GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lecture 5: Markov Chain Monte Carlo (MCMC)

Hai Shu, PhD

10/11/2022

Topics

- Bayesian inference for normal data with a semi-conjugate prior distribution
- Bayesian inference for non-closed form posterior distribution
  - Discrete grid approximation
  - Rejection sampling
  - Importance sampling
- Bayesian inference for normal data with a semi-conjugate prior distribution via MCMC
  - Sampling from the conditional distributions
  - Gibbs sampling
  - Markov Chain Monte Carlo (MCMC)
- Monte Carlo vs. Markov chain Monte Carlo (MCMC)
- MCMC diagnostics

Normal data

- We have looked at Bayesian inference for data $y_1, \ldots, y_n$ such that $y_1, \ldots, y_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ in different cases:
  - when the variance σ^2 is assumed known and $p(\mu) = N(\mu_0, \tau_0^2)$ $\longrightarrow p(\mu | y_1, \ldots, y_n, \sigma^2)$
  - when the mean μ is assumed known and $p(\sigma^2) = \text{Inverse Gamma}(a, b)$ $\longrightarrow p(\sigma^2 | y_1, \ldots, y_n, \mu)$
  - when both μ and σ^2 are unknown $\longrightarrow p(\mu, \sigma^2 | y_1, \ldots, y_n)$
    1. noninformative and improper prior $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$
    2. conjugate prior

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu | \sigma^2) \cdot p(\sigma^2) \\ &= N(\mu_0, \tfrac{\sigma^2}{\kappa_0}) \cdot \text{Inverse Gamma}(\tfrac{\nu_0}{2}, \tfrac{\nu_0 \sigma_0^2}{2}) \\ &= N(\mu_0, \tfrac{\sigma^2}{\kappa_0}) \cdot \text{Inverse } \chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

Normal data

- We have seen that if $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$, we can approximate the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ via Monte Carlo methods by sampling from
  - the marginal posterior distribution
    $p(\sigma^2 | y_1, \ldots, y_n) = \text{InverseGamma}(\frac{(n-1)}{2}, \frac{(n-1)s^2}{2}) = \text{Inverse } \chi^2(n-1, s^2)$
  - the conditional posterior distribution
    $p(\mu | \sigma^2, y_1, \ldots, y_n) = N(\bar{y}, \frac{\sigma^2}{n})$
- We have seen that if $p(\mu, \sigma^2) = N(\mu_0, \frac{\sigma^2}{\kappa_0}) \cdot \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$, we can approximate the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ via Monte Carlo methods by sampling from
  - the marginal posterior distribution
    $p(\sigma^2 | y_1, \ldots, y_n) = \text{InverseGamma}(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}) = \text{Inverse } \chi^2(\nu_n, \sigma_n^2)$
  - the conditional posterior distribution
    $p(\mu | \sigma^2, y_1, \ldots, y_n) = N(\mu_n, \frac{\sigma^2}{\kappa_n})$

Normal data: Semi-conjugate prior

- Now we consider the case where the prior distribution for μ and σ^2 is specified assuming that μ and σ^2 are independent, that is:

$$p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2)$$

  where $p(\mu) = N(\mu_0, \tau_0^2)$ and $p(\sigma^2) = \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$.
- The prior independence of μ on σ^2 makes sense in situations where the prior distribution of μ is NOT obtained from a fixed number of prior observations from the population with variance σ^2.
- Example: Suppose μ is the unknown weight of a particular student. The data $y_1, \ldots, y_n$ are weighins on a particular scale with unknown variance σ^2.
  The prior information consists of a visual inspection: the student looks to weigh about 150 pounds with a subjective 95% probability that the weight is in the range $[150 \pm 20]$.
  Then, we can express our prior on μ as:
  $p(\mu) = N(\mu_0 = 150, \tau_0^2 = 10^2)$.

Normal data: Semi-conjugate prior

- The joint prior distribution is:

$$p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2)$$

$$\propto \frac{1}{\tau_0} \exp\left(-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right) \cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right)$$

and is not conjugate. Infact, the likelihood is:

$$p(y_1, \ldots, y_n | \mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right)$$

and the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ is:

$$p(\mu, \sigma^2 | y_1, \ldots, y_n) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right)$$

$$\cdot \frac{1}{\tau_0} \exp\left(-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right) \cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right)$$

Normal data: Semi-conjugate prior

- The joint posterior distribution does not follow any standard parametric form.
- We can proceed as before: we can approximate the joint posterior distribution by simulating from it using the decomposition of the joint posterior distribution in

$$p(\mu, \sigma^2 | y_1, \ldots, y_n) = p(\mu | \sigma^2, y_1, \ldots, y_n) \cdot p(\sigma^2 | y_1, \ldots, y_n)$$

that is, by considering

- the conditional posterior distribution $p(\mu | \sigma^2, y_1, \ldots, y_n)$ of μ given σ^2 and the data $y_1, \ldots, y_n$
- the marginal posterior distribution $p(\sigma^2 | y_1, \ldots, y_n)$ of σ^2 given the data $y_1, \ldots, y_n$.

Normal data: Semi-conjugate prior

- Since

$$
\begin{aligned}
p(\mu, \sigma^2 | y_1, \ldots, y_n) &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right) \\
&\cdot \frac{1}{\tau_0} \exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right) \cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)
\end{aligned}
$$

the conditional posterior distribution of μ is:

$$
\begin{aligned}
p(\mu | \sigma^2, y_1, \ldots, y_n) &\propto \exp\left(-\frac{1}{2\sigma^2} n(\bar{y}-\mu)^2\right) \cdot \exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\frac{(\bar{y}-\mu)^2}{\frac{\sigma^2}{n}} + \frac{(\mu-\mu_0)^2}{\tau_0^2}\right]\right)
\end{aligned}
$$

Normal data: Semi-conjugate prior

- We have already seen (Lecture 3) what is the conditional posterior distribution of μ when we assume that σ^2 is known and $p(\mu) = N(\mu_0, \tau_0^2)$:

$$p(\mu|\sigma^2, y_1, \ldots, y_n) \sim N(\mu_n, \tau_n^2)$$

where

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \qquad \mu_n = \tau_n^2 \cdot \left(\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma^2} \bar{y} \right)$$

- So, we have found that if $y_1, \ldots, y_n|\mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ and $p(\mu, \sigma^2) = N(\mu; \mu_0; \tau_0^2) \cdot \text{InverseGamma}(\sigma^2; \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$, then:

$$p(\mu|\sigma^2, y_1, \ldots, y_n) \sim N(\mu_n, \tau_n^2)$$

Normal data: Semi-conjugate prior

- Now, we need to derive the marginal posterior distribution $p(\sigma^2|y_1,\ldots,y_n)$. This is derived by integrating out μ from the joint posterior distribution $p(\mu,\sigma^2|y_1,\ldots,y_n)$, that is:

$$\begin{aligned} p(\sigma^2|y_1,\ldots,y_n) &= \int_{-\infty}^{\infty} p(\mu,\sigma^2|y_1,\ldots,y_n)d\mu \\ &\propto \int_{-\infty}^{\infty} p(y_1,\ldots,y_n|\mu,\sigma^2)p(\mu,\sigma^2)d\mu \\ &\propto \int_{-\infty}^{\infty} \Big\{ \left[\prod_i N(y_i;\mu,\sigma^2)\right] \cdot N(\mu;\mu_0,\tau_0^2) \\ &\cdot \quad \text{Inverse Gamma}(\sigma^2;\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2}) \Big\} d\mu \end{aligned}$$

- This integral can be computed in closed form because the integrand, considered as a function of μ, is proportional to a normal density.
- However, the integral does not yield a density we know. Additionally, to derive the marginal posterior density $p(\sigma^2|y_1,\ldots,y_n)$, we need to compute the proportionality constant.

Approximate joint posterior

- A convenient way to figure out the proportionality constant is to use the fact that:

$$p(\sigma^2|y_1,\ldots,y_n) = \frac{p(\mu,\sigma^2|y_1,\ldots,y_n)}{p(\mu|\sigma^2,y_1,\ldots,y_n)}$$

- However, rather than evaluating the integral, we can approximate the joint posterior distribution $p(\mu,\sigma^2|y_1,\ldots,y_n)$ using numerical methods.
  We choose a grid of values evenly spaced for μ,σ^2: $\{\mu_1,\mu_2,\ldots,\mu_G\}$ and $\{\sigma_1^2,\sigma_2^2,\ldots,\sigma_H^2\}$. Then, we approximate $p(\mu,\sigma^2|y_1,\ldots,y_n)$ at (μ_j,σ_k^2) for $j=1,\ldots,G$ and $k=1,\ldots,H$ using the following formula:

$$p_D(\mu_j,\sigma_k^2|y_1,\ldots,y_n) = \frac{p(\mu_j,\sigma_k^2|y_1,\ldots,y_n)}{\sum_{j=1}^{G}\sum_{k=1}^{H} p(\mu_j,\sigma_k^2|y_1,\ldots,y_n)} = \frac{\frac{p(y_1,\ldots,y_n|\mu_j,\sigma_k^2)p(\mu_j)p(\sigma_k^2)}{p(y_1,\ldots,y_n)}}{\sum_{j=1}^{G}\sum_{k=1}^{H}\frac{p(y_1,\ldots,y_n|\mu_j,\sigma_k^2)p(\mu_j)p(\sigma_k^2)}{p(y_1,\ldots,y_n)}}$$

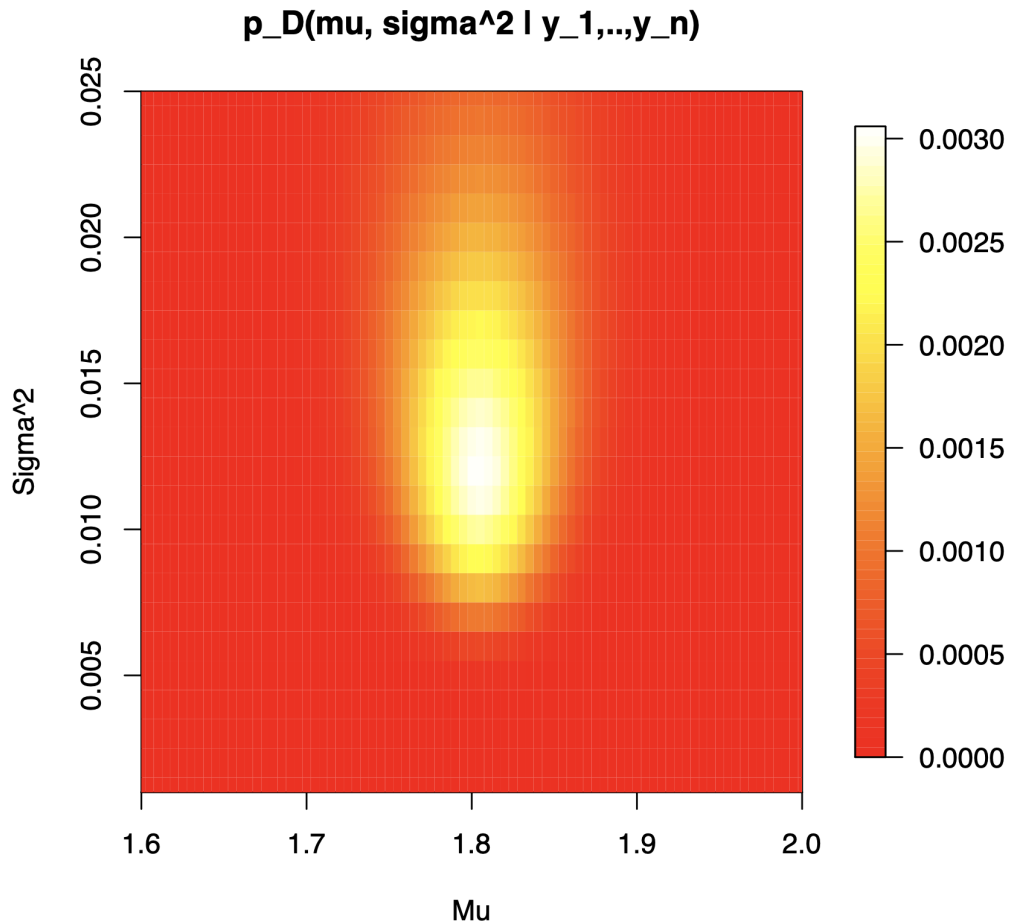
$$= \frac{p(y_1,\ldots,y_n|\mu_j,\sigma_k^2)p(\mu_j)p(\sigma_k^2)}{\sum_{j=1}^{G}\sum_{k=1}^{H} p(y_1,\ldots,y_n|\mu_j,\sigma_k^2)p(\mu_j)p(\sigma_k^2)}$$

Approximate joint posterior: example

- Note that $p_D(\mu, \sigma^2 | y_1, \ldots, y_n)$ is an approximation to the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ and is also a real joint probability distribution since it sums up to 1.
  Infact, it **IS** the true joint posterior distribution of μ and σ^2 if μ and σ^2 had discrete priors and could only take a finite number of values.
- Example: Let's consider the data on the length of 9 members of a species of midge, $y_1 = 1.64$, $y_2 = 1.70$, $y_3 = 1.72$, $y_4 = 1.74$, $y_5 = 1.82$, $y_6 = 1.82$, $y_7 = 1.82$, $y_8 = 1.90$ and $y_9 = 2.08$ with $\bar{y} = 1.804$ and $s^2 = 0.017$.
- Let the priors on μ and σ^2 be: $p(\mu) = N(\mu_0, \tau_0^2) = N(1.9, 0.95^2)$ and $p(\sigma^2) = \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) = \text{InverseGamma}(\frac{1}{2}, \frac{0.01}{2})$.
- Let's consider a grid of 100 values for μ and σ^2:
  - $\{\mu_1, \mu_2, \ldots, \mu_G\} = \{1.505, 1.510, \ldots, 2.00\}$
  - $\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_H^2\} = \{0.001, 0.002, \ldots, 0.10\}$

Approximate joint posterior: example

Plot of the approximate joint posterior distribution $p_D(\mu, \sigma^2 | y_1, \ldots, y_n)$



