Some Examples on Gibbs Sampling and Metropolis-Hastings methods

Gibbs Sampler

- Sample a multidimensional probability distribution from *conditional densities*.
- Suppose $d = 2$, $\theta = (\theta_1, \theta_2)$. Set an initial point, $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$.
  1. Generate a value $\theta_2^{(1)}$ from $\pi(\theta_2 | \theta_1 = \theta_1^{(0)}, X = x)$.
  2. Generate a value $\theta_1^{(1)}$ from $\pi(\theta_1 | \theta_2 = \theta_2^{(1)}, X = x)$.
- Two steps give a new value $\theta^{(1)} = (\theta_1^{(0)}, \theta_2^{(0)})$
- After several iterations: $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots,$
- Samples (after training period) correspond to $\pi(\theta_1, \theta_2 | x)$.

Example: Binomial-Beta

- One observation, $X \sim Binomial(n, \theta)$ and $\pi(\theta) = Beta(a, b)$
- Bayes theorem,

$$\pi(\theta|x) \propto \theta^x (1-\theta)^{n-x} \theta^{a-1} (1-\theta)^{b-1}$$

- The joint density for (x, θ),

$$f(x;\theta)\pi(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

- What is the marginal distribution of X? Its a Binomial-Beta distribution,

$$f(x) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)}, x = 0, 1, 2, \ldots, n$$

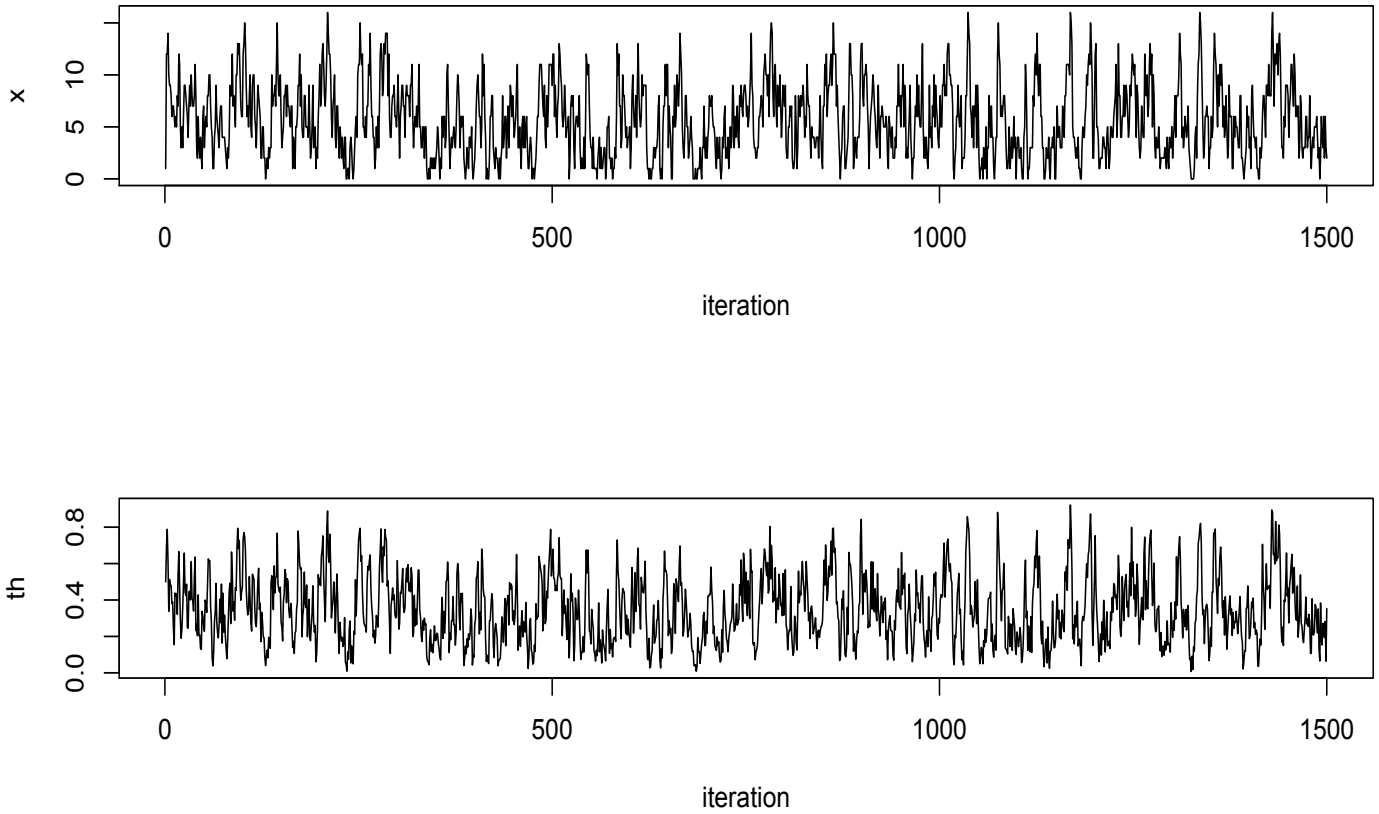
- Via Gibbs sampling. Set initial value $(x^{(0)}, \theta^{(0)})$.
- For $i = 1, 2, \ldots M$
  1. Sample $x^{(i)} \sim f(x|\theta^{(i-1)}) = \mathit{Binomial}(n, \theta^{(i-1)})$
  2. Sample $\theta^{(i)} \sim \pi(\theta|x^{(i)}) = \mathit{Beta}(a + x^{(i)}, b + n - x^{(i)})$
  3. Repeat (1) and (2) many times.
- The process produces samples of (x, θ).
- Approximate $f(x)$ with the values $x^{(i)}$
- For example,

