}{\int \underbrace{[y|\theta]}_{\text{likelihood}} \underbrace{[\theta]}_{\text{prior}} d\theta} \\
 R &= \frac{[\theta^{*k+1}|y]}{[\theta^k|y]}
 \end{aligned}$$

# The cunning idea behind Metropolis updates

$$\begin{aligned}
 [\theta^{*k+1}|y] &= \frac{\overbrace{[y|\theta^{*k+1}]}^{\text{likelihood}} \overbrace{[\theta^{*k+1}]}^{\text{prior}}}{\int \underbrace{[y|\theta]}_{\text{likelihood}} \underbrace{[\theta]}_{\text{prior}} d\theta} \\
 [\theta^k|y] &= \frac{\overbrace{[y|\theta^k]}^{\text{likelihood}} \overbrace{[\theta^k]}^{\text{prior}}}{\int \underbrace{[y|\theta]}_{\text{likelihood}} \underbrace{[\theta]}_{\text{prior}} d\theta} \\
 R &= \frac{[\theta^{*k+1}|y]}{[\theta^k|y]}
 \end{aligned}$$

## When do we keep the proposal?

$$P_R = \min(1, R)$$

Keep  $\theta^{*k+1}$  as the next value in the chain with probability  $P_R$  and keep  $\theta^k$  with probability  $1 - P_R$ .

## When do we keep the proposal?

1. Calculate  $R$  based on likelihoods and priors.
2. Draw a random number,  $U$  from uniform distribution  $0,1$  If  $R > U$ , we keep the proposal  $\theta^{*k+1}$  as the next value in the chain.
3. Otherwise, we retain  $\theta^k$  as the next value.

## A simple example for one parameter

- ▶ Mary is interested in estimating the proportion of coal-fired power plants that fail to meet regulations for emissions of lead.
- ▶ She is not very ambitious, so she only checks 12 plants, 3 of which are non-compliant. She assumes there is no prior knowledge of this proportion.
- ▶ How could she calculate the parameters of the posterior distribution of the proportion of non-compliant plants on the back of a cocktail napkin?

