Markov chain Monte Carlo I

Bayesian Modeling for Socio-Environmental Data

N. Thompson Hobbs

August 4, 2016

SESYNC

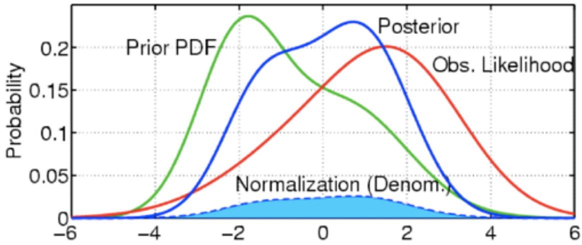
The MCMC algorithm

- Some intuition
- Accept-reject sampling with Metropolis algorithm
- Introduction to full-conditional distributions
- Gibbs sampling (exercise)
- Metropolis-Hastings algorithm
- Implementing accept-reject sampling

MCMC learning outcomes

1. Develop a big picture understanding of how MCMC allows us to approximate the marginal posterior distribution for 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. Be able to write full-conditional distributions.

Remember the marginal distribution of the data



We have simple solutions for the posterior for simple models:

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

Problems of high dimension do not have simple solutions:

$$[\theta_1, \theta_2, \theta_3, \theta_4, z_i, y_i, u_i] = \frac{[y_i|\theta_1 z_i][u_i|\theta_2, z_i][z_i|\theta_3, \theta_4][\theta_1][\theta_2][\theta_3][\theta_4]}{\int\int\int\int [y_i|\theta_1 z_i][u_i|\theta_2, z_i][z_i|\theta_3, \theta_4][\theta_1][\theta_2][\theta_3][\theta_4]\, d\theta_1\, d\theta_2\, d\theta_3\, d\theta_4}$$

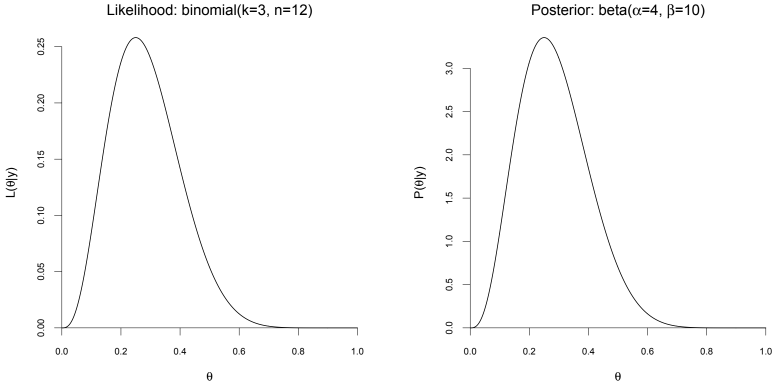
What we are doing in MCMC?

Recall that the posterior distribution is proportional to the joint:

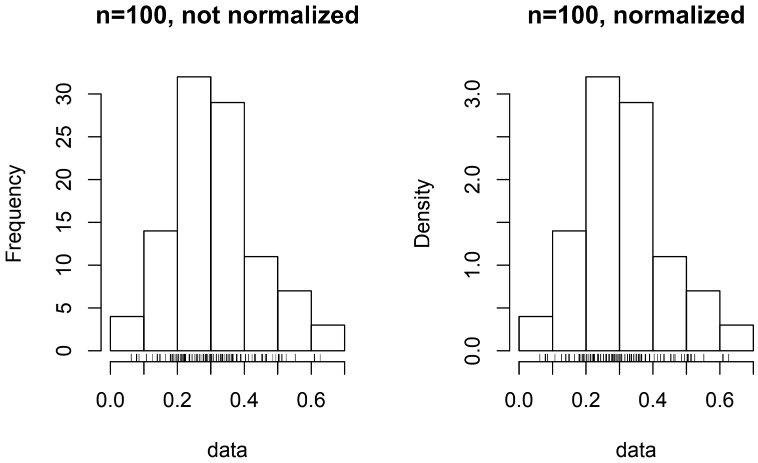
$$[\boldsymbol{\theta}|y] \propto [y|\boldsymbol{\theta}][\boldsymbol{\theta}], \tag{1}$$

because the marginal distribution of the data $\int [y|\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{\theta}$ is a constant after the data have been observed.