$$f(x) = P(X = x) \approx (\#\ of\ X^{(i)} = x)/M$$

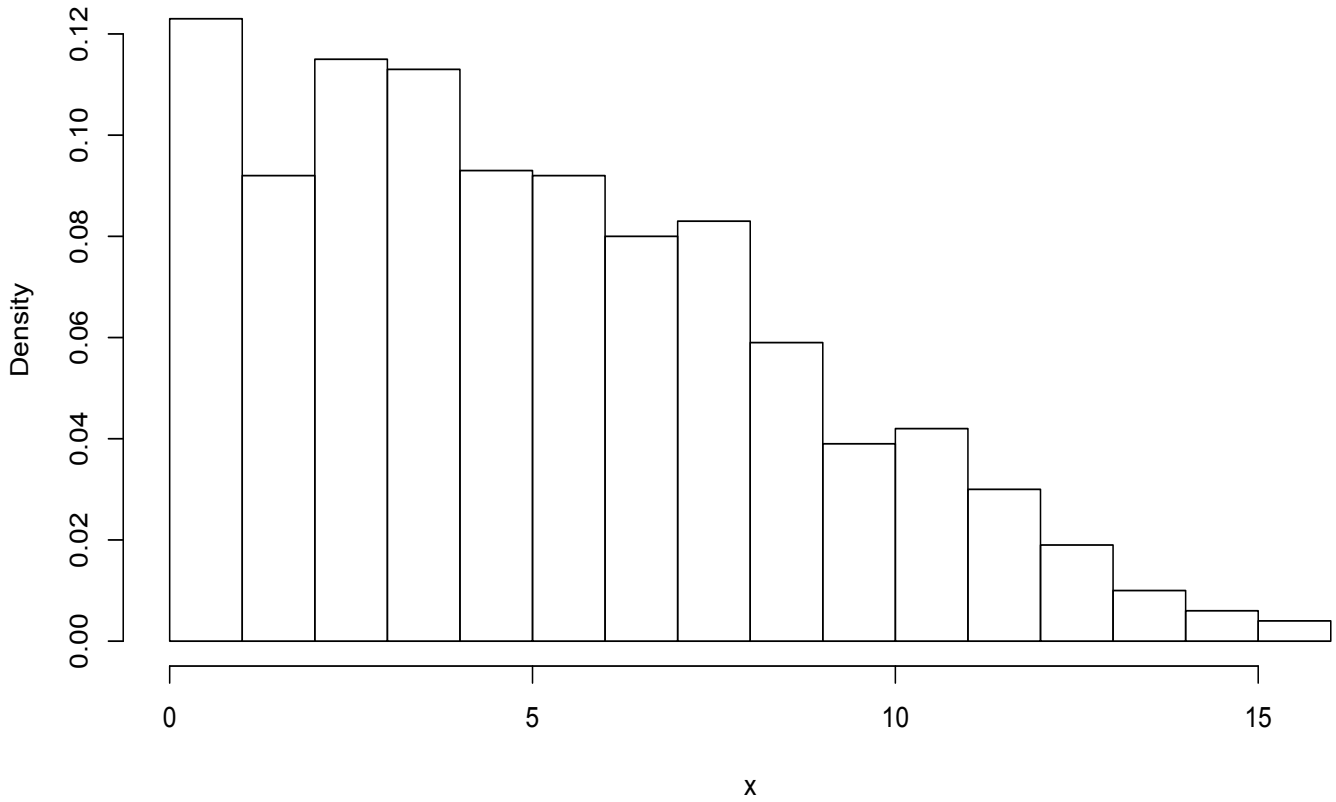
- We may also,
  - (1) Sample $\theta^{(i)} \sim \pi(\theta|x^{(i-1)})$
  - (2) Sample $x^{(i)} \sim f(x|\theta^{(i)})$.

```
# Example 1:Beta-Binomial simulation
# Prior parameters
a=2; b=4; n=16
# prepare objects to save iterations
it =1500
x= rep(NA,it); th=rep(NA,it)
# set initial value
x[1]=1; th[1]=0.5
# Perform Gibbs iterations
for (i in 2:it)
{
  x[i] =rbinom(1,size=n,prob=th[i-1])
  th[i] = rbeta(1,a+x[i],b+n-x[i])
}
```