Approximate marginal posterior

- From the approximate discrete joint posterior distribution $p_D(\mu, \sigma^2|y_1, \ldots, y_n)$, we can derive the approximate discrete marginal posterior distributions $p_D(\mu|y_1, \ldots, y_n)$ and $p_D(\sigma^2|y_1, \ldots, y_n)$.
- Using the grids of values $\{\mu_1, \ldots, \mu_G\}$ and $\{\sigma_1^2, \ldots, \sigma_H^2\}$, we define $p_D(\mu_j|y_1, \ldots, y_n)$ as:

$$p_D(\mu_j|y_1, \ldots, y_n) = \sum_{k=1}^{H} p_D(\mu_j, \sigma_k^2|y_1, \ldots, y_n) \qquad j = 1, \ldots, G$$

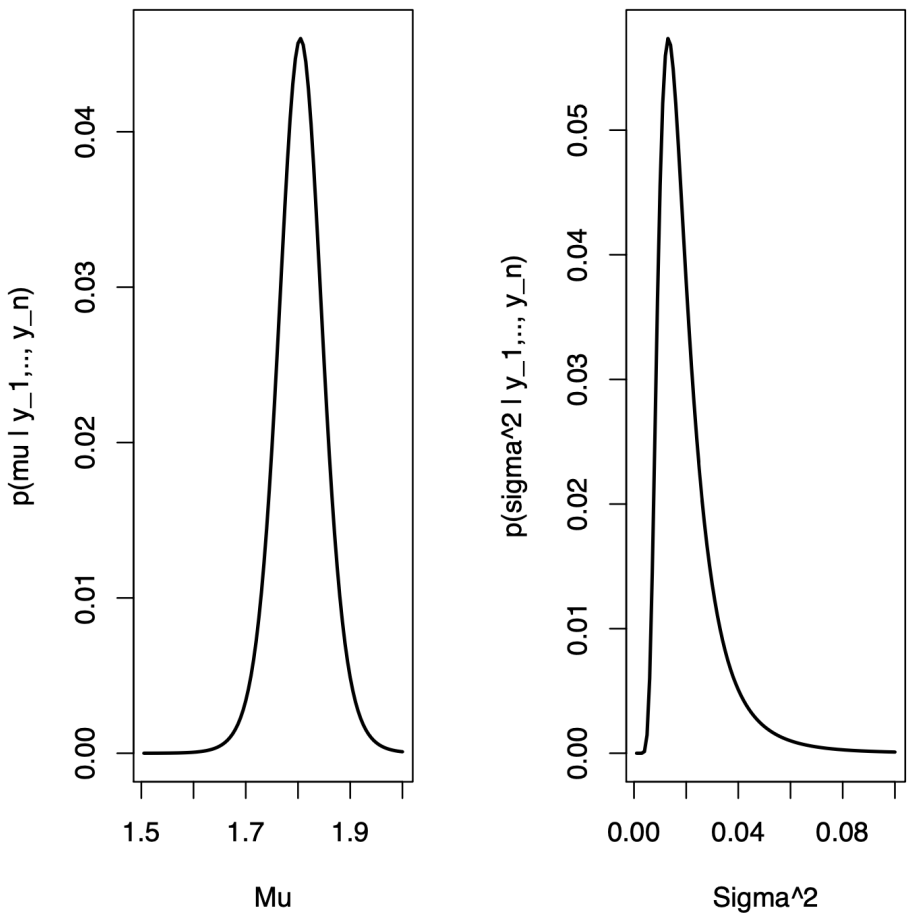
  and similarly for $p_D(\sigma_k^2|y_1, \ldots, y_n)$, $k = 1, \ldots, H$.
- Note that here there is no need to normalize here since

$$\sum_{j=1}^{G} p_D(\mu_j|y_1, \ldots, y_n) = \sum_{j=1}^{G} \left[\sum_{k=1}^{H} p_D(\mu_j, \sigma_k^2|y_1, \ldots, y_n) \right] = 1$$

  by the definition of $p_D(\mu_j, \sigma_k^2|y_1, \ldots, y_n)$, $j = 1, \ldots, J$ and $k = 1, \ldots, H$. Similarly for $p_D(\sigma^2|y_1, \ldots, y_n)$.

Approximate marginal posterior

Plot of the approximate marginal posterior distributions $p_D(\mu|y_1,\ldots,y_n)$ and $p_D(\sigma^2|y_1,\ldots,y_n)$



Non-closed form of marginal posterior

- In the case of normal data with semi-conjugate prior, the marginal posterior distribution for σ^2 does not have a closed form that we know.
- How do we perform Bayesian inference in this case?
- Let's consider the case of one-parameter models (this will very easily translate to the case of multi-parameter models): suppose we have $y_1, \ldots, y_n | \theta \overset{iid}{\sim} p(y|\theta)$ and $p(\theta)$ prior distribution on θ.
- Suppose we have derived the posterior distribution $p(\theta|y_1, \ldots, y_n)$ but it does not have a closed form that we recognize and from which we can sample.
- What are approaches that we can take to perform inference in this case?
    1. Discrete grid approximation
    2. Rejection sampling
    3. Importance sampling

Rejection sampling

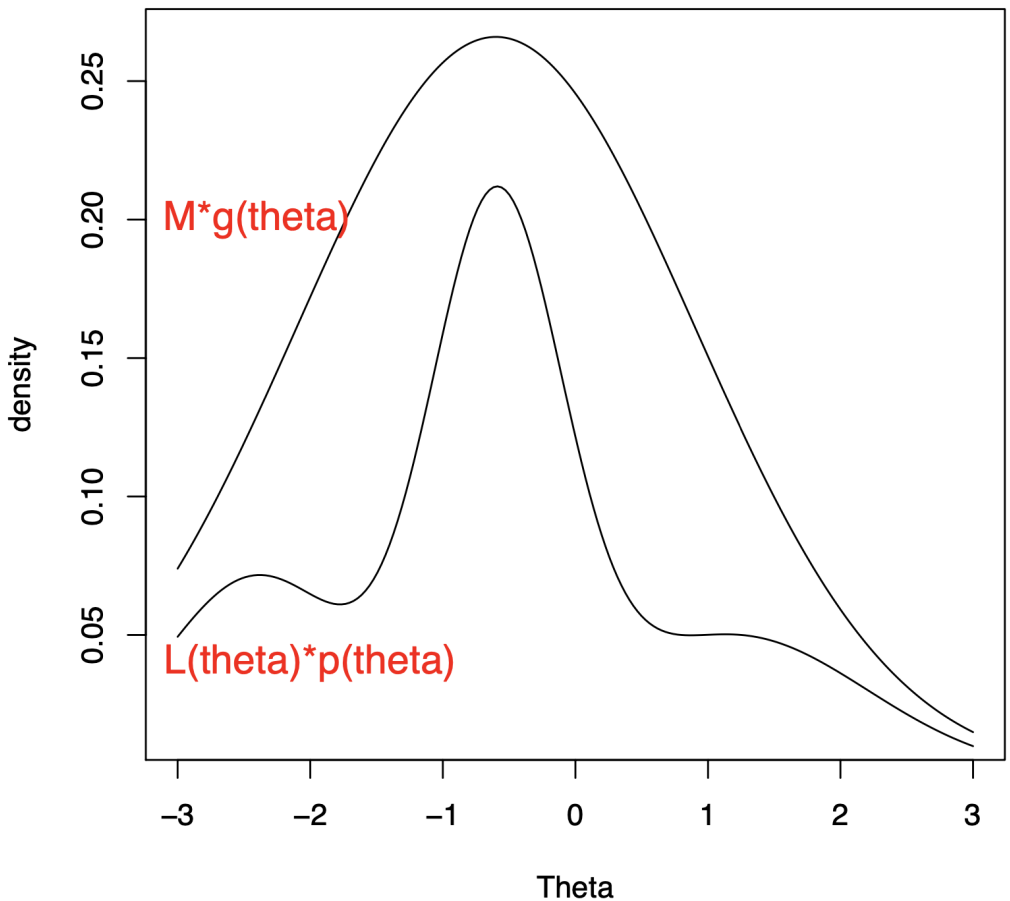
- A method that can be used to sample from the posterior distribution $p(\theta|y_1, \ldots, y_n)$ but that does not require to know the constant of proportionality is the method of rejection sampling.
- Suppose that there exists an envelope function $g(\theta)$ from which it is easy to sample and suppose there exists a constant $M > 0$ such that

$$p(y_1, \ldots, y_n|\theta) \cdot p(\theta) = \left[\prod_{i=1}^{n} p(y_i|\theta)\right] \cdot p(\theta) = L(\theta) \cdot p(\theta) < Mg(\theta)$$

for all θ where $L(\theta)$ is the likelihood function.

Rejection sampling

Plot of the envelope function $M \cdot g(\theta)$ and of the un-normalized posterior distribution $L(\theta) \cdot p(\theta)$.



Rejection sampling

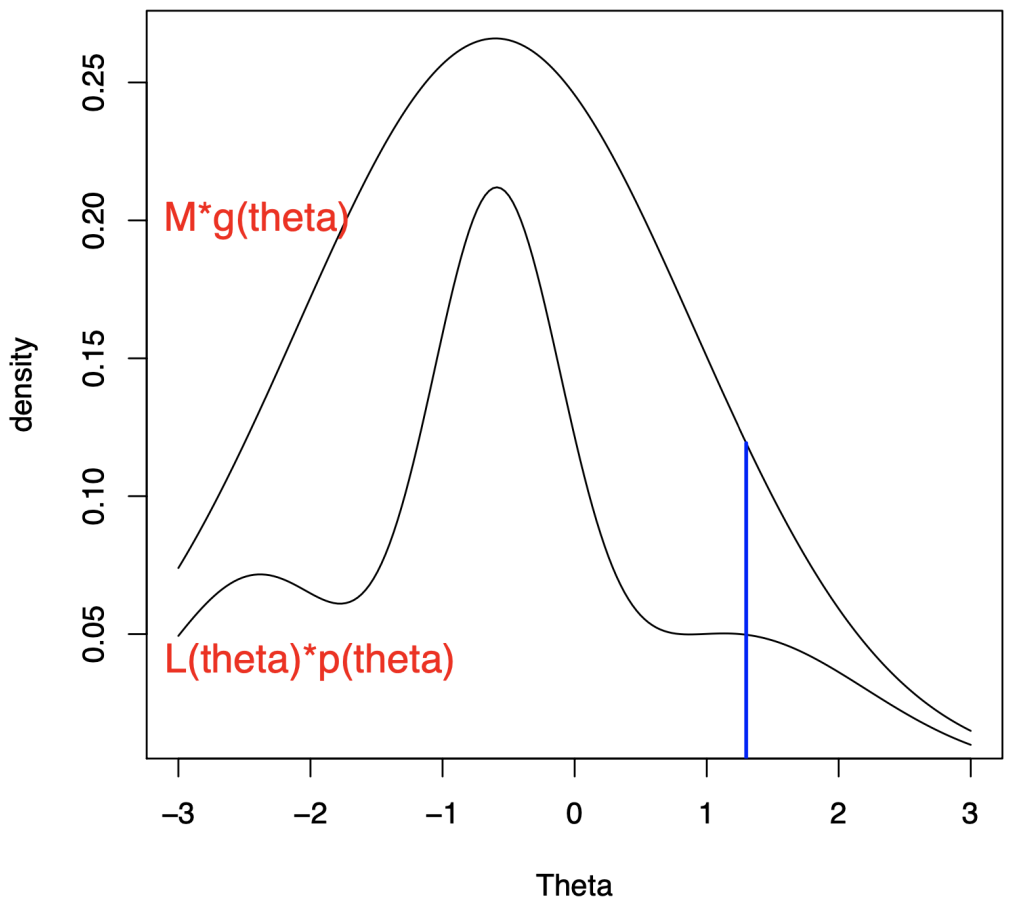
- Then, the rejection sampling algorithm works as follows:
  - Choose a number B of simulations
  - Sample a value $\theta^{(i)} \sim g(\theta)$
  - Sample a value $u^{(i)} \sim \text{Uniform}[0,1]$
  - If $M \cdot u^{(i)} \cdot g(\theta^{(i)}) < L(\theta^{(i)}) \cdot p(\theta^{(i)})$, accept $\theta^{(i)}$, otherwise reject it.
- Repeat the last 3 steps until a large enough sample $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$ is generated. This is a sample from the normalized posterior distribution $p(\theta|y_1, \ldots, y_n)$:

$$p(\theta|y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n|\theta) \cdot p(\theta)}{p(y_1, \ldots, y_n)} = \frac{L(\theta) \cdot p(\theta)}{\int_{\theta \in \Theta} L(\theta) \cdot p(\theta) d\theta}$$

- Then, we can use the sample to compute posterior summaries (posterior mean, posterior variance, credible intervals, etc. etc.).

Rejection sampling: why it works

Plot of the envelope function $M \cdot g(\theta)$ and of the un-normalized posterior distribution $L(\theta) \cdot p(\theta)$.



Rejection sampling: why it works

- For a formal proof of why rejection sampling works, look at Devroye (1986; p. 40-42).
- Here is an informal justification in the case of a univariate θ. Consider a fairly large sample of values generated from the density $g(\theta)$. Then it's possible to rescale the histogram, so that this is the curve $Mg(\theta)$.
- Now, consider a value $\theta^{(i)}$. The rejection step has the effect of slicing off the top portion of the line $x = \theta^{(i)}$ since only those points for which $M \cdot u^{(i)} \cdot g(\theta^{(i)})$ is below the lower curve are retained.
- Since this is true for any value $\theta^{(i)}$, the histogram of the accepted $\theta^{(i)}$ will mimic the shape of the lower curve $L(\theta) \cdot p(\theta)$ which is proportional to the posterior distribution $p(\theta|y_1, \ldots, y_n)$.