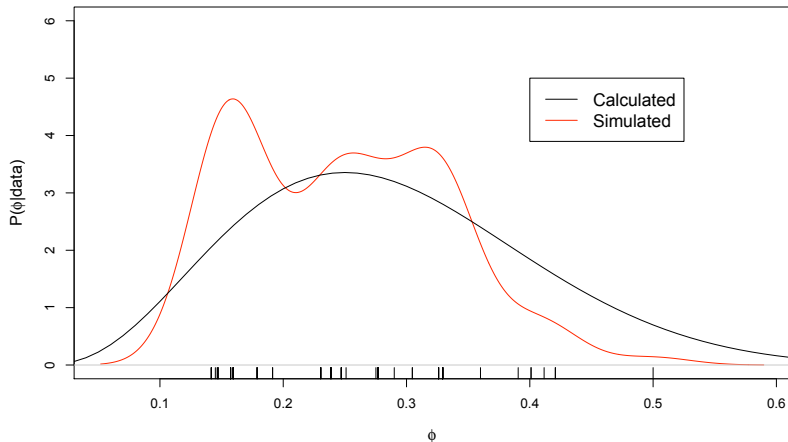
# The model

$$[\phi|y] \propto \text{binomial}(y|n, \phi) \text{beta}(\phi|1, 1)$$



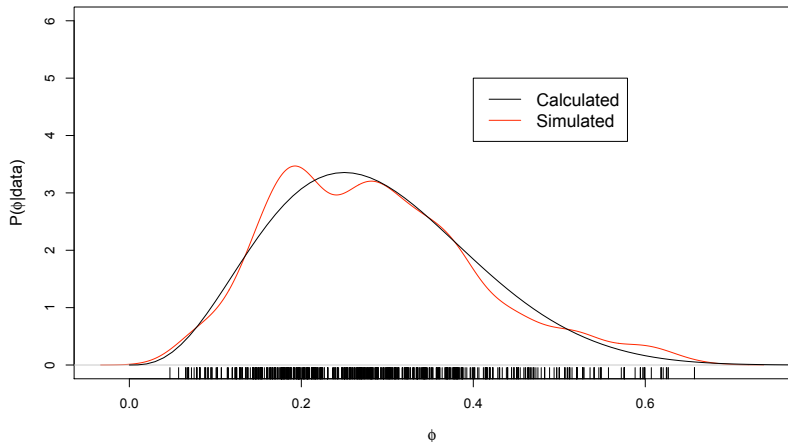
# Sampling from the posterior

**Simulated and Calculated Distribution, iterations = 100**



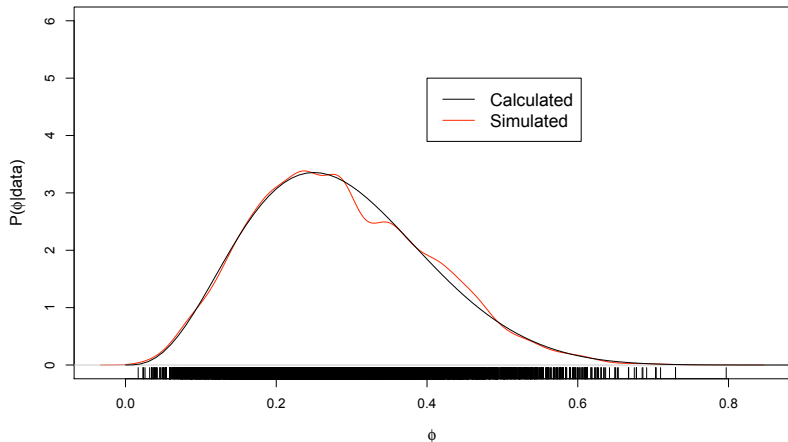
# Sampling from the posterior

Simulated and Calculated Distribution, iterations = 1000



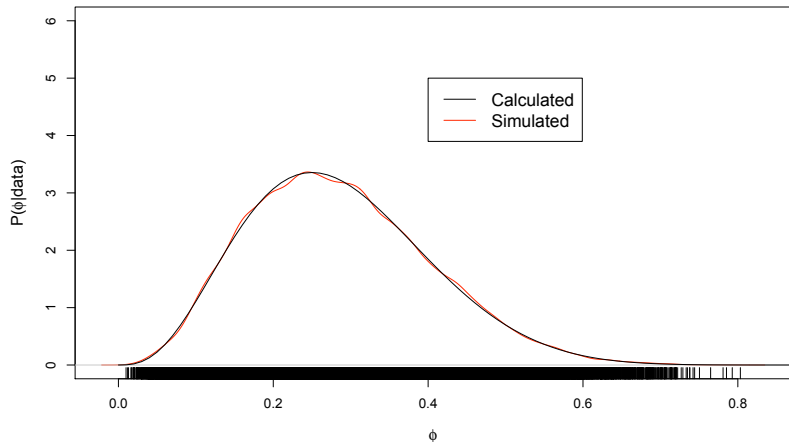
# Sampling from the posterior

**Simulated and Calculated Distribution, iterations = 10000**

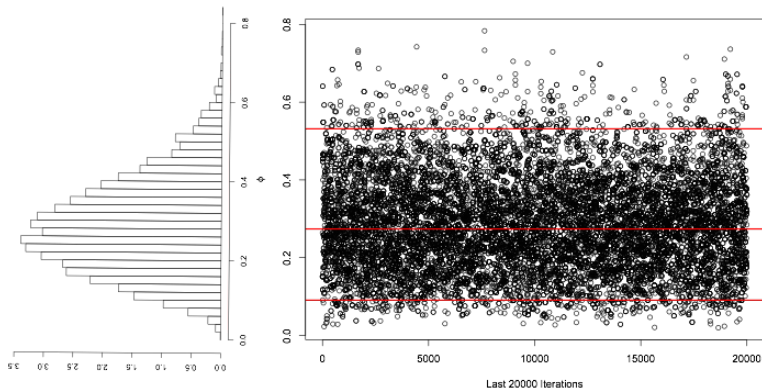


# Sampling from the posterior

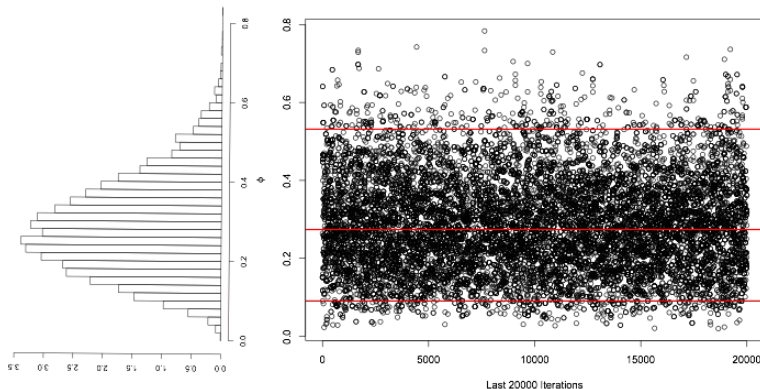
**Simulated and Calculated Distribution, iterations = 100000**



# Sampling from the posterior

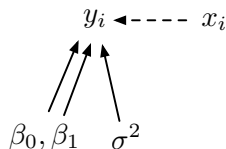


# Sampling from the posterior



The chain has *converged* when adding more samples does not change the shape of the posterior distribution. We throw away samples that are accumulated before convergence (burn-in).

# Intuition for MCMC for multi-parameter models



$$g(\beta_0, \beta_1, x_i) = \beta_0 + \beta_1 x_i$$

$$[\beta_0, \beta_1, \sigma^2 \mid y_i] \propto [\beta_0, \beta_1, \sigma^2, y_i]$$

factoring rhs using DAG:

$$[\beta_0, \beta_1, \sigma^2 \mid y_i] \propto [y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2][\beta_0][\beta_1][\sigma^2]$$