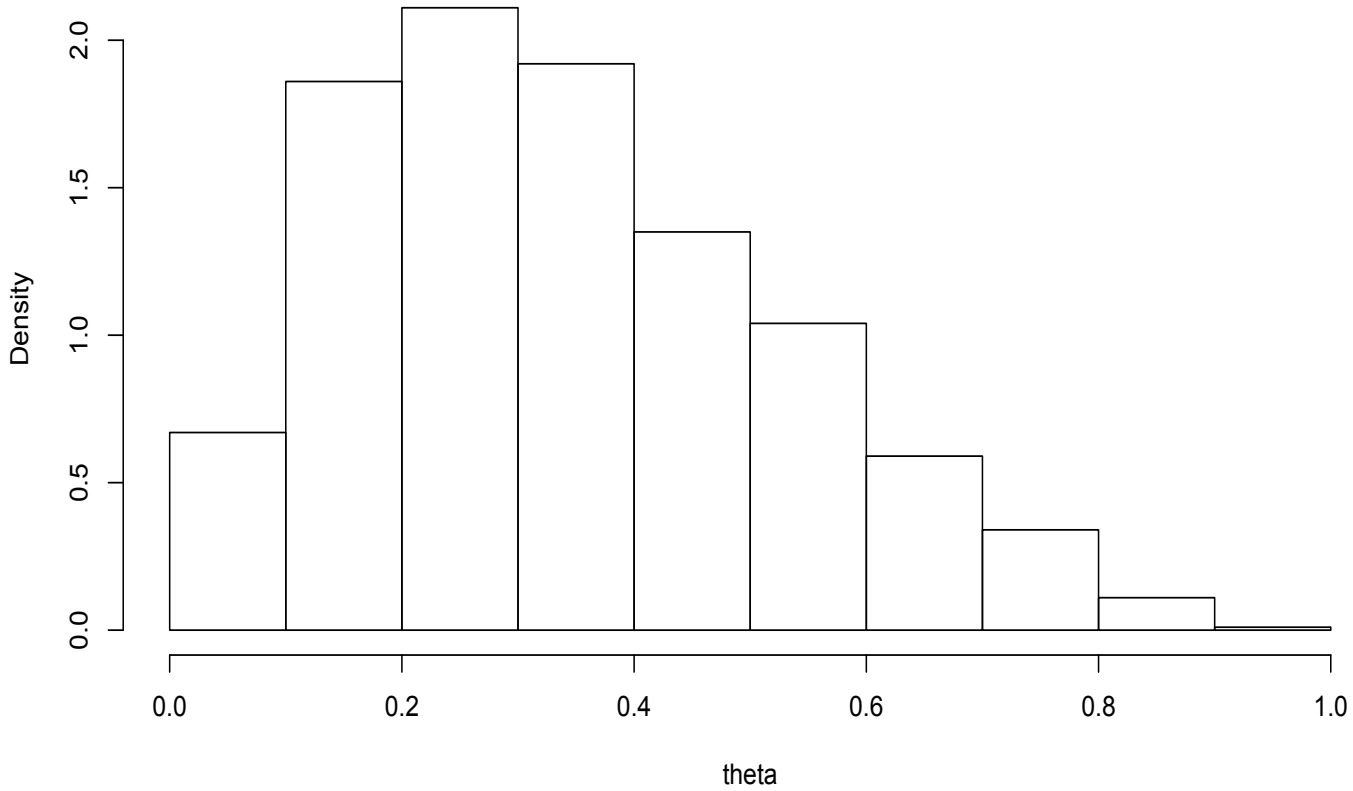
Trace plots for simulations of x and θ



Histogram *x* values (f(x))



Histogram θ values ($\pi(\theta)$)



- In d-dimensions, we iteratively use

$$\pi_i(\theta_i|\theta_{-i}, x)$$

- where $\theta_{-i} = (\theta_1, \theta_2, \ldots, \theta_{i_1}, \theta_{i+1}, \ldots, \theta_d)$.
- Gibbs sampling is an example of *Markov Chain Monte Carlo* (MCMC) methods.
  - A new point $\theta^{(s+1)}$ depends on $\theta^{(s)}$
  - *Monte Carlo*: *pseudo-random* values.
- The *limit* or *stationary* distribution is $\pi(\theta|x)$.