Rejection sampling: choosing M and $g(\theta)$

- How to choose M and $g(\theta)$ in rejection sampling?
- The constant M determines how many candidate values are rejected. Therefore, by intuition it is best to choose M as small as possible.
- Formally: if Z is a random variable that counts the number of iterations necessary to get an accepted value of $\theta^{(i)}$, then Z is a geometric random variable with parameter
  $p =$ probability of acceptance where

$$p = \frac{c}{M}$$

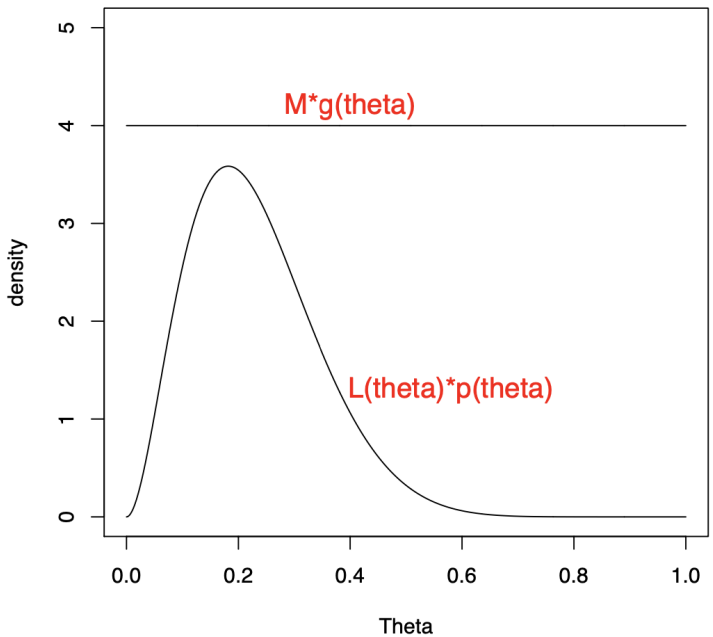
  with c proportionality constant in the posterior distribution $p(\theta|y_1, \ldots, y_n)$.
- Since $E(Z) = \frac{1}{p} = \frac{M}{c}$, if we want to minimize the number of wasted $\theta^{(i)}$ samples, we need to take M quite small. The best choice would be $M = c$.

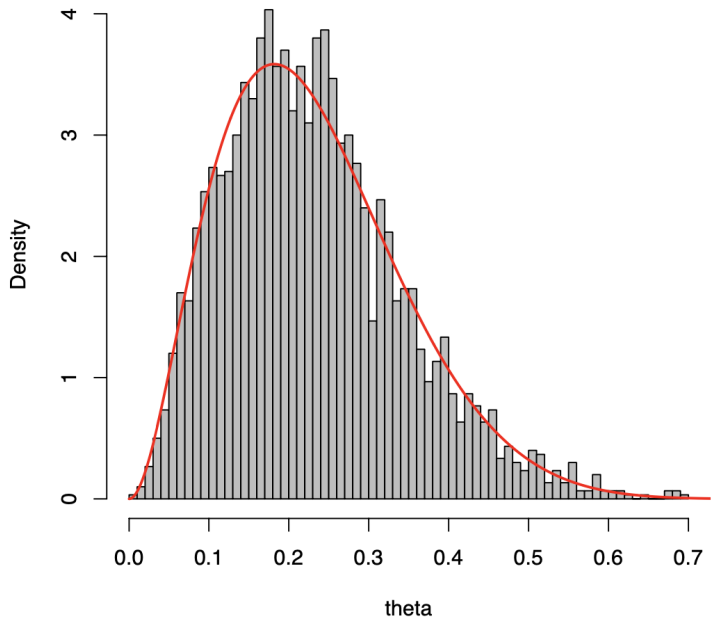
Rejection sampling: choosing M and $g(\theta)$

- For the envelope density $g(\theta)$, the most important characteristics this density should have, are:
  - it needs to satisfy the condition $L(\theta)p(\theta) < Mg(\theta)$. If this condition is violated, the sample of accepted $\theta^{(i)}$ is not a sample from the posterior distribution $p(\theta|y_1, \ldots, y_n)$
  - it should have heavier tails than $L(\theta) \cdot p(\theta)$
  - it should have many sharp peaks to ensure that there are many rejection candidates regions across the domain.

Rejection sampling: example

Suppose we want to sample from $L(\theta) \cdot p(\theta) = \text{Beta}(\theta; 3, 10)$. We use rejection sampling until we obtain a sample of size $N = 3,000$.





Importance sampling

- The last method we are going to consider is importance sampling.
- Suppose that we have derived the posterior distribution $p(\theta|y_1,\ldots,y_n)$:

$$p(\theta|y_1,\ldots,y_n) = \frac{p(y_1,\ldots,y_n|\theta)\cdot p(\theta)}{p(y_1,\ldots,y_n)} = \frac{p(y_1,\ldots,y_n|\theta)\cdot p(\theta)}{\int_{\theta\in\Theta} p(y_1,\ldots,y_n|\theta)\cdot p(\theta)d\theta}$$

$$= \frac{L(\theta)\cdot p(\theta)}{\int_{\theta\in\Theta} L(\theta)\cdot p(\theta)d\theta}$$

and we want to summarize it through the posterior mean $E(\theta|y_1,\ldots,y_n)$:

$$E(\theta|y_1,\ldots,y_n) = \frac{\int_{\theta\in\Theta} \theta\cdot L(\theta)\cdot p(\theta)d\theta}{\int_{\theta\in\Theta} L(\theta)\cdot p(\theta)d\theta}$$

- If the posterior distribution $p(\theta|y_1,\ldots,y_n)$ does not have a closed form, we can compute $E(\theta|y_1,\ldots,y_n)$, and in general any expectation $E(h(\theta)|y_1,\ldots,y_n)$ where $h(\theta)$ is a function of θ, via importance sampling.

Importance sampling

- Suppose there exists a function $g(\theta)$, called the importance function, such that:
  1. $g(\theta) > 0$ whenever $L(\theta) \cdot p(\theta) \neq 0$
  2. $g(\theta)$ approximates or is very close to the normalized likelihood times the prior, $c \cdot L(\theta) \cdot p(\theta)$, but with heavier tails
  3. $g(\theta)$ is easy to sample from
- We define the weight function $w(\theta) = \frac{L(\theta) \cdot p(\theta)}{g(\theta)}$.
  For any function $h(\theta)$

$$
\begin{aligned}
E(h(\theta)|y_1, \ldots, y_n) &= \int_{\theta \in \Theta} h(\theta) \cdot p(\theta|y_1, \ldots, y_n) d\theta \\
&= \frac{\int_{\theta \in \Theta} h(\theta) \cdot L(\theta) \cdot p(\theta) d\theta}{\int_{\theta \in \Theta} L(\theta) \cdot p(\theta) d\theta} = \frac{\int_{\theta \in \Theta} h(\theta) \cdot \frac{L(\theta) \cdot p(\theta)}{g(\theta)} \cdot g(\theta) d\theta}{\int_{\theta \in \Theta} \frac{L(\theta) \cdot p(\theta)}{g(\theta)} g(\theta) d\theta} \\
&= \frac{\int_{\theta \in \Theta} h(\theta) \cdot w(\theta) \cdot g(\theta) d\theta}{\int_{\theta \in \Theta} w(\theta) \cdot g(\theta) d\theta}
\end{aligned}
$$

Importance sampling

- Thus:

$$E(h(\theta)|y_1,\ldots,y_n) = \frac{\int_{\theta\in\Theta} h(\theta)\cdot w(\theta)\cdot g(\theta)d\theta}{\int_{\theta\in\Theta} w(\theta)\cdot g(\theta)d\theta}$$

  where $w(\theta) = \frac{L(\theta)\cdot p(\theta)}{g(\theta)}$.

- We approximate $E(h(\theta)|y_1,\ldots,y_n)$ via Monte Carlo methods, by sampling B values $\theta^{(i)} \sim g(\theta)$ and setting:

$$E(h(\theta)|y_1,\ldots,y_n) \approx \frac{\frac{1}{N}\cdot\sum_{i=1}^{B} h(\theta^{(i)})\cdot w(\theta^{(i)})}{\frac{1}{N}\cdot\sum_{i=1}^{B} w(\theta^{(i)})}$$

Importance sampling

- $E(h(\theta)|y_1,\ldots,y_n) \approx \frac{\frac{1}{N}\cdot\sum_{i=1}^{B} h(\theta^{(i)})\cdot w(\theta^{(i)})}{\frac{1}{N}\cdot\sum_{i=1}^{B} w(\theta^{(i)})}$
- The quality of the approximation obtained via importance sampling depends on the importance function $g(\theta)$:
  - If $g(\theta)$ is a good approximation to the normalized likelihood times prior $c\cdot L(\theta)\cdot p(\theta)$, then the weights are going to be all approximately equal and that minimize the variance of the numerator and the denominator.
  - If $g(\theta)$ is a poor approximation to the normalized likelihood times prior $c\cdot L(\theta)\cdot p(\theta)$, then many weights are approximately equal to 0 and a few $\theta^{(i)}$ would dominate the sums, producing an innacurate approximation.
- Finding the importance function is the challenging part of importance sampling!

Normal data with a semi-conjugate prior

- On pp. 5–15 , we have looked at Bayesian inference for data $y_1, \ldots, y_n$ such that $y_1, \ldots, y_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ when:

$$p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2)$$

  where $p(\mu) = N(\mu_0, \tau_0^2)$ and $p(\sigma^2) = \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$.
- We have seen that in this case, the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ is not a distribution we know.
- We have looked at two ways to approximate the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$:
  1. numerical methods: approximating $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ with $p_D(\mu_j, \sigma_k^2 | y_1, \ldots, y_n)$, $j = 1, \ldots, G$ and $k = 1, \ldots, H$, where μ_j and σ_k^2 are in a dense grid of values $\{\mu_1, \ldots, \mu_G\}$ and $\{\sigma_1^2, \ldots, \sigma_H^2\}$.
  2. Monte Carlo methods: approximating $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ with the empirical distribution (or histograms) of samples $(\mu^{(i)}, \sigma^{2(i)})$, $i = 1, \ldots, B$.

Normal data with a semi-conjugate prior

- To approximate the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ via Monte Carlo methods we would proceed as follows:
  - choose a number B of samples;
  - for i from 1 to B:
    - sample $\sigma^{2(i)}$ from $p(\sigma^2 | y_1, \ldots, y_n)$
    - sample $\mu^{(i)}$ from $p(\mu | \sigma^{2(i)}, y_1, \ldots, y_n)$
- We have seen on page 10 that the marginal posterior distribution $p(\sigma^2 | y_1, \ldots, y_n)$ is not a distribution that we know.
- We will now see a different way to approximate the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$.

Conditional posterior distribution of σ^2

- Remember the sampling model and the prior:
  $y_1, \ldots, y_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ and $p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2)$ where
  $p(\mu) = N(\mu_0, \tau_0^2)$ and $p(\sigma^2) = \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$.
- The problem, as we saw, is with the marginal posterior distribution $p(\sigma^2 | y_1, \ldots, y_n)$.
- Now, suppose that we knew the value of μ. Then, the conditional posterior distribution $p(\sigma^2 | \mu, y_1, \ldots, y_n)$ is given by:

$$
\begin{aligned}
p(\sigma^2 | \mu, y_1, \ldots, y_n) &= \frac{p(y_1, \ldots, y_n, \mu, \sigma^2)}{p(y_1, \ldots, y_n, \mu)} = \frac{p(y_1, \ldots, y_n | \mu, \sigma^2) \cdot p(\mu, \sigma^2)}{\int p(y_1, \ldots, y_n, \mu, \sigma^2) d\sigma^2} \\
&= \frac{p(y_1, \ldots, y_n | \mu, \sigma^2) \cdot p(\sigma^2)}{\int p(y_1, \ldots, y_n | \mu, \sigma^2) \cdot p(\sigma^2) d\sigma^2} \propto p(y_1, \ldots, y_n | \mu, \sigma^2) \cdot p(\sigma^2) \\
&= \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \left[(n-1)s^2 + n(\bar{y} - \mu)^2 \right] \right) \\
&\cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \cdot \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2} \right)
\end{aligned}
$$

Conditional posterior distribution of σ^2

$$
\begin{aligned}
p(\sigma^2|\mu, y_1, \ldots, y_n) &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right) \\
&\cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \cdot \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \\
&\propto \frac{1}{(\sigma^2)^{\frac{n+\nu_0}{2}+1}} \cdot \exp\left(-\frac{1}{\sigma^2}\frac{\left[(n-1)s^2 + n(\bar{y}-\mu)^2 + \nu_0\sigma_0^2\right]}{2}\right)
\end{aligned}
$$

which is the kernel of an Inverse Gamma$(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2(\mu)}{2})$ distribution.

- Since:

$$
\begin{aligned}
p(\sigma^2|y_1, \ldots, y_n, \mu) &= \text{Inverse Gamma}(\tfrac{n+\nu_0}{2}, \tfrac{(n-1)s^2 + n(\bar{y}-\mu)^2 + \nu_0\sigma_0^2}{2}) \\
&= \text{Inverse Gamma}(\tfrac{n+\nu_0}{2}, \tfrac{\sum_{i=1}^n (y_i-\mu)^2 + \nu_0\sigma_0^2}{2}).
\end{aligned}
$$

it follows that $\nu_n = n + \nu_0$ and

$$
\sigma_n^2(\mu) = \frac{1}{\nu_n} \cdot \left[(n-1)s^2 + n(\bar{y}-\mu)^2 + \nu_0\sigma_0^2\right] = \frac{1}{\nu_n} \cdot \left[\sum_{i=1}^n (y_i-\mu)^2 + \nu_0\sigma_0^2\right].
$$

Full conditional distributions

- Therefore, if:

$$\begin{cases} y_1, \ldots, y_n \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2) \\ \mu, \sigma^2 \sim p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2) \\ \quad = N(\mu; \mu_0, \tau_0^2) \cdot \text{Inverse Gamma}(\sigma^2; \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) \end{cases}$$

  the two conditional posterior distributions are:
  - $p(\mu | y_1, \ldots, y_n, \sigma^2) \sim N(\mu; \mu_n, \tau_n^2)$
  - $p(\sigma^2 | y_1, \ldots, y_n, \mu) \sim \text{Inverse Gamma}(\sigma^2; \frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$
- The two conditional distributions above are called full conditional distributions of μ and σ^2, respectively, because they are conditional distributions **given** everything else.
- Since both distributions are known, if we knew μ, then we could easily sample from the conditional posterior distribution $p(\sigma^2 | y_1, \ldots, y_n, \mu)$. And similarly, if σ^2 was known, then we could easily sample from the conditional posterior distribution $p(\mu | y_1, \ldots, y_n, \sigma^2)$.