joint for all data :

$$[\beta_0, \beta_1, \sigma^2 \mid \mathbf{y}] \propto \prod_{i=1}^n [y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2][\beta_0][\beta_1][\sigma^2]$$

choose specific distributions:

$$\begin{aligned} [\beta_0, \beta_1, \sigma^2 \mid \mathbf{y}] &\propto \prod_{i=1}^n \text{normal}(y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2) \\ &\times \text{normal}(\beta_0 \mid 0, 10000) \text{normal}(\beta_1 \mid 0, 10000) \\ &\times \text{uniform}(\sigma^2 \mid 0, 500) \end{aligned}$$

# Intuition for MCMC for multi-parameter models

$$\begin{aligned}
 [\beta_0, \beta_1, \sigma^2 \mid \mathbf{y}] &\propto \prod_{i=1}^n \text{normal}(y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2) \\
 &\times \text{normal}(\beta_0 \mid 0, 10000) \text{normal}(\beta_1 \mid 0, 10000) \\
 &\times \text{uniform}(\sigma^2 \mid 0, 100)
 \end{aligned}$$

1. Set initial values for  $\beta_0, \beta_1, \sigma^2$
2. Assume  $\beta_1, \sigma^2$  are known and constant. Make a draw for  $\beta_0$ . Store the draw.
3. Assume  $\beta_0, \sigma^2$  are known and constant. Make a draw for  $\beta_1$ . Store the draw.
4. Assume  $\beta_0, \beta_1$  are known and constant. Make a draw for  $\sigma^2$ . Store the draw.
5. Do this many times. The stored values for each parameter approximate its marginal posterior distribution after convergence.