- Gibbs sampling assumes *direct* simulation from full conditionals.
- If full conditional simulation is not possible, then we could use *Metropolis-Hastings*.

Example: Binomial-Beta-Poisson

- Treat n as unknown with $\pi(n) = Poisson(\lambda)$, ($\lambda$ known).
- $X \sim Binomial(n, \theta)$; $\pi(\theta) = Beta(a, b)$
- The joint distribution for (X, θ, n) is:

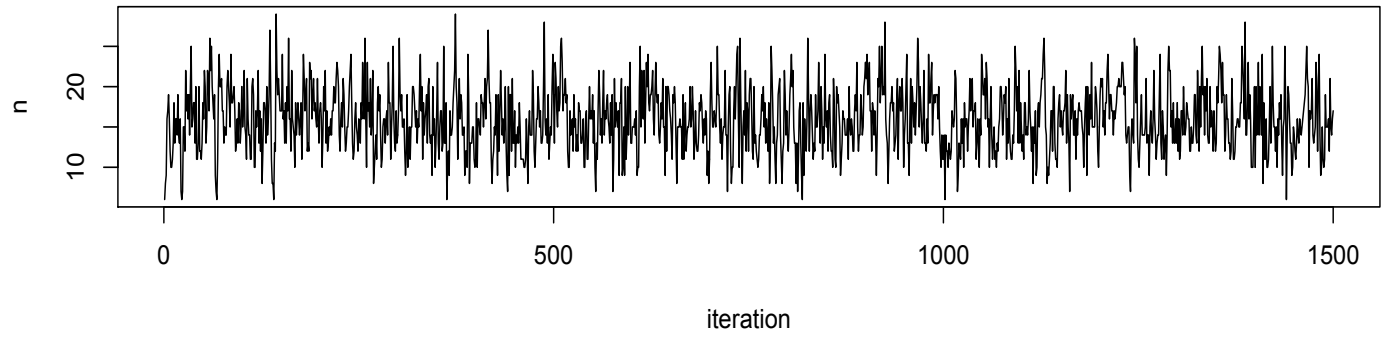
$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\left(e^{-\lambda}\frac{\lambda^n}{n!}\right)$$