Sampling from conditional distributions

- Suppose that we were given a sampled value $\sigma^{2(1)}$ from the posterior marginal distribution $p(\sigma^2|y_1,\ldots,y_n)$.
- Then, we could sample $\mu^{(1)}$ from the conditional posterior distribution

$$\mu^{(1)} \sim p(\mu|y_1,\ldots,y_n,\sigma^{2(1)})$$

- $(\mu^{(1)},\sigma^{2(1)})$ is a sample from the joint posterior distribution $p(\mu,\sigma^2|y_1,\ldots,y_n)$
- If we only look at $\mu^{(1)}$, this can be considered as a sample from the marginal posterior distribution $p(\mu|y_1,\ldots,y_n)$.
- Since we know the conditional posterior distribution $p(\sigma^2|y_1,\ldots,y_n,\mu)$, we could use $\mu^{(1)}$ and sample $\sigma^{2(2)}$ from the conditional posterior distribution $p(\sigma^2|y_1,\ldots,y_n,\mu)$:

$$\sigma^{2(2)} \sim p(\sigma^2|y_1,\ldots,y_n,\mu^{(1)})$$

- $(\mu^{(1)},\sigma^{2(2)})$ is a sample from the joint posterior distribution $p(\mu,\sigma^2|y_1,\ldots,y_n)$

Sampling from conditional distributions

- Since $(\mu^{(1)}, \sigma^{2(2)})$ is a sample from the joint posterior distribution $p(\mu, \sigma^2|y_1, \ldots, y_n)$, $\sigma^{2(2)}$ is a sampled value from the marginal posterior distribution $p(\sigma^2|y_1, \ldots, y_n)$
- We can then use $\sigma^{2(2)}$ and sample a new value $\mu^{(2)}$ from the conditional posterior distribution $p(\mu|y_1, \ldots, y_n, \sigma^2)$.
- Alternating the sampling from the two full conditional distributions, we obtain a sample $\left\{(\mu^{(1)}, \sigma^{2(1)}), (\mu^{(2)}, \sigma^{2(2)}), \ldots\right\}$
- Thus, it seems that if we can have a value $\sigma^{2(1)}$, then we can generate samples from the posterior joint distribution.
- This is the basic idea behind the Gibbs sampling algorithm.

Gibbs sampling algorithm

- More specifically, for the problem we considered:

$$\begin{cases} y_1, \ldots, y_n \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2) \\ \mu, \sigma^2 \sim p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2) \\ \quad = N(\mu; \mu_0, \tau_0^2) \cdot \text{Inverse Gamma}(\sigma^2; \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) \end{cases}$$

  let $\phi = (\mu, \sigma^2)$ be a two-dimensional parameter vector.

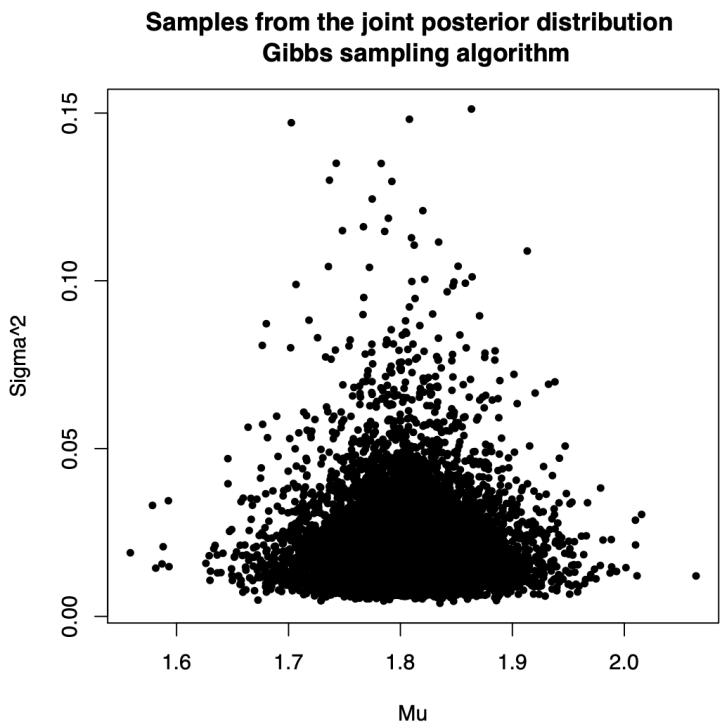
- In a Gibbs sampling algorithm, **given** a current state of the parameter vector $\phi^{(s)} = (\mu^{(s)}, \sigma^{2(s)})$, we generate a new parameter vector $\phi^{(s+1)}$ as follows:
  1. sample $\mu^{(s+1)}$ from $p(\mu | y_1, \ldots, y_n, \sigma^{2(s)})$
  2. sample $\sigma^{2(s+1)}$ from $p(\sigma^2 | y_1, \ldots, y_n, \mu^{(s+1)})$
  3. update $\phi^{(s)} = (\mu^{(s)}, \sigma^{2(s)})$ to $\phi^{(s+1)} = (\mu^{(s+1)}, \sigma^{2(s+1)})$
- The sampled values $\{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(S)}\}$ are a sample from the joint posterior distribution $p(\phi | y_1, \ldots, y_n) = p(\mu, \sigma^2 | y_1, \ldots, y_n)$. We can derive posterior summaries (e.g. mean, median, intervals) as we have done previously for Monte Carlo samples.

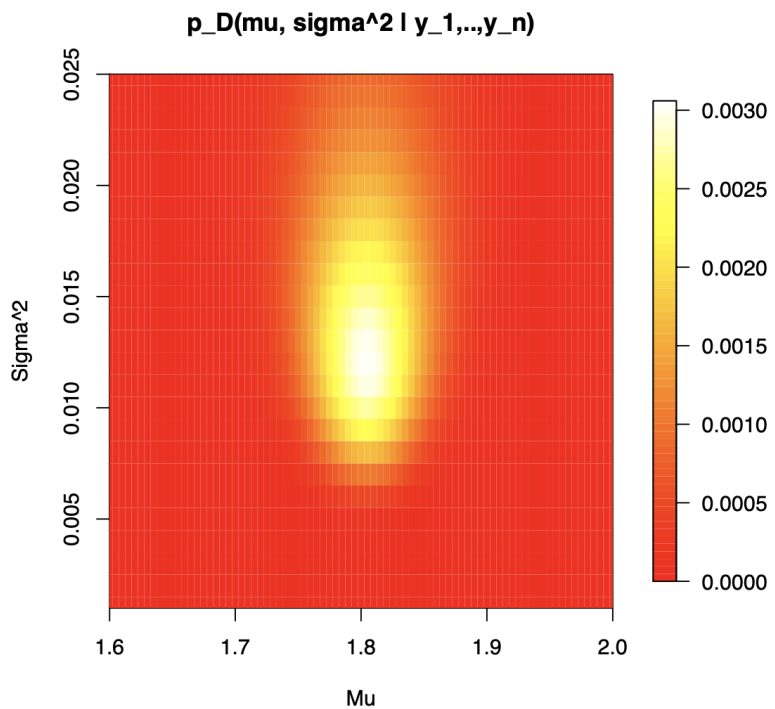
Gibbs sampling algorithm: example

- Let's consider again the example of the midge wing length: $y_1 = 1.64$, $y_2 = 1.70$, $y_3 = 1.72$, $y_4 = 1.74$, $y_5 = 1.82$, $y_6 = 1.82$, $y_7 = 1.82$, $y_8 = 1.90$ and $y_9 = 2.08$.
- Let the priors on μ and σ^2 be, respectively: $p(\mu) = N(\mu_0, \tau_0^2) = N(1.9, 0.95^2)$ and $p(\sigma^2) = \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) = \text{InverseGamma}(\frac{1}{2}, \frac{0.01}{2})$.
- We run the Gibbs sampling algorithm, sampling $S = 10,000$ times from the full conditional distributions with initial value for σ^2 and μ, respectively, the sample mean and the sample variance.
- We can use the samples $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$ to infer upon the posterior distribution: we can look at the posterior marginal distribution of μ, σ^2 or any other function of these two parameters.
- We can also compute posterior summaries for these two posterior marginal distributions: posterior means, posterior medians, 95% credible intervals.

Gibbs sampling algorithm: example

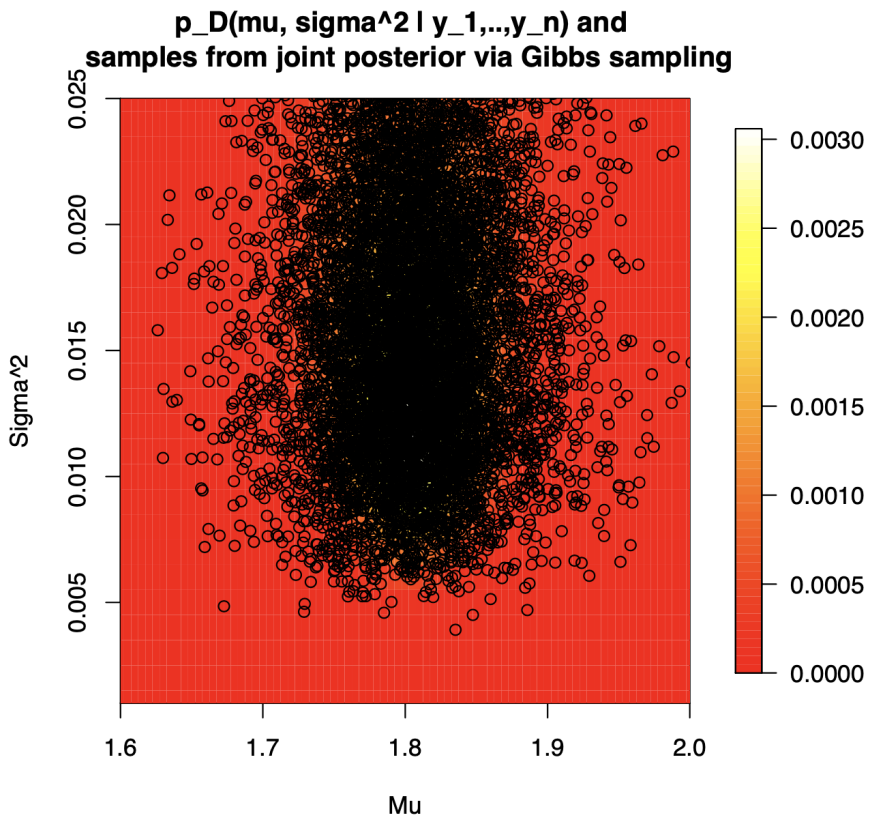
- Plot of the samples from the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ obtained via Gibbs sampling and plot of the discrete approximation $p_D(\mu_j, \sigma_k^2 | y_1, \ldots, y_n)$ obtained using numerical methods (see pp. 11-13).





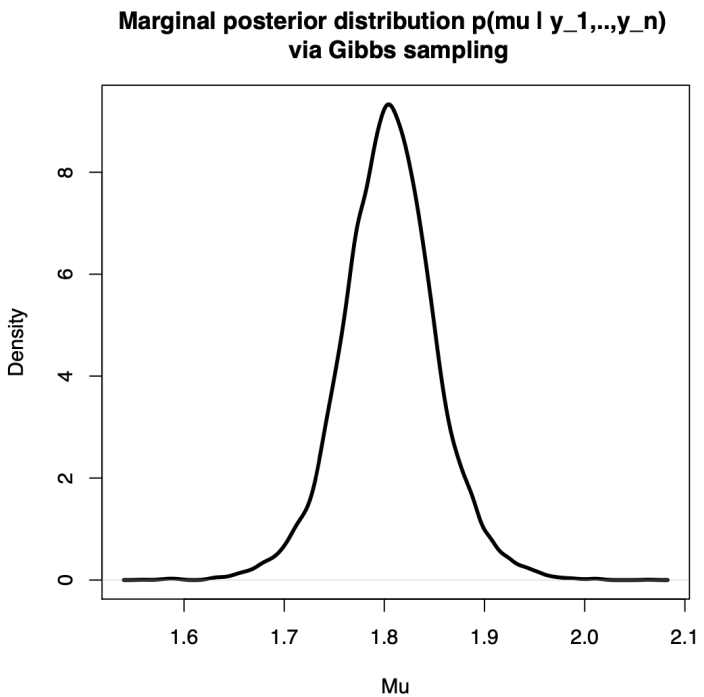
Gibbs sampling algorithm: example

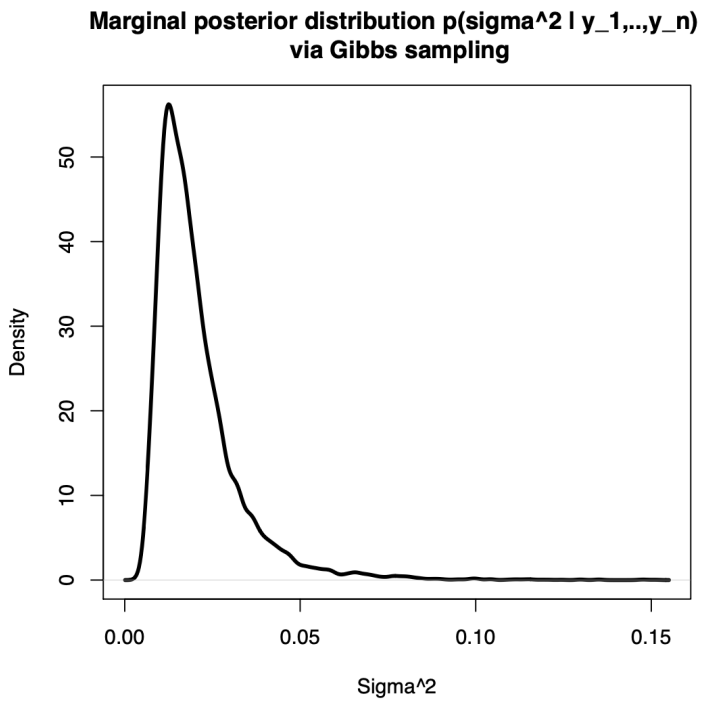
- Plot of the discrete approximation $p_D(\mu_j, \sigma_k^2 | y_1, \ldots, y_n)$ to the joint posterior distribution $p(\mu, \sigma^2 | y_1, \ldots, y_n)$ obtained using numerical methods (see pp. 11–13) with overlaid the sampled values from the joint posteriorior distribution via Gibbs sampling.