What we are doing in MCMC?



What we are doing in MCMC?

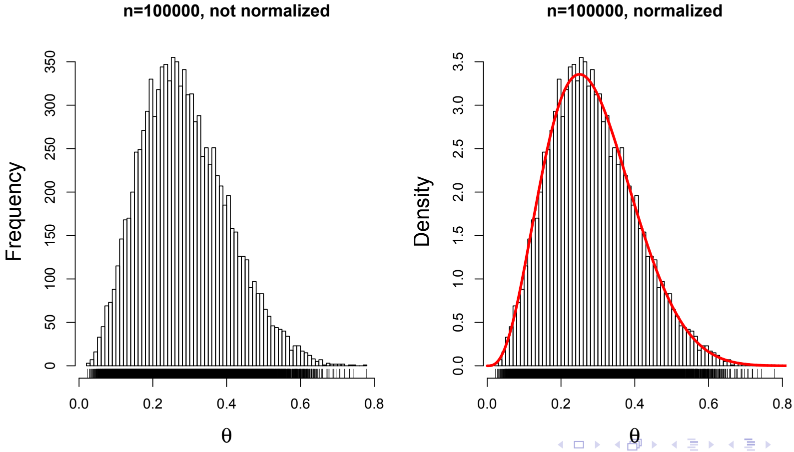


What are we doing in MCMC?

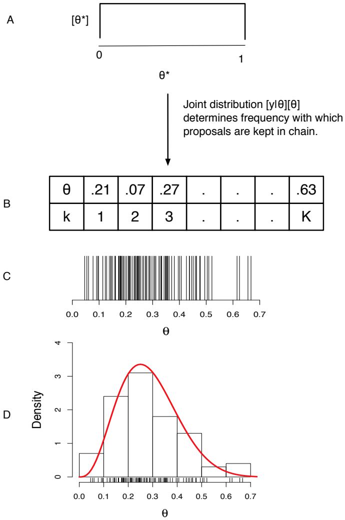
- The posterior distribution is unknown, but the likelihood is known as a likelihood profile and we know the priors.
- We want to accumulate many, many values that represent random samples proportionate to their density in the posterior distribution.
- MCMC generates these samples using the likelihood 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.



What are we doing in MCMC?



Algorithms for drawing samples from the posterior

1. Accept-reject samplers
   1.1 Metropolis: requires a symmetric proposal distribution (e.g., normal, uniform)
   1.2 Metropolis-Hastings: allows asymmetric proposal distributions (e.g., beta, gamma, lognormal)
2. Gibbs: accepts all proposals because they come directly from the posterior using conjugates.

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

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

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
Proposal $\boldsymbol{\theta}^{*\,k+1}$		$\boldsymbol{\theta}^{*\,2}$	$\boldsymbol{\theta}^{*\,3}$	$\boldsymbol{\theta}^{*\,4}$
Test		$P(\boldsymbol{\theta}^{*\,2}) > P\left(\boldsymbol{\theta}^{1}\right)$	$P(\boldsymbol{\theta}^{2}) > P\left(\boldsymbol{\theta}^{*\,3}\right)$	$P(\boldsymbol{\theta}^{3}) > P\left(\boldsymbol{\theta}^{*\,4}\right.$
Chain$(\boldsymbol{\theta}^{k})$	$\boldsymbol{\theta}^{1}$	$\boldsymbol{\theta}^{2} = \boldsymbol{\theta}^{*\,2}$	$\boldsymbol{\theta}^{3} = \boldsymbol{\theta}^{2}$	$\boldsymbol{\theta}_{4} = \boldsymbol{\theta}_{3}$