# Implementing MCMC for multiple parameters

- ▶ Write an expression for the posterior and joint distribution using a DAG as a guide. Always.
- ▶ If you are using MCMC software (e.g. JAGS) use expression for the posterior and joint distribution as template for writing code. You are done.
- ▶ If you are writing your own MCMC sampler *or* you simply want to understand what JAGS is doing for you:
  - ▶ Decompose the expression of the multivariate joint distribution into a series of univariate distributions called *full-conditional distributions*.
  - ▶ Choose a sampling method for each full-conditional distribution.
  - ▶ Cycle through each unobserved quantity, sampling from its full-conditional distribution, treating the others as if they were known and constant.
  - ▶ The accumulated samples approximate the marginal posterior distribution of each unobserved quantity.
  - ▶ Note that this takes a complex, multivariate problem and turns it into a series of simple, univariate problems that we solve, as in the example above, one at a time.

## Definition of full-conditional distribution

Let  $\boldsymbol{\theta}$  be a vector of length  $k$  containing all of the unobserved quantities we seek to understand. Let  $\boldsymbol{\theta}_{-j}$  be a vector of length  $k - 1$  that contains all of the unobserved quantities *except*  $\theta_j$ . The full-conditional distribution of  $\theta_j$  is

$$[\theta_j | y, \boldsymbol{\theta}_{-j}],$$

which we notate as

$$[\theta_j | \cdot].$$

It is the posterior distribution of  $\theta_j$  conditional on all of the other parameters and the data, which we assume are *known*.

## Writing full-conditional distributions

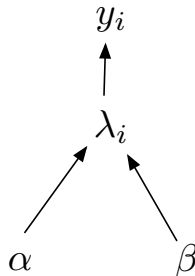
- ▶ You will have one full-conditional for each unobserved quantity in the posterior.
- ▶ For each unobserved quantity, write the distributions where it appears.
- ▶ Ignore the other distributions.
- ▶ Simple.

## Example

- ▶ Clark 2003 considered the problem of modeling fecundity of spotted owls and the implication of individual variation in fecundity for population growth rate.
- ▶ Data were number of offspring produced by per pair of owls with sample size  $n = 197$ .

Clark, J. S. 2003. Uncertainty and variability in demography and population growth: A hierarchical approach. *Ecology* 84:1370-1381.

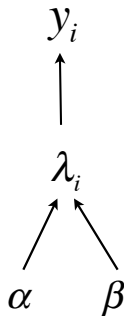
# Example



$$\begin{aligned}
 [\boldsymbol{\lambda}, \alpha, \beta | \mathbf{y}] &\propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \\
 &\times \text{gamma}(\alpha | .001, .001) \text{gamma}(\beta | .001, .001)
 \end{aligned}$$

# Full-conditionals

$$[\boldsymbol{\lambda}, \alpha, \beta | \mathbf{y}] \propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001) \text{gamma}(\alpha | .001, .001)$$



We use the multivariate joint distribution to find univariate full-conditionals for all unobserved quantities.

How many full conditionals are there?

# Writing full-conditional distributions

- ▶ You will have one full-conditional for each unobserved quantity in the posterior.
- ▶ For each unobserved quantity, write the distributions (including products) where it appears.
- ▶ Ignore the other distributions.
- ▶ Simple.