Gibbs sampling algorithm: example

- Approximation to the posterior marginal distributions $p(\mu|y_1, \ldots, y_n)$ and $p(\sigma^2|y_1, \ldots, y_n)$ using the samples from the joint posterior distribution obtained via Gibbs sampling.





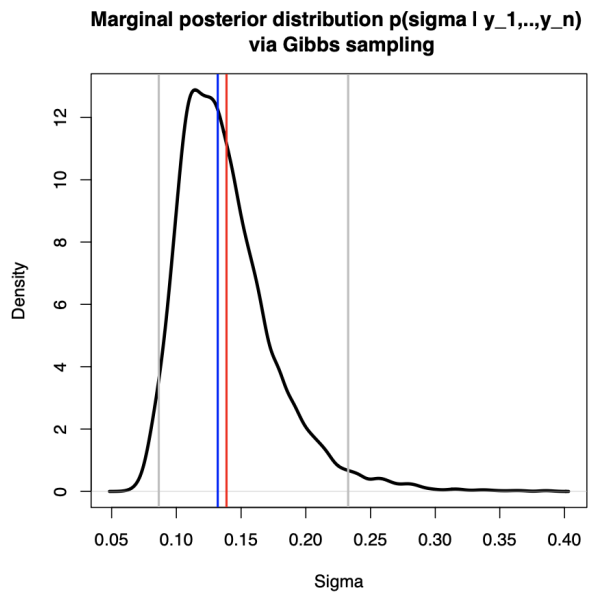
Gibbs sampling algorithm: example

- We can summarize these posterior distributions reporting posterior summaries.
  For example:

Parameter	Posterior Mean	Posterior Median	95% Credible Interval
μ	1.80	1.80	(1.7; 1.9)
σ^2	0.021	0.017	(0.007; 0.05)

Gibbs sampling algorithm: example

- We can use the samples from the joint posterior distribution to obtain approximation to the posterior distribution of any function of the two parameters μ and σ^2.
- Approximation to the marginal posterior distribution $p(\sigma|y_1, \ldots, y_n)$ using samples from the joint posterior distribution obtained via Gibbs sampling.



- The posterior mean, posterior median and 95% credible interval for σ are, respectively: 0.14, 0.13 and $[0.086; 0.233]$

Gibbs sampling algorithm

- In general, the Gibbs sampling algorithm works as follows: suppose that we have observations $y_1, \ldots, y_n | \phi \overset{iid}{\sim} p(y|\phi)$ where $\phi = (\phi_1, \ldots, \phi_p)$ and we place a prior $p(\phi)$ on ϕ.
- We are interested in the joint posterior distribution $p(\phi|y_1, \ldots, y_n)$.
- **Given** an initial value, a starting point, $\phi^{(0)} = \left(\phi_1^{(0)}, \phi_2^{(0)}, \ldots, \phi_p^{(0)}\right)$, the Gibbs sampling algorithm provides an approximation to the posterior distribution $p(\phi|y_1, \ldots, y_n)$ by sampling iteratively from the full conditional distributions.
  Precisely, at the s-th iteration, the algorithm proceeds as follows:
  - sample $\phi_1^{(s)}$ from $p(\phi_1|y_1, \ldots, y_n, \phi_2^{(s-1)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)})$
  - sample $\phi_2^{(s)}$ from $p(\phi_2|y_1, \ldots, y_n, \phi_1^{(s)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)})$
  - sample $\phi_3^{(s)}$ from $p(\phi_3|y_1, \ldots, y_n, \phi_1^{(s)}, \phi_2^{(s)}, \phi_4^{(s-1)}, \ldots, \phi_p^{(s-1)})$
  - $\ldots$
  - sample $\phi_{p-1}^{(s)}$ from $p(\phi_{p-1}|y_1, \ldots, y_n, \phi_1^{(s)}, \phi_2^{(s)}, \phi_3^{(s)}, \ldots, \phi_{p-2}^{(s)}, \phi_p^{(s-1)})$
  - sample $\phi_p^{(s)}$ from $p(\phi_p|y_1, \ldots, y_n, \phi_1^{(s)}, \phi_2^{(s)}, \phi_3^{(s)}, \ldots, \phi_{p-1}^{(s)})$

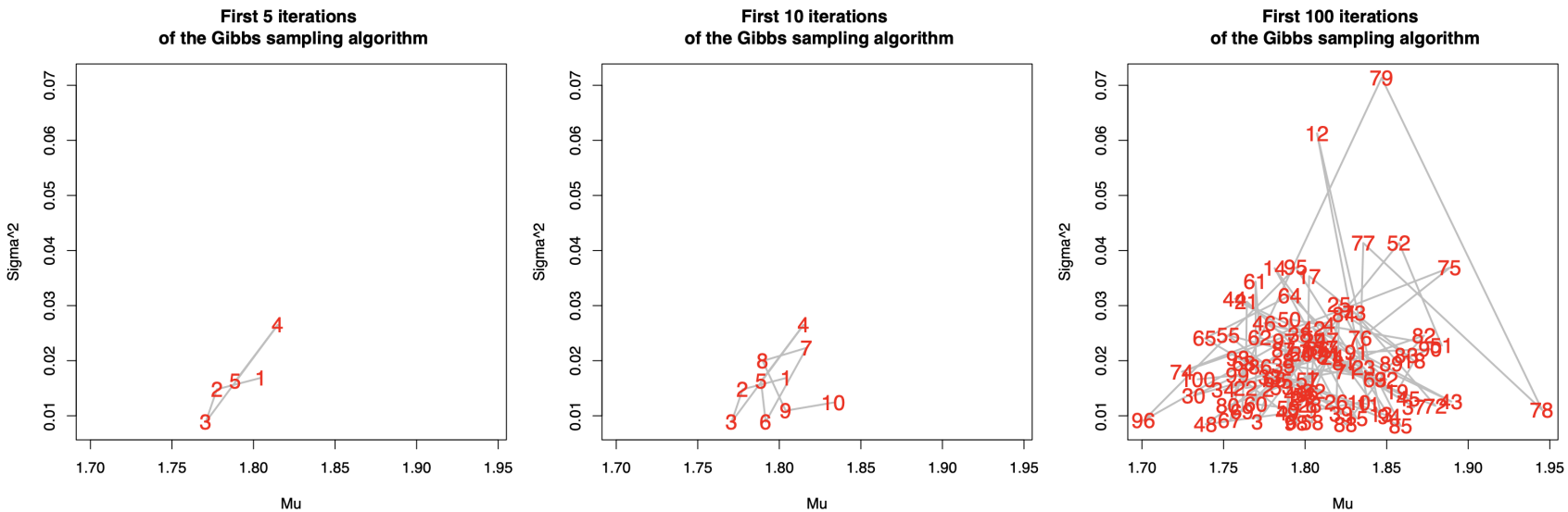
Gibbs sampling algorithm

- Thus, the algorithm generates a dependent sequence of vectors:

$$
\begin{aligned}
\phi^{(1)} &= \left(\phi_1^{(1)}, \ldots, \phi_p^{(1)}\right) \\
\phi^{(2)} &= \left(\phi_1^{(2)}, \ldots, \phi_p^{(2)}\right) \\
\phi^{(3)} &= \left(\phi_1^{(3)}, \ldots, \phi_p^{(3)}\right) \\
\vdots \quad &\vdots \quad \vdots \\
\phi^{(S)} &= \left(\phi_1^{(S)}, \ldots, \phi_p^{(S)}\right)
\end{aligned}
$$

Gibbs sampling algorithm: example

- Plot of the first 5, 10 and 100 iterations of the Gibbs sampling algorithm ran on the midge data



Gibbs sampling algorithm

- The Gibbs sampling algorithm generates a dependent sequence of vectors.
- At each iteration s, the s-th value $\phi^{(s)}$ depends on the "past", $\phi^{(0)}, \phi^{(1)}, \ldots, \phi^{(s-1)}$ only through $\phi^{(s-1)}$.
- A sequence of dependent random variables/random vectors that satisfy this property, is said to enjoy the Markov property and the sequence is called a Markov chain.
- We are now going to digress a little bit from Gibbs sampling and we are going to take a look at a series of definitions and theorems that are at the basis of MCMC (=Markov Chain Monte Carlo) methods.
- This will show us why the Gibbs sampling algorithm works and why we can approximate joint posterior distribution by sampling from the full conditionals

Markov chain

- A sequence of random variables $\{X_0, X_1, \ldots\}$ having the same sample space $\mathcal{X}$ and indexed by some set T is called a stochastic process.
- Example: X_t could be the exchange rate of the US dollar in the European market starting from January 1, 2000. So, X_0 refers to January 1, 2000, X_1 refers to the exchange rate in January 2, 2000 and so forth.
- A stochastic process $\{X_0, X_1, \ldots,\}$ is called a Markov chain if it satisfies the Markov property:

$$P(X_{n+1} \in A | X_0, X_1, \ldots, X_n) = P(X_{n+1} \in A | X_n)$$

- Typically, the Markov chains that we will consider in our applications, have as sample space (a subset of) $\mathbb{R}^d$. However, it is much easier to illustrate the main properties of Markov chains in the discrete case, that is, when the sample space $\mathcal{X}$ is either finite or countably infinite.
- In the latter case, $\mathcal{X}$ can be for example the set of non-negative integers: $\mathcal{X} = \mathbb{N} \cup \{0\}$

Transition probabilities

- Let's assume $\mathcal{X}$ is discrete and $\{X_0, X_1, \ldots\}$ is a Markov chain with sample space $\mathcal{X}$. Then the Markov property in this case is:

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = P(X_{n+1} = j | X_n = i) = p_{ij}$$

- The p_{ij} where $i, j \in \mathcal{X}$ are called the one-step transition probabilities.
- They satisfy the following conditions:

$$p_{ij} \geq 0 \qquad \sum_{j \in \mathcal{X}} p_{ij} = 1$$

- Similarly, we can define the t-step transition probability $p_{ij}(t)$:

$$p_{ij}(t) = P(X_t = j | X_0 = i)$$

- For example:

$$\begin{aligned} p_{ij}(2) &= P(X_2 = j | X_0 = i) = \textstyle\sum_{k \in \mathcal{X}} P(X_2 = j, X_1 = k | X_0 = i) \\ &= \textstyle\sum_{k \in \mathcal{X}} P(X_2 = j | X_1 = k, X_0 = i) \cdot P(X_1 = k | X_0 = i) \\ &= \textstyle\sum_{k \in \mathcal{X}} P(X_2 = j | X_1 = k) \cdot P(X_1 = k | X_0 = i) = \sum_{k \in \mathcal{X}} p_{kj} \cdot p_{ik}. \end{aligned}$$

Stationary distribution

- Let $\{X_0, X_1, \ldots\}$ be a Markov chain with sample space $\mathcal{X}$ and with transition probabilities p_{ij}.
- Let π be a probability distribution on $\mathcal{X}$, that is, for all $i \in \mathcal{X}$, $\pi(i) \geq 0$ and $\sum_{i \in \mathcal{X}} \pi(i) = 1$.
  Then, π is called a stationary distribution for the Markov chain $\{X_0, X_1, \ldots\}$ if

$$\sum_{i \in \mathcal{X}} \pi(i) \cdot p_{ij} = \pi(j)$$

- In other words, a probability distribution π is a stationary distribution if the expected value of transitioning into state j from any state i is $\pi(j)$ for any $j \in \mathcal{X}$.
- Our goal is to construct a Markov chain whose stationary distribution is the posterior distribution $p(\phi | y_1, \ldots, y_n)$.
- A Markov chain that satisfies certain properties admits a stationary distribution.

Irreducibility and recurrence

- A Markov chain $\{X_0, X_1, \ldots\}$ is said irreducible if for all i, j there exists a $t > 0$ such that $p_{ij}(t) > 0$.
- Therefore, in an irreducible Markov chain it is possible to reach any state starting from a state $i \in \mathcal{X}$.
- Suppose that the initial state of a Markov chain $\{X_0, X_1, \ldots\}$ is i, that is:

$$X_0 = i.$$

The time to first return to state i, τ_{ii} is defined as:

$$\tau_{ii} = \min\{t > 0 : X_t = i | X_0 = i\}$$

- An irreducible Markov chain $\{X_0, X_1, \ldots\}$ is said to be recurrent if

$$P(\tau_{ii} < \infty) = 1 \qquad \text{for all } i \in \mathcal{X}$$

- In a recurrent Markov chain we are almost certain that the chain will return to its initial state in a finite amount of time. The only time when this will not happen is if the Markov chain starts in a pathological state, and this is a set with measure 0.

Positive recurrence

- An irreducible Markov chain $\{X_0, X_1, \ldots\}$ is called positive recurrent if

$$E(\tau_{ii}) < \infty \qquad \text{for all } i \in \mathcal{X}$$

- A positive recurrent Markov chain is expected to return to its initial state in a finite amount of time.
- An irreducible Markov chain $\{X_0, X_1, \ldots\}$ is positive recurrent if and only if there exists a stationary distribution π for the chain.
- Let S be a non-null set of non-zero integers, we denote with gcd(S) the greatest common divisor of the elements in S
- A state $i \in \mathcal{S}$ is said to have period k if

$$k = \gcd\{t > 0 : p_{ii}(t) > 0\}$$

- A state i has period k if any return to state i must happen at times that are multiple of k.

Aperiodicity and ergodicity

- A state $i \in \mathcal{X}$ is said to be aperiodic if $k = 1$.
- This implies that the Markov chain does not return to state i in a predictable manner.
- An irreducible Markov chain $\{X_0, X_1, \ldots\}$ is called aperiodic if every state $i \in \mathcal{S}$ is aperiodic.
- If a Markov chain is aperiodic this means that the chain does **NOT** oscillate between states in a predictable manner.
- A Markov chain $\{X_0, X_1, \ldots\}$ is called ergodic if it is positive recurrent and aperiodic.
- If a Markov chain $\{X_0, X_1, \ldots\}$ is irreducible and ergodic, then it has a unique stationary distribution π given by:

$$\pi(j) = \lim_{t \to \infty} p_{ij}(t) = \lim_{t \to \infty} P(X_t = j | X_0 = i)$$

Aperiodicity and ergodicity

- The previous result tells us that if we have a Markov chain $\{X_0, X_1, \ldots\}$, then:
  - if we can reach every state in $\mathcal{X}$ from any other state in a finite amount of time (irreducible)
  - if there is no predictable pattern in the chain (aperiodic)
  - if we expect to return to any state in a finite amount of time (positive recurrent)

  then, no matter where we start the chain, the chain will eventually reach its stationary distribution π. Once the stationary distribution is reached, all the states will be distributed according to π.
- Thus, our goal is to generate irreducible and ergodic Markov chains that admit the posterior distribution as stationary distribution.
- Then, in light of what we have seen, no matter where we start the chain, we eventually will be sampling from the stationary distribution.
- The stationary distribution is what the textbook calls the target distribution.