Metropolis Updates

$$[\boldsymbol{\theta}^{*k+1}|y] = \frac{\overbrace{[y|\boldsymbol{\theta}^{*k+1}]}^{\text{likelihood}}\overbrace{[\boldsymbol{\theta}^{*k+1}]}^{\text{prior}}}{\int [y|\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{\theta}}$$

$$[\boldsymbol{\theta}^{k}|y] = \frac{\overbrace{[y|\boldsymbol{\theta}^{k}]}^{\text{likelihood}}\ \overbrace{[\boldsymbol{\theta}^{k}]}^{\text{prior}}}{\int [y|\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{\theta}}$$

$$R = \frac{[\boldsymbol{\theta}^{*k+1}|y]}{[\boldsymbol{\theta}^{k}|y]}$$

Metropolis Updates

$$[\theta^{*k+1}|y] = \frac{\overbrace{[y|\theta^{*k+1}]}^{\text{likelihood}}\overbrace{[\theta^{*k+1}]}^{\text{prior}}}{\cancel{\int [y|\theta][\theta]d\theta}}$$

$$[\theta^{k}|y] = \frac{\overbrace{[y|\theta^{k}]}^{\text{likelihood}}\overbrace{[\theta^{k}]}^{\text{prior}}}{\cancel{\int [y|\theta][\theta]d\theta}}$$

$$R = \frac{[\theta^{*k+1}|y]}{[\theta^{k}|y]}$$

When do we keep the proposal?

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

Keep θ^{*k+1} as the next value in the chain with probability P_R and keep θ^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 θ^{*k+1} as the next value in the chain.
3. Otherwise, we retain θ^{k} as the next value.

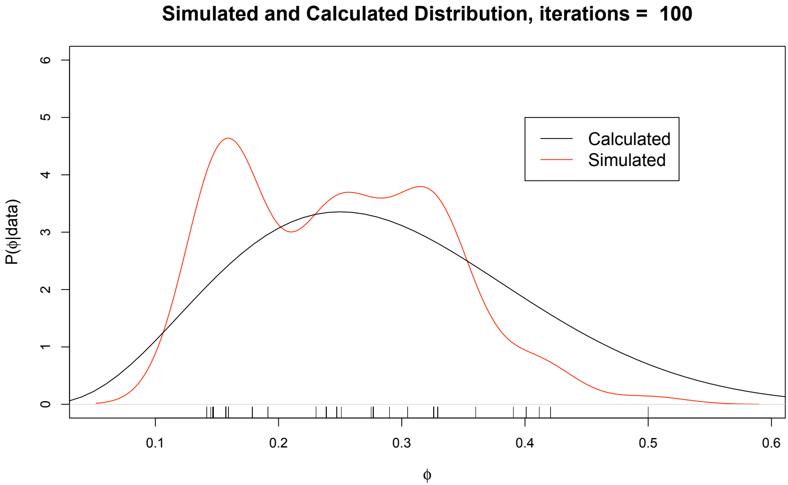
A simple example for one parameter

- Grace is interested in estimating the prevalence of *Chytrid* fungus in a population of frogs.
- She is sort of lazy, so she only samples 12 of them, of which 3 have the fungus.
- What is her best estimate of prevalence?
- How would she calculate the parameters of the posterior on the back of a cocktail napkin?