## Full-conditional for each $\lambda_i$

$$[\lambda, \alpha, \beta | \mathbf{y}] \propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001) \text{gamma}(\alpha | .001, .001)$$

Writing the full-conditional distribution for  $\lambda_i$ :

$$[\lambda_i | .] \propto \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta)$$



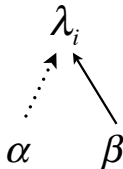


## Full-conditional for $\beta$

$$[\lambda, \alpha, \beta | \mathbf{y}] \propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001) \text{gamma}(\alpha | .001, .001)$$

Writing the full-conditional distribution for  $\beta$ :

$$[\beta | \cdot] \propto \prod_{i=1}^n \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001)$$

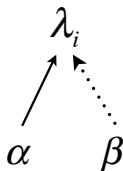


## Full-conditional for $\alpha$

$$[\lambda, \alpha, \beta | \mathbf{y}] \propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001) \text{gamma}(\alpha | .001, .001)$$

Writing the full-conditional distribution for  $\alpha$ :

$$[\alpha | \cdot] \propto \prod_{i=1}^n \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\alpha | .001, .001)$$



## Full-conditionals for the model

Posterior and joint:

$$\begin{aligned}
 [\boldsymbol{\lambda}, \alpha, \beta | \mathbf{y}] &\propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \\
 &\times \text{gamma}(\alpha | .001, .001) \text{gamma}(\beta | .001, .001)
 \end{aligned}$$

Full conditionals:

$$[\lambda_i | \cdot] \propto \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta)$$

$$[\beta | \cdot] \propto \prod_{i=1}^n \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001)$$

$$[\alpha | \cdot] \propto \prod_{i=1}^n \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\alpha | .001, .001)$$

# Implementing MCMC for multiple parameters and latent quantities

- ▶ Write an expression for the posterior and joint distribution using a DAG as a guide. Always.
- ▶ If you are using MCMC software (e.g. JAGS) use expression for posterior and joint as template for writing code.
- ▶ If you are writing your own MCMC sampler:
  - ▶ Decompose the expression of the multivariate joint distribution into a series of univariate distributions called *full-conditional distributions*.
  - ▶ Choose a sampling method for each full-conditional distribution.
  - ▶ Cycle through each unobserved quantity, sampling from the its full-conditional distribution, treating the others as if they were known and constant.
  - ▶ Note that this takes a complex, multivariate problem and turns it into a series of simple, univariate problems that we solve, as in the example above, one at a time.

# Choosing a sampling method

1. Accept-reject:
  - 1.1 Metropolis: requires a symmetric proposal distribution (e.g., normal, uniform). This is what we used above in the example for one parameter.
  - 1.2 Metropolis-Hastings: allows asymmetric proposal distributions (e.g., beta, gamma, lognormal). A minor modification of Metropolis. See optional notes.
2. Gibbs: accepts all proposals because they are especially well chosen. Requires conjugates. In lab today.

## Why do you need to understand conjugate priors?

- ▶ A easy way to find parameters of posterior distributions for simple problems.
- ▶ Critical to understanding Gibbs updates in Markov chain Monte Carlo as you are about to learn.

# What are conjugate priors?

Assume we have a likelihood and a prior:

$$\overbrace{[\theta|y]}^{\text{posterior}} = \frac{\overbrace{[y|\theta]}^{\text{likelihood}} \overbrace{[\theta]}^{\text{prior}}}{[y]}.$$

If the form of the distribution of the posterior

$$[\theta|y]$$

is the same as the form of the distribution of the prior,

$$[\theta]$$

then the likelihood and the prior are said to be conjugates

$$\underbrace{[y|\theta][\theta]}$$

conjugates

and the prior is called a conjugate prior for  $\theta$ .

## Gibbs updates

When priors and likelihoods are conjugate, we *know* all but one of the parameters of the full-conditional because they are *assumed to be known* at each iteration. We make a draw of the single unknown *directly* from its posterior distribution as if the other parameters were fixed.

Wickedly clever.