Ergodic theorem

- If $\{X_0, X_1, \ldots\}$ is an irreducible and ergodic Markov chain with stationary distribution π and f is a function that is absolutely integrable with respect to π, that is:

$$E_\pi(|f|) < \infty$$

then, if f_N denotes the ergodic average $f_N = \frac{\sum_{i=1}^{N} f(X_i)}{N}$, it results that:

$$f_N \to E_\pi(f) \qquad \text{almost surely}$$

where $E_\pi f = \sum_{i \in \mathcal{X}} f(i) \cdot \pi(i)$.

- If the sample space $\mathcal{X}$ is not discrete, then: $E_\pi f = \int_{x \in \mathcal{X}} f(x) \cdot \pi(x) dx$.

Putting it altogether

- Now we link the theory of Markov chain with the Gibbs sampling algorithm.
- We have seen that the Gibbs sampling algorithm produces a Markov chain: $\phi^{(0)}, \phi^{(1)}, \ldots$.
- Here, now the sample space $\mathcal{X}$ is not discrete, but it is continuos and "equal" to $\mathbb{R}^d$.
- The t-step transition probabilities are then:

$$p_{AB}(t) = P(\phi^{(t)} \in A | \phi^{(0)} \in B)$$

- From previous results, we know that if the Gibbs sampling algorithm produces a Markov chain that is irreducible and ergodic, then the chain $\phi^{(0)}, \phi^{(1)}, \ldots$ admits a unique stationary distribution given by:

$$\lim_{t \to \infty} p_{AB}(t) = \lim_{t \to \infty} P(\phi^{(t)} \in A | \phi^{(0)} \in B)$$

Markov chain Monte Carlo (MCMC)

- The Gibbs sampling algorithm does produce such a chain and its stationary distribution is the posterior distribution $p(\phi|y_1,\ldots,y_n)$
- Therefore, no matter what the starting value $\phi^{(0)}$ is, the sampling distribution of $\phi^{(s)}$ will eventually approach the stationary distribution.
- By the ergodic theorem, then for any function g that is absolutely integrable with respect to the posterior distribution $p(\phi|y_1,\ldots,y_n)$:

$$\frac{1}{S}\sum_{s=1}^{S} g(\phi^{(s)}) \to E_{p(\phi|\text{data})}[g(\phi)] = \int g(\phi)p(\phi|y_1,\ldots,y_n)d\phi \qquad \text{as } S \to \infty$$

- This means that we can approximate the posterior mean, posterior median, credible intervals using sample averages as done in the case of Monte Carlo approximation.
- For this reason, we call this approximation of the posterior distribution by sampling from the full conditional distributions Markov chain Monte Carlo (MCMC) approximation and the procedure to obtain it an MCMC algorithm.

Bayesian modeling inference

- In performing inference in a Bayesian context, often the algorithmic part is confused with the statistical modeling part. To clarify things let's review the steps of Bayesian modeling and inference.
- *Bayesian modeling*
  - we specify a sampling model for the data: $y_1, \ldots, y_n | \phi_1, \ldots, \phi_p \sim p(y|\phi)$ where $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$ is a p-dimensional parameter vector with sample space $\Phi \subset \mathbb{R}^p$
  - we specify a prior on the parameter vector ϕ.
- *Bayesian inference*
  - we derive the posterior distribution $p(\phi|y_1, \ldots, y_n)$ of ϕ given the data using Bayes' theorem:

$$p(\phi|y_1, \ldots, y_n) = \frac{p(y_1,\ldots,y_n|\phi)\cdot p(\phi)}{p(y_1,\ldots,y_n)} = \frac{p(y_1,\ldots,y_n|\phi)\cdot p(\phi)}{\int_{\phi\in\Phi} p(y_1,\ldots,y_n|\phi)\cdot p(\phi) d\phi}$$

$$\propto p(y_1, \ldots, y_n|\phi) \cdot p(\phi)$$

Bayesian inference

- Once the posterior distribution has been derived (up to a proportionality constant, equal to $\frac{1}{p(y_1,\ldots,y_n)} = \frac{1}{\int_{\phi\in\Phi} p(y_1,\ldots,y_n|\phi)\cdot p(\phi)d\phi}$, two situations can arise
  1. the posterior distribution $p(\phi|y_1,\ldots,y_n)$ is a known distribution (i.e. we recognize the kernel of a known distribution $\Longrightarrow$ we implicitly derive the proportionality constant)
  2. the posterior distribution $p(\phi|y_1,\ldots,y_n)$ is **NOT** a known distribution
- In the first case, we can obtain posterior summaries of the posterior distributions using results for the known distribution.
  We can derive summaries, such as: posterior means, posterior medians, credible/confidence intervals, posterior quantiles, posterior variances/standard deviations, etc.

Approximation to $p(\phi|y_1, \ldots, y_n)$

- If we can't identify the posterior distribution with a known distribution, we need to approximate the posterior distribution $p(\phi|y_1, \ldots, y_n)$.
  We can do this in two ways:
  1. numerical techniques or numerical approximation
  2. stochastic simulation or Monte Carlo approximation
- Numerical approximation of the posterior distribution $p(\phi|y_1, \ldots, y_n)$ consists mainly in approximating the density with a discretized version $p_D(\phi|y_1, \ldots, y_n)$, obtained by choosing a dense grid over the sample space $\Phi \subset \mathbb{R}^p$, evaluating $p(y_1, \ldots, y_n|\phi) \cdot p(\phi)$ over the grid and normalizing it to obtain a density that sums up to 1 over the grid
  - it works well when the dimension p of the parameter vector is not very large;
  - it is computationally more demanding than Monte Carlo approximation when p is large.

Monte Carlo approximation to $p(\phi|y_1, \ldots, y_n)$

- Monte Carlo approximation to the posterior distribution consists into approximating $p(\phi|y_1, \ldots, y_n)$ by simulation.
- We can distinguish two different ways of doing that:
  1. generating independent samples directly from the posterior distribution $p(\phi|y_1, \ldots, y_n)$
     This is what we do in Monte Carlo approximation to $p(\phi|y_1, \ldots, y_n)$
  2. generating dependent samples, in particular a Markov chain, that admits the posterior distribution $p(\phi|y_1, \ldots, y_n)$ as its stationary distribution
     [the algorithm needs to generate an irreducible and ergodic Markov chain whose t-step transition probabilities have as limit for $t \to \infty$ the posterior distribution $p(\phi|y_1, \ldots, y_n)$]
     This is what we do when we perform Markov chain Monte Carlo approximation to $p(\phi|y_1, \ldots, y_n)$
     The Gibbs sampling algorithm is an MCMC algorithm where the Markov chain is obtained by sampling from the full conditional distributions.

MC vs MCMC approximation

- Monte Carlo (MC) and Markov chain Monte Carlo (MCMC) approximation are different for two main aspects
  - MC generates independent samples while MCMC generates dependent samples
  - the samples generated via MC **are** distributed according to the posterior distribution; the distribution of the samples generated via MCMC algorithm instead **converges** to the posterior distribution.
- We illustrate the difference between MC approximation and MCMC approximation by looking at the example of approximating a joint distribution (not posterior) of two random variables, a discrete one, δ, and a continuous, θ.

MC vs MCMC approximation: example

- Let δ be a discrete random variable that can take values $\{1,2,3\}$ with probability, respectively:

$$P(\delta=1)=0.45 \qquad P(\delta=2)=0.1 \qquad P(\delta=3)=0.45$$

- Let θ be a continuos random variable that, conditionally on δ, has distribution $N(\theta;\mu_\delta,\sigma^2_\delta)$ with $\delta=1,2,3$ where:

$$(\mu_1,\sigma^2_1)=(-3,\frac{1}{3}) \qquad (\mu_2,\sigma^2_2)=(0,\frac{1}{3}) \qquad (\mu_3,\sigma^2_3)=(\ 3\ ,\frac{1}{3})$$

- So, we have a mixture of normal densities where δ denotes the mixture membership.
- In fact, the marginal distribution of θ is given by:

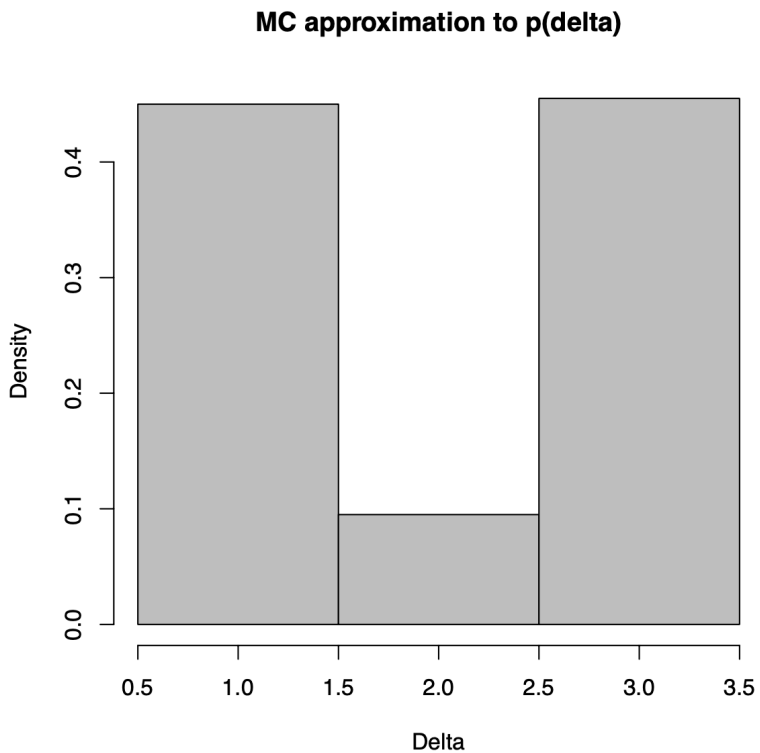
$$\begin{aligned} p(\theta) &= \textstyle\sum_\delta p(\theta,\delta)=\sum_\delta p(\theta|\delta)\cdot p(\delta) \\ &= 0.45\cdot N(\theta;-3,\tfrac{1}{3})+0.1\cdot N(\theta;0,\tfrac{1}{3})+0.45\cdot N(\theta;3,\tfrac{1}{3}) \end{aligned}$$

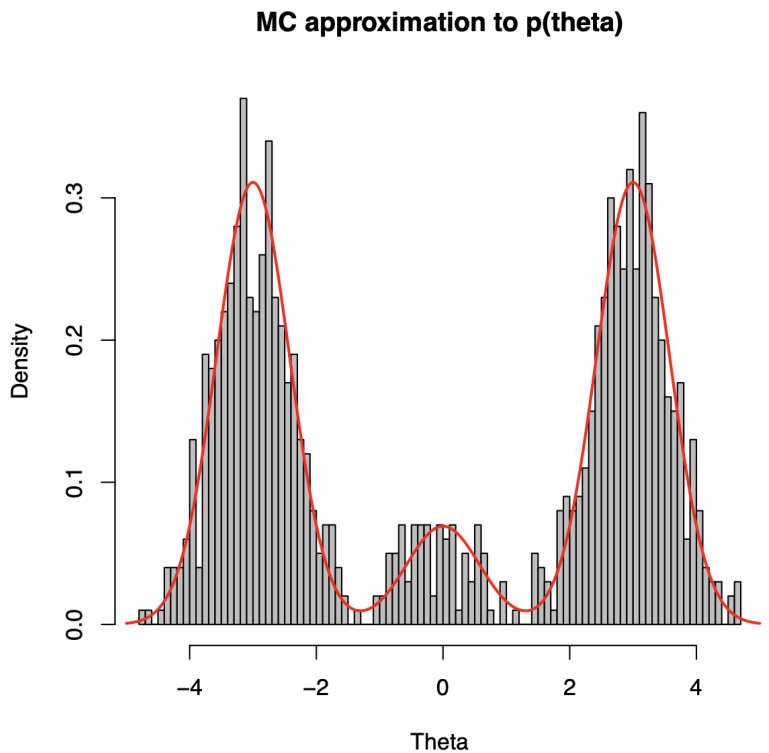
MC vs MCMC approximation: example

- We can approximate the distribution $p(\theta, \delta)$ via Monte Carlo very easily: we choose a number of iterations B and we alternate between the two steps:
  - sample $\delta^{(i)}$ from a discrete distribution on $\{1, 2, 3\}$
  - sample $\theta^{(i)}$ from the corresponding normal distribution $N(\mu_{\delta^{(i)}}, \sigma^2_{\delta^{(i)}})$
- We can then look at the samples only for θ: $(\theta^{(1)}, \ldots, \theta^{(B)})$: their empirical distribution provide an approximation to $p(\theta)$

MC vs MCMC approximation: example

Monte Carlo approximation to $p(\delta)$ and $p(\theta)$: here $B = 1,000$.