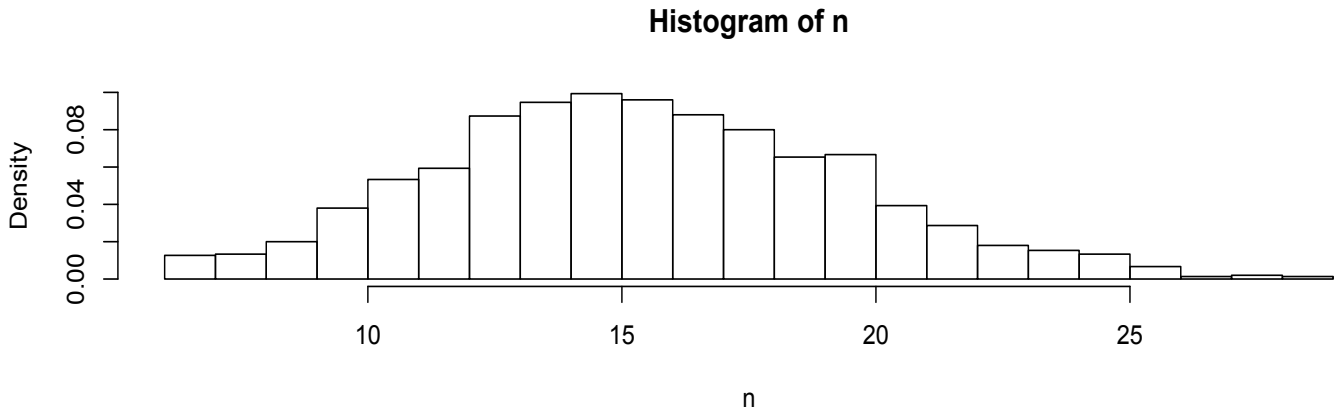
- $x = 0, 1, \ldots, n$; $0 < \theta < 1$; $n = 0, 1, 2, \ldots$.
- Again, interested on $f(x)$ but impossible to find in analytic-closed form.
- Alternatively (get rid of constants),

$$f(\theta, x, n) \propto \binom{n}{x}\theta^{a+x-1}(1-\theta)^{b+n-x-1}\frac{\lambda^n}{n!}$$