# Gamma-Poisson conjugate relationship for $\lambda$

The conjugate prior distribution for a Poisson likelihood is  $\text{gamma}(\lambda|\alpha, \beta)$ . Given  $n$  observations  $y_i$  of new data, the posterior distribution of  $\lambda$  is

$$[\lambda|\mathbf{y}] = \text{gamma} \left( \underbrace{\overbrace{\alpha_0}^{\text{The prior } \alpha} + \sum_{i=1}^n y_i}_{\text{The new } \alpha}, \underbrace{\overbrace{\beta_0}^{\text{The prior } \beta} + n}_{\text{The new } \beta} \right). \quad (3)$$

## Sampling from the Poisson-gamma conjugate:

Full conditional:

$$[\lambda_i | \cdot] \propto \text{Poisson}(y_i | \lambda_i) \text{gamma}(\lambda_i | \alpha, \beta) \quad (4)$$

Gibbs sample:

$$[\lambda_i^k | y_i] = \text{gamma} \left( \overbrace{\alpha^{k-1}}^{\text{The current } \alpha} + y_i, \overbrace{\beta^{k-1}}^{\text{The current } \beta} + 1 \right). \quad (5)$$

In R, this would be:

```
shape[k] = alpha[k-1] + y_i
rate[k] = beta[k-1] + 1
lambda[k,i] = rgamma(1, shape[k], rate[k])
```

## Gamma-gamma conjugate relationship

The conjugate prior distribution for the  $\beta$  parameter (rate) in a gamma likelihood  $\text{gamma}(y_i | \alpha, \beta)$  is a gamma distribution  $\text{gamma}(\beta | \alpha_0, \beta_0)$ . Given  $n$  observations  $y_i$  of new data, the posterior distribution of  $\beta$  (assuming that  $\alpha$  (shape) is known) is given by:

$$[\beta | \mathbf{y}] = \text{gamma} \left( \underbrace{\underbrace{\alpha_0}_{\text{The prior } \alpha} + n\alpha}_{\text{The new } \alpha}, \underbrace{\underbrace{\beta_0}_{\text{The prior } \beta} + \sum_{i=1}^n y_i}_{\text{The new } \beta} \right). \quad (6)$$

We can substitute any “known” quantity for  $\mathbf{y}$ , e.g.,  $\boldsymbol{\lambda}$  in MCMC.

## Sampling from the gamma-gamma conjugate:

The full conditional:

$$[\beta|.] \propto \prod_{i=1}^n \text{gamma}(\lambda_i | \alpha, \beta) \text{gamma}(\beta | .001, .001)$$

Gibbs sample:

$$\beta^k \sim \text{gamma}\left(.001 + n\alpha^{k-1}, .001 + \sum_{i=1}^n \lambda_i^k\right)$$

In R this would be:

```
shape[k] = .001 + length(y) * alpha[k-1]
rate[k] = .001 + sum(lambda[,k])
beta[k] = rgamma(1, shape[k], rate[k])
```

# MCMC algorithm

1. Iterate over  $i = 1 \dots 197$
2. At each  $i$ , make a draw from

$$\lambda_i^k \sim \underbrace{\text{gamma}(\alpha^{k-1} + y_i, \beta^{k-1} + 1)}_{\text{Gibbs update using gamma - Poisson conjugate for each } \lambda_i} \quad (7)$$

$$\beta^k \sim \underbrace{\text{gamma}(.001 + \alpha^{k-1} n, .001 + \sum_{i=1}^n \lambda_i^k)}_{\text{Gibbs update using gamma - gamma conjugate for } \beta} \quad (8)$$

$$\alpha^k \propto \underbrace{\prod_{i=1}^n \text{gamma}(\lambda_i^k | \alpha^{k-1}, \beta^k) \text{gamma}(\alpha^{k-1} | .001, .001)}_{\text{No conjugate for } \alpha. \text{ Use Metropolis - Hastings update}} \quad (9)$$

3. Repeat for  $k = 1 \dots K$  iterations, storing  $\lambda_i^k, \alpha^k$  and  $\beta^k$ . Store the value of each parameter at each iteration in a vector.

# Inference from MCMC

Make inference on each unobserved quantity using the elements of their vectors stored after convergence. These post-convergence vectors, (i.e., the “rug” described above) approximate the marginal posterior distributions of unobserved quantities.