MC vs MCMC approximation: example

- Now, let's construct a Gibbs sampling algorithm to approximate the distribution $p(\theta, \delta)$.
- By definition, the algorithm alternates sampling from the two full conditionals.
  We know the full conditional distribution for θ: this is $p(\theta|\delta) = N(\theta; \mu_\delta; \sigma^2_\delta)$.
- The full condtional for δ is $p(\delta|\theta)$:

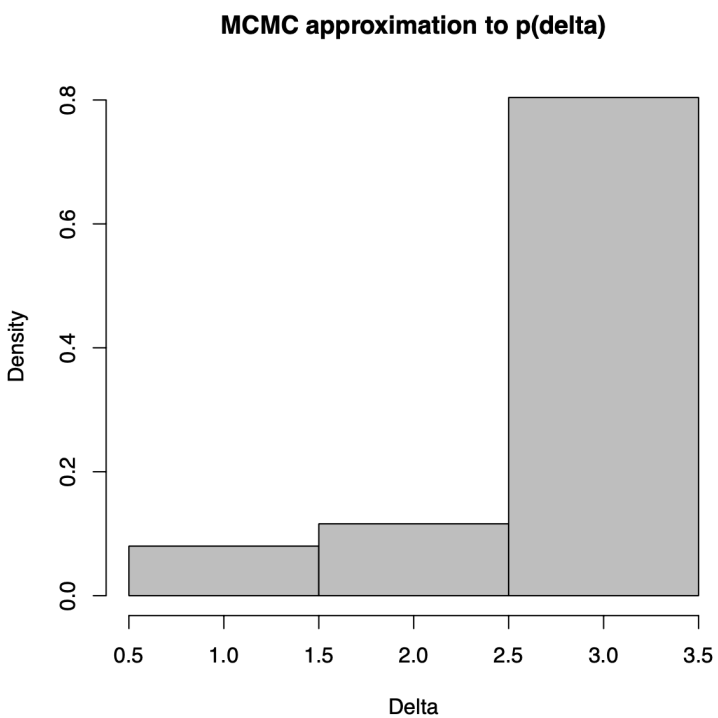
$$\begin{aligned} p(\delta|\theta) &= \frac{p(\delta,\theta)}{p(\theta)} = \frac{p(\theta|\delta)\cdot p(\delta)}{p(\theta)} \\ &= \frac{p(\theta|\delta)\cdot p(\delta)}{\sum_\delta p(\theta|\delta)\cdot p(\delta)} \qquad \delta = 1,2,3 \end{aligned}$$

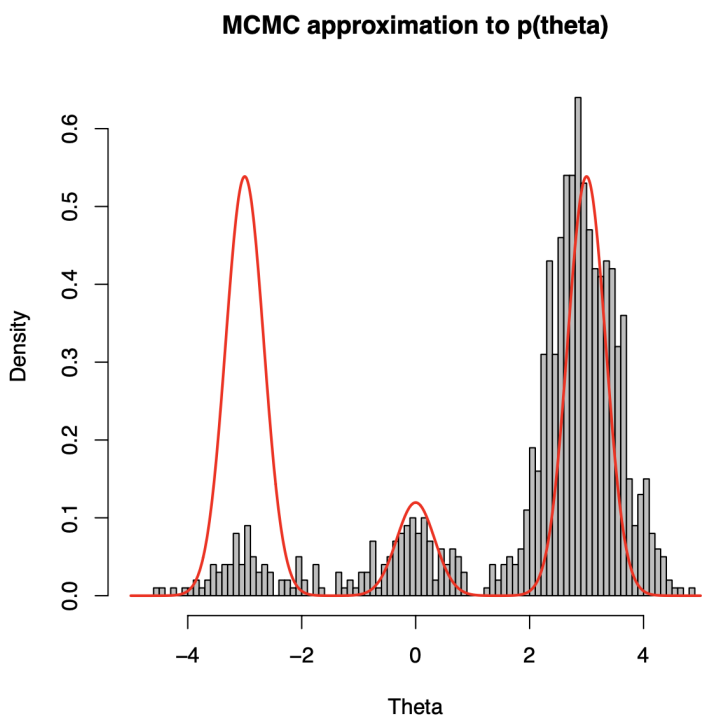
- Therefore: for $\delta = 1$

$$\begin{aligned} p(\delta = 1|\theta) &= \frac{p(\theta|\delta{=}1)\cdot p(\delta{=}1)}{\sum_\delta p(\theta|\delta)\cdot p(\delta)} \\ &= \frac{0.45\cdot N(\theta; -3, \frac{1}{3})}{0.45\cdot N(\theta; -3, \frac{1}{3}) + 0.1\cdot N(\theta; 0, \frac{1}{3}) + 0.45\cdot N(\theta; 3, \frac{1}{3}).} \end{aligned}$$

MC vs MCMC approximation: example

- We run our Gibbs sampling algorithm for $S = 1,000$ iterations and look at the MCMC approximation of the marginal distribution $p(\delta)$ and $p(\theta)$ by looking at the empirical distribution of the samples $(\delta^{(1)}, \ldots, \delta^{(S)})$ and $(\theta^{(1)}, \ldots, \theta^{(S)})$, respectively.





MC vs MCMC approximation: example

- The approximation provided by the Monte Carlo method is pretty good, but the approximation provided by the MCMC algorithm is not great. What happened?
- The problem is that the while Monte Carlo method generates samples by sampling directly and independently from the target distribution, the MCMC algorithm generates a Markov chain whose transition probabilities converge to the target distribution.
- So, when we use an MCMC algorithm we need to check whether the Markov chain **has achieved its stationary distribution i.e. whether it has converged**.
- There are different ways to assess that.

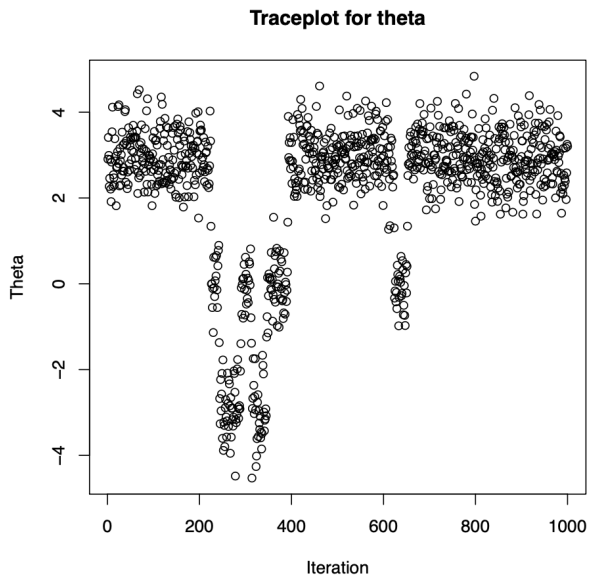
MCMC diagnostics

- The first and easiest diagnostic that one can consider is a traceplot, that is, a plot of the sampled values for each parameter/random variable vs. the iteration number.
- This plot can show if a parameter gets "stuck" in certain regions of the sample space. If that happens, we say that the Markov chain display stickiness, that there is a high autocorrelation in the Markov chain or that the Markov chain has poor mixing.
- The autocorrelation is the correlation between consecutive values in a sequence. In R, the autocorrelation in a sequence of values $\left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}\right)$ can be computed using the function `acf` on the vector with the sequence of values.
- In particular, the lag-t autocorrelation function for a sequence $\left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}\right)$ is given by:

$$\text{acf}_t(\theta) = \frac{\frac{1}{S}\sum_{j=1}^{S-t}(\theta_j - \bar{\theta})(\theta_{j+t} - \bar{\theta})}{\frac{1}{S}\sum_{j=1}^{S}(\theta_j - \bar{\theta})^2}$$

where $\bar{\theta}$ is the sample mean for the sequence $\left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}\right)$.

MCMC diagnostics



- The chain is sticky and has poor mixing: θ gets stuck in certain regions and rarely moves among all the regions.
- The lag-1 autocorrelation for the sequence of sampled θ values obtained via Gibbs sampling is 0.9 and it is approximately 0.6 at lag 29!
- Conversely, the lag-1 autocorrelation for the sequence of sampled θ values obtained via Monte Carlo sampling is -0.01, i.e. the sampled θ values are independent.

MCMC diagnostics

- Mixing refers to the degree of autocorrelation in the sequence of sampled values. A Monte Carlo sampler always produces sequences of sampled values that have good mixing since they have virtually no autocorrelation.
- Poor mixing affects inference.
- Suppose that we have generated samples $\left(\theta_{MC}^{(1)}, \ldots, \theta_{MC}^{(B)}\right)$ using a Monte Carlo sampler and we have generated samples $\left(\theta_{MCMC}^{(1)}, \ldots, \theta_{MCMC}^{(S)}\right)$ using an MCMC algorithm. In both cases, the two algorithms are providing an approximation to a posterior distribution.
- Then, we know that we can approximate the posterior mean $\theta_0 = \int \theta \cdot p(\theta|\text{data}) d\theta$ using the sample mean $\bar{\theta}_{MC}$ of $\left(\theta_{MC}^{(1)}, \ldots, \theta_{MC}^{(B)}\right)$ and, by the ergodic theorem, we know we can approximate θ_0 via the sample mean $\bar{\theta}_{MCMC}$ of $\left(\theta_{MCMC}^{(1)}, \ldots, \theta_{MCMC}^{(S)}\right)$.

MCMC diagnostics

- The variability of the estimate $\bar{\theta}_{MC}$ provided by the Monte Carlo approximation is $\text{Var}_{\text{MC}}\left[\bar{\theta}_{MC}\right]$:

$$\text{Var}_{\text{MC}}\left[\bar{\theta}_{MC}\right] = E\left(\bar{\theta}_{MC} - \theta_0\right)^2 = \frac{\text{Var}(\theta_{MC})}{B}$$

  where $\text{Var}(\theta_{MC})$ is the variance of the sequence $\left(\theta_{MC}^{(1)}, \ldots, \theta_{MC}^{(B)}\right)$.

- $\sqrt{\text{Var}_{\text{MC}}\left[\bar{\theta}\right]}$ is the Monte Carlo standard error and it gets smaller as the number of simulations, B increases. It also provides a standard error for the estimator.
- A 95% confidence interval for the posterior mean θ_0 using a Monte Carlo approximation is given approximately by:

$$\bar{\theta}_{MC} \pm 1.96\sqrt{\text{Var}_{\text{MC}}\left[\bar{\theta}_{MC}\right]}$$

- This interval gets tighter as we increase B.

MCMC diagnostics

- On the other hand, the variability of the estimate $\bar{\theta}_{MCMC}$ provided by the Markov chain Monte Carlo approximation is:

$$\begin{aligned}
\text{Var}_{\text{MCMC}}\left[\bar{\theta}_{MCMC}\right] &= E\left[\bar{\theta}_{MCMC}-\theta_0\right]^2 = E\left[\left(\frac{1}{S}\left(\sum_{i=1}^{S}\theta_{MCMC}^{(i)}-\theta_0\right)\right)^2\right] \\
&= \frac{1}{S^2}E\Bigg[\sum_{i=1}^{S}\left(\theta_{MCMC}^{(i)}-\theta_0\right)^2 + \\
&+ \sum_{i\neq j}\left(\theta_{MCMC}^{(i)}-\theta_0\right)\left(\theta_{MCMC}^{(j)}-\theta_0\right)\Bigg] \\
&= \frac{1}{S}\text{Var}(\theta_{MCMC}) + \frac{S\cdot(S-1)}{S^2}\text{Cov}(\theta_{MCMC}) \\
&> \frac{1}{S}\text{Var}(\theta_{MCMC})
\end{aligned}$$

since $\text{Cov}(\theta_{MCMC}) > 0$ as the Markov chain is generally positively correlated.

MCMC diagnostics

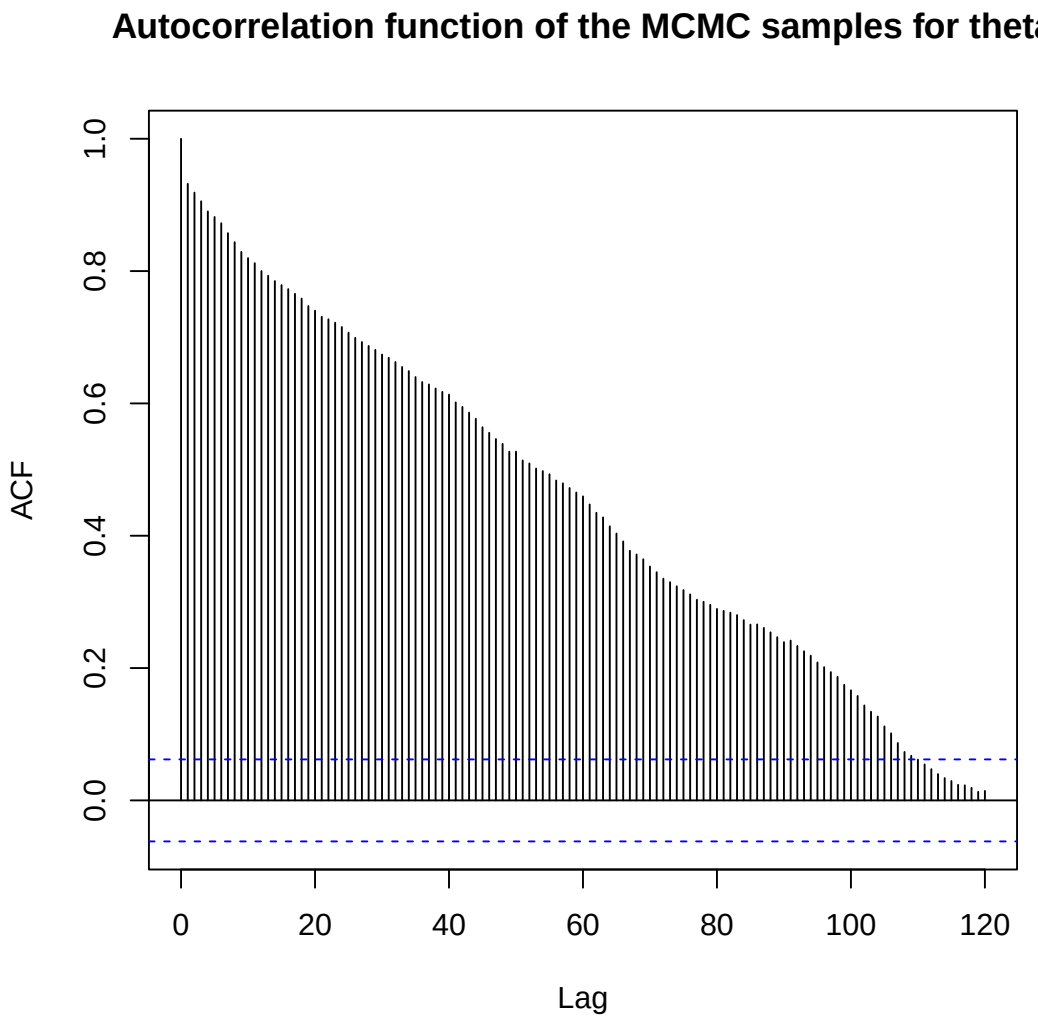
- Therefore, the variability in the estimate $\bar{\theta}_{MCMC}$ of the posterior mean provided by the Markov chain Monte Carlo approximation is **larger** than that of the MC estimate $\bar{\theta}_{MC}$.
- The higher the autocorrelation, the more MCMC samples need to be taken to attain a certain level of precision for our approximation.
- One way to measure the number of MCMC samples needed to achieve a certain precision is to compute the effective sample size.
- This can be computed in `R` using the command `effectiveSize` in the `coda` pacakge.
- The effective sample size calculates the number of independent Monte Carlo samples necessary to give the same precision as Monte Carlo samples.
- For example, in the MCMC sample we have generated earlier, the effective sample size is: 7.16. This means that the precision in the 1,000 MCMC samples of θ is the same as the precision that we would have obtained in approximately 7 independent sample of θ!