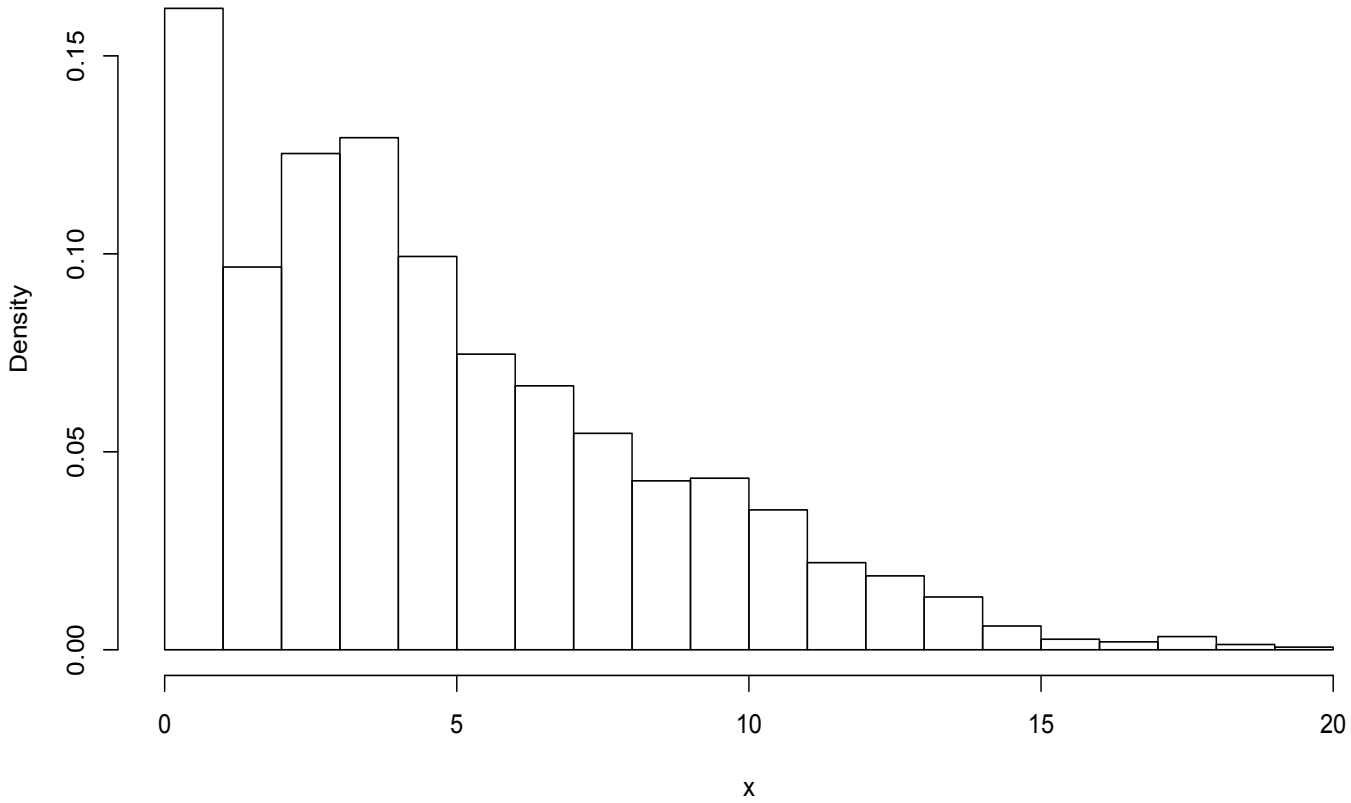
- For Gibbs Sampling, find full conditionals:
- $f(x|\theta, n) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x} \propto Binomial(n, \theta)$,
- $\pi(\theta|x, n) \propto \theta^{a+x-1}(1-\theta)^{b+n-x-1} \propto Beta(a+x, b+n-x)$,
- $\pi(n|\theta, x) \propto \binom{n}{x}\frac{\lambda^n}{n!}(1-\theta)^{n-x} \propto \frac{[\lambda(1-\theta)]^{n-x}}{(n-x)!}; n = x, x+1, x+2, \ldots,$
- If we set, $z = n - x$, then $z \sim Poisson(\lambda(1-\theta))$
- Set $(x^{(0)}, \theta^{(0)}, n^{(0)})$. For $i = 1, 2, 3, \ldots,$
  - Sample $x^{(i)} \sim Binomial(n^{(i-1)}, \theta^{(i-1)})$.
  - Sample $\theta^{(i)} \sim Beta(a + x^{(i)}, b + n^{(i-1)} - x^{(i)})$.
  - Sample $n^{(i)} = x^{(i)} + z, z \sim Poisson(\lambda(1-\theta^{(i)}))$
  - Repeat until *convergence* is reached.