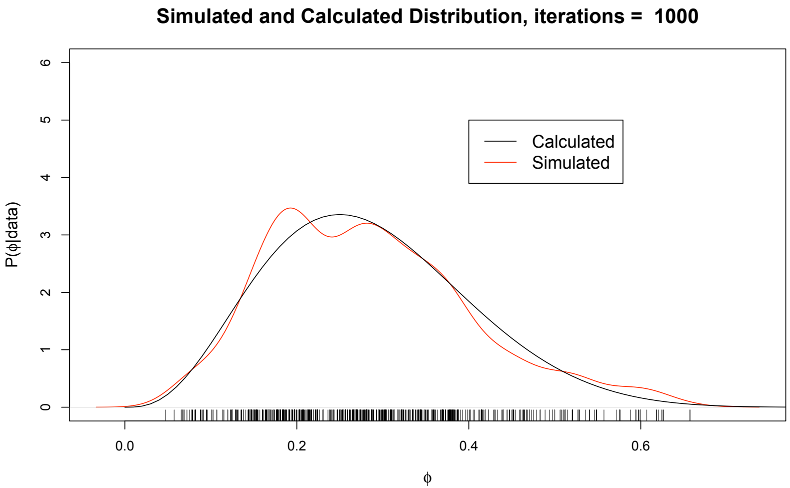
The model

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

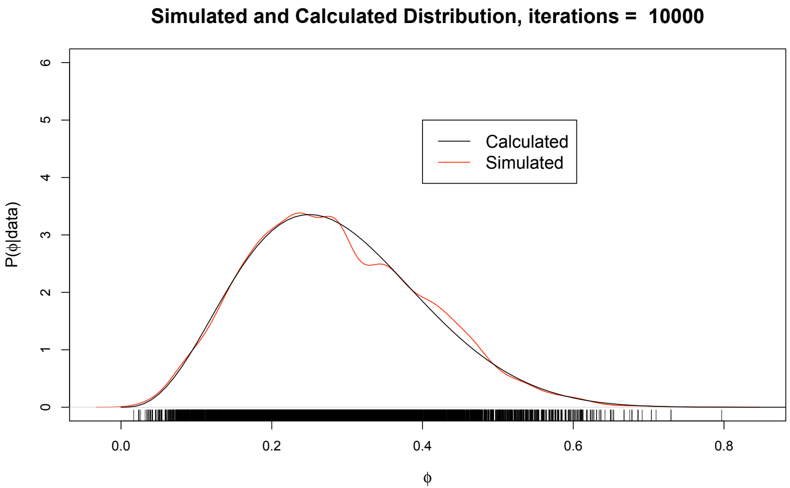
Simulation


Simulated and Calculated Distribution, iterations = 100
Calculated
Simulated
P(ϕ|data)
0
1
2
3
4
5
6
0.1
0.2
0.3
0.4
0.5
0.6
ϕ

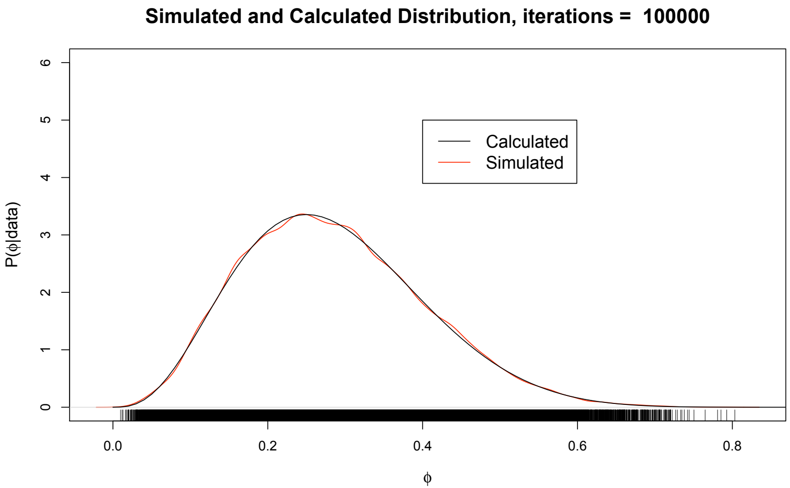
Simulation


Simulated and Calculated Distribution, iterations = 1000
Calculated
Simulated
P(ϕ|data)
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
ϕ

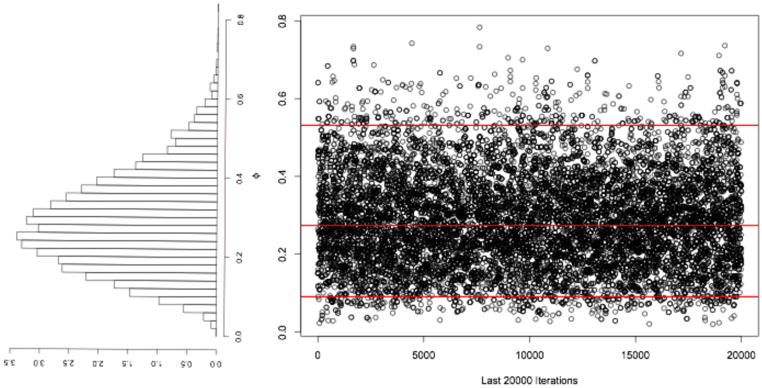
Simulation


Simulated and Calculated Distribution, iterations = 10000
Calculated
Simulated
P(ϕ|data)
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
ϕ