MCMC diagnostics

- A high autocorrelation in the MCMC samples means that the Markov chain moves in the sample space very slowly, taking a lot of time to reach the stationary distribution.
- One way to decrease the autocorrelation in the MCMC samples is to run the MCMC chain for a large number S of iterations and then "thin" it or sub-sample it by taking a subsample every b iteration. This procedure is called thinning and the goal is to generate an independent sample.
- How to choose the thinning number T? One way is to produce a plot of the autocorrelation function and take T to be equal to the lag for which the autocorrelation is not significantly different from 0.

MCMC diagnostics

- Plot of the autocorrelation function for the MCMC samples of θ.
- Based on this plot, we could take $T \approx 110$



MCMC diagnostics

- The goal in MCMC methods is to produce a Markov chain that has the posterior distribution as its stationary distribution.
- If the chain has reached its stationary distribution, then samples in one part of the chain have a similar distribution than samples in other parts of the chain.
- Thus, a way to check if the Markov chain has reached its stationary distribution is to take subsample from the chain and verify if the distributions of the subsamples are the same.
- This might not be always easy to do, especially if the parameter vector has large dimension.
- Another strategy is to run the MCMC algorithm several times using different starting values. After a certain amount of time, called burnin, **if** the chains have reached the stationary distribution, they should all converge around the same distributions and the traceplots for each parameter should wiggle around the same values.

MCMC diagnostics

- When we make posterior inference using the output of an MCMC algorithm and we compute for example, posterior means, posterior medians, credible intervals, etc., we don't use the entire MCMC output. We don't want to use those samples where the Markov chain had not reached the stationary distribution.
  Thus, we discard the samples in the burnin period.
- So, if the first b samples are discarded for burnin, then the posterior mean for θ is estimated by:

$$\bar{\theta}_{MCMC} = \frac{1}{S-b} \sum_{i=b+1}^{S} \theta_{MCMC}^{(i)}$$

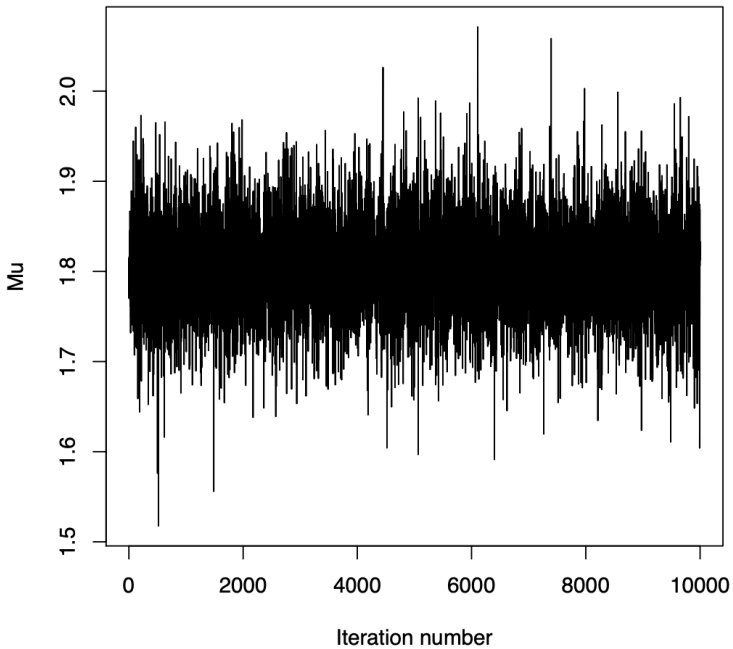
and similarly for all the other posterior summaries.

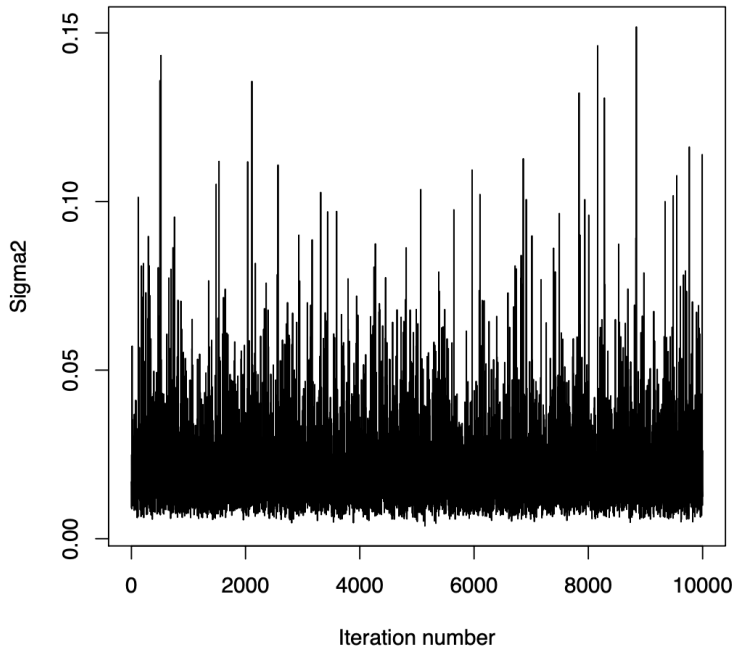
Midge wing length example

- The midge wing length data refers to measurements of wings for 9 flies. Their wings measured, respectively, $y_1 = 1.64$, $y_2 = 1.70$, $y_3 = 1.72$, $y_4 = 1.74$, $y_5 = 1.82$, $y_6 = 1.82$, $y_7 = 1.82$, $y_8 = 1.90$ and $y_9 = 2.08$ millimeters.
- Previously, we used the following priors on μ and σ^2:
  $p(\mu) = N(\mu_0, \tau_0^2) = N(1.9, 0.95^2)$ and
  $p(\sigma^2) = \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) = \text{InverseGamma}(\frac{1}{2}, \frac{0.01}{2})$.
- We used as initial values for μ and σ^2, respectively, $\bar{y} = 1.80$ and $s^2 = 0.017$.
- We ran our Gibbs sampling algorithm in `R` for 10,000 iterations.

Midge wing length example

- Trace plots for μ and σ^2

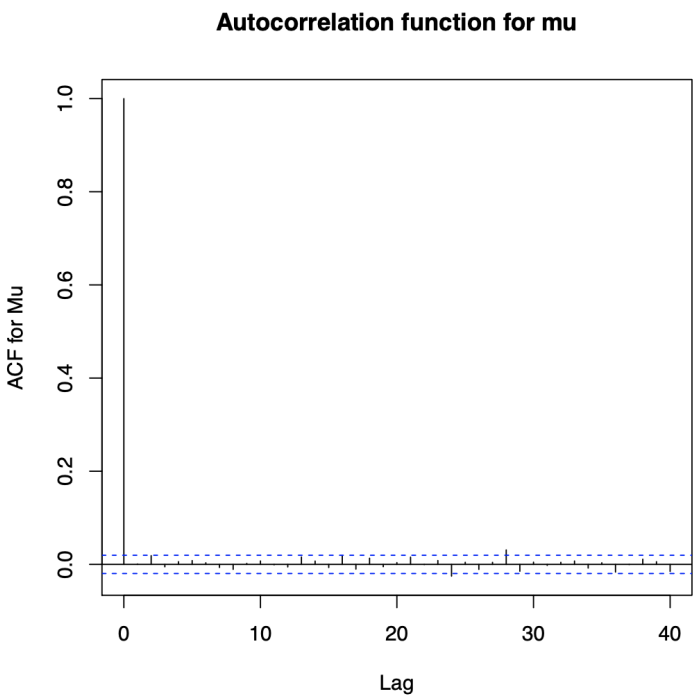


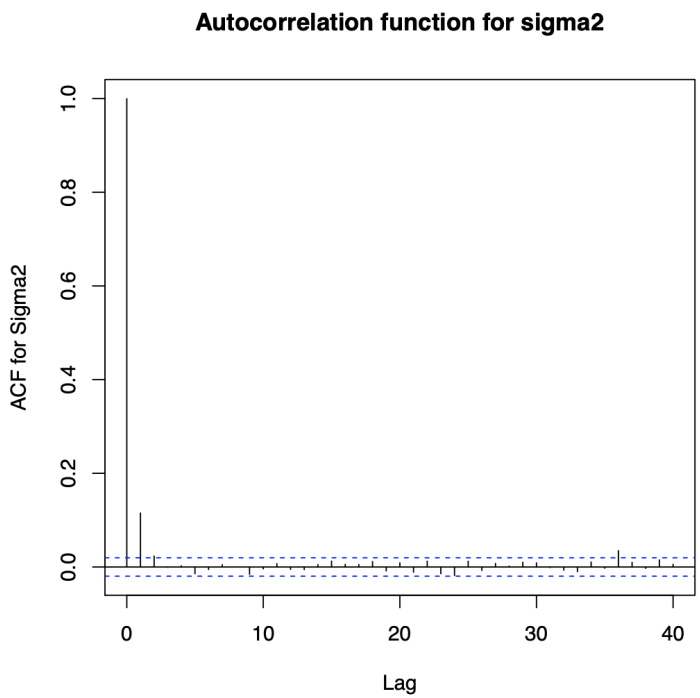


- Looking at the trace plots, we can deduce that either the Markov chains has reached stationarity quickly or there is a high level of autocorrelation in the samples.

Midge wing length example

- Autocorrelation functions for μ and σ^2 MCMC samples.

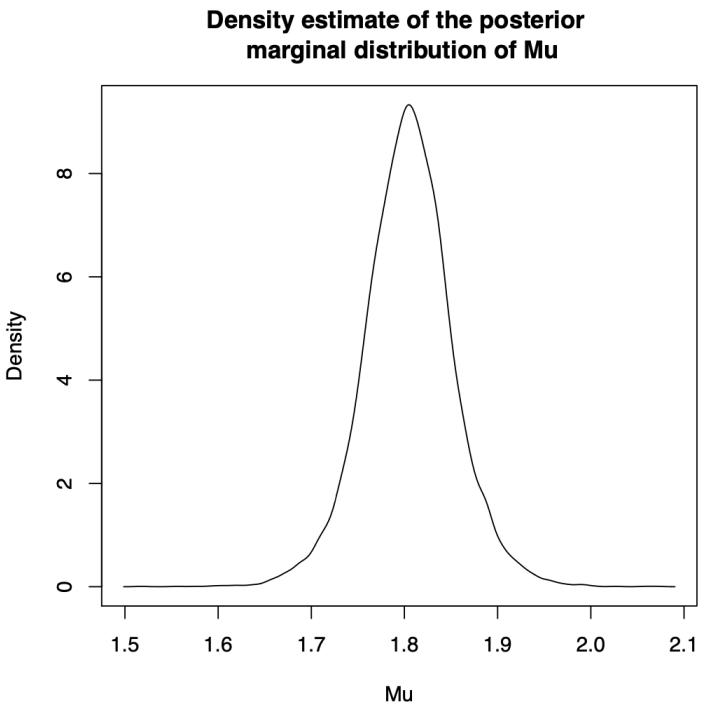


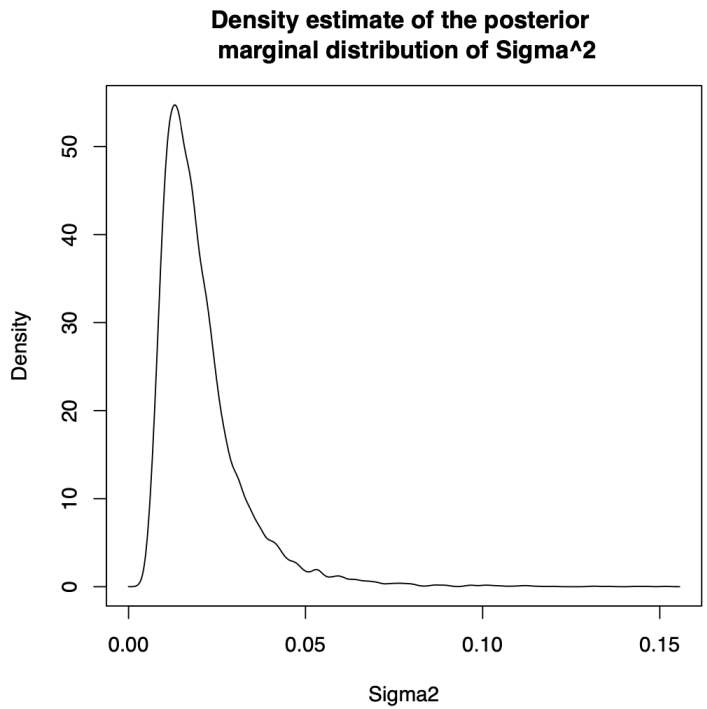


- There is almost no autocorrelation in the samples. Therefore, we do not need to use thinning in this case.

Midge wing length example

- We throw away the first 500 MCMC samples for burnin. The kernel density estimates of the posterior marginal distributions of μ and σ^2 are based on the sampled MCMC values of μ and σ^2 after burnin.





Midge wing length example

- We summarize the two posterior distributions via posterior mean posterior median, 95% credible interval and posterior standard deviation.

Parameter	Posterior Mean	Posterior Median	95% Credible Interval	Posterior standard deviation
μ	1.80	1.80	(1.71; 1.90)	0.048
σ^2	0.021	0.017	(0.007; 0.05)	0.012

- If we need to report an estimate for μ or σ^2, we use either the posterior mean or the posterior median, depending on whether the posterior marginal appears to be skewed or not.
Thus, in this case our estimates for μ or σ^2 are: 1.80 and 0.017.