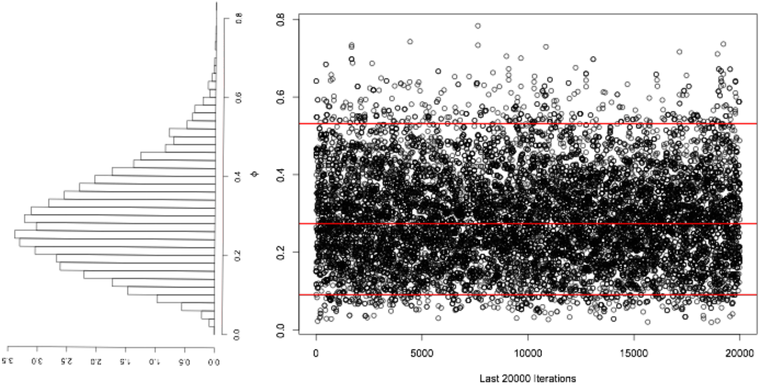
Simulation


Simulated and Calculated Distribution, iterations = 100000
Calculated
Simulated
P(ϕ|data)
ϕ
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6

Simulation


0.0
0.2
0.4
0.6
0.8
φ
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
5000
10000
15000
20000
Last 20000 Iterations

Simulation



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).

Multiple parameters and latent quantities

- We write out an expression for the posterior and joint distribution using a DAG as a guide.
- We decompose the expression of the multivariate joint distribution into a series of univariate distributions called *full-conditional distributions*. We choose a sampling method for each one.
- We then 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 problem and turns it into a series of simple problems that we solve, as in the example above, one at a time.

Multiple parameters and latent quantities

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* θ_j. The full-conditional distribution of θ_j is

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

which we notate as

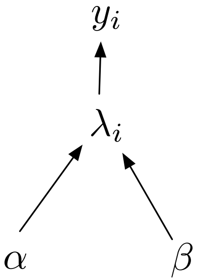
$$[\theta_j|\cdot].$$

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

Example

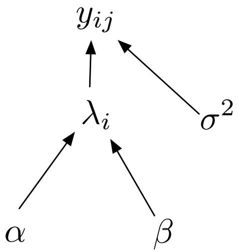
- Clark 2003 considered the problem of modeling fecundity of spotted owls and the implication of individual variation in fecundity for population growth rate.
- Data were number of offspring produced by per pair of owls, sample size = 197.

Example

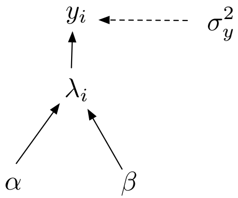


$$
\begin{aligned}
[\boldsymbol{\lambda}, \alpha, \beta | \mathbf{y}] &\propto \prod_{i=1}^{n} \text{Poisson}\left(y_i | \lambda_i\right) \text{gamma}\left(\lambda_i | \alpha, \beta\right) \\
&\times \text{gamma}\left(\alpha | .001, .001\right) \text{gamma}\left(\beta | .001, .001\right)
\end{aligned}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

A Bayesian network digression



A Bayesian network digression



Example

Posterior and joint:

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

Full conditionals:

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

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

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

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

Using a Gibbs sampler, we can estimate the posterior distribution of each unobserved quantity based on the densities in which it appears :

$$[\lambda_i|\cdot] \propto \underbrace{\text{gamma}(.001+y_i,.001+1)}_{\text{Gibbs step using gamma - Poisson conjugate for } each\lambda_i}$$

$$[\beta|\cdot] \propto \underbrace{\text{gamma}(.001+\alpha n,.001+\sum_{i=1}^{n}\lambda_i)}_{\text{Gibbs step using gamma - gamma conjugate for } \beta}$$

$$[\alpha|\cdot] \propto \underbrace{\prod_{i=1}^{n}\text{gamma}(\lambda_i|\alpha,\beta)\,\text{gamma}(\alpha|.001,.001)}_{\text{No conguate for } \alpha\text{. Use Metropolis - Hastings update}}$$