Trace plot and histogram of *n* values





Histogram of *x* values



Example with Bivariate Normal Distribution

- $\mathbf{x} = (x_1, x_2)$, $\mu = (\mu_1, \mu_2)$ and Σ is a 2×2 covariance matrix with diagonal entries σ_1^2, σ_2^2 and off-diagonals $\sigma_{1,2}$. The pdf is

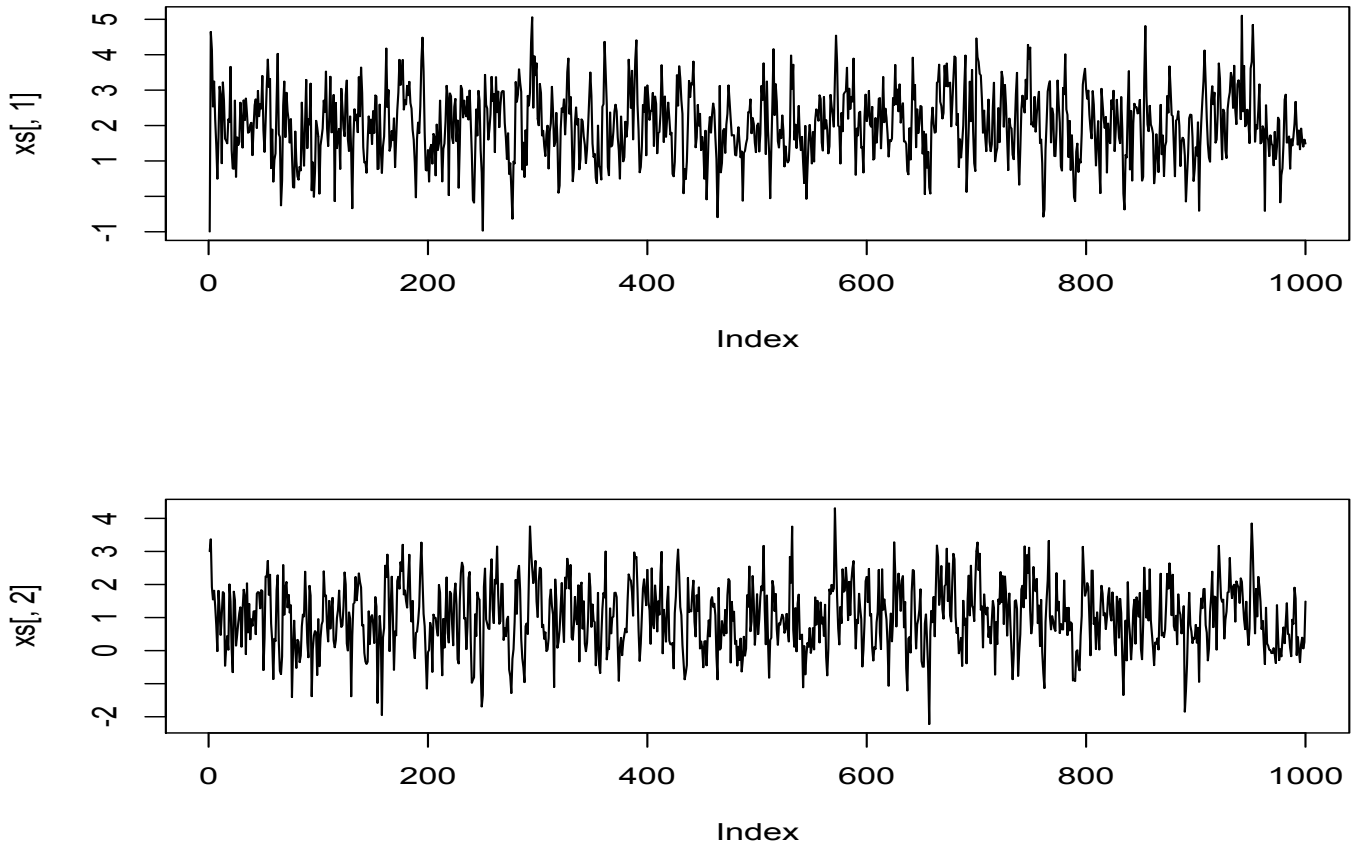
$$f(\mathbf{x}|\mu, \Sigma) \propto |\Sigma|^{-1/2} exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

- Marginal distributions: $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$.
- For *Gibbs sampling*, we need $f(x_1|x_2)$ and $f(x_2|x_1)$. In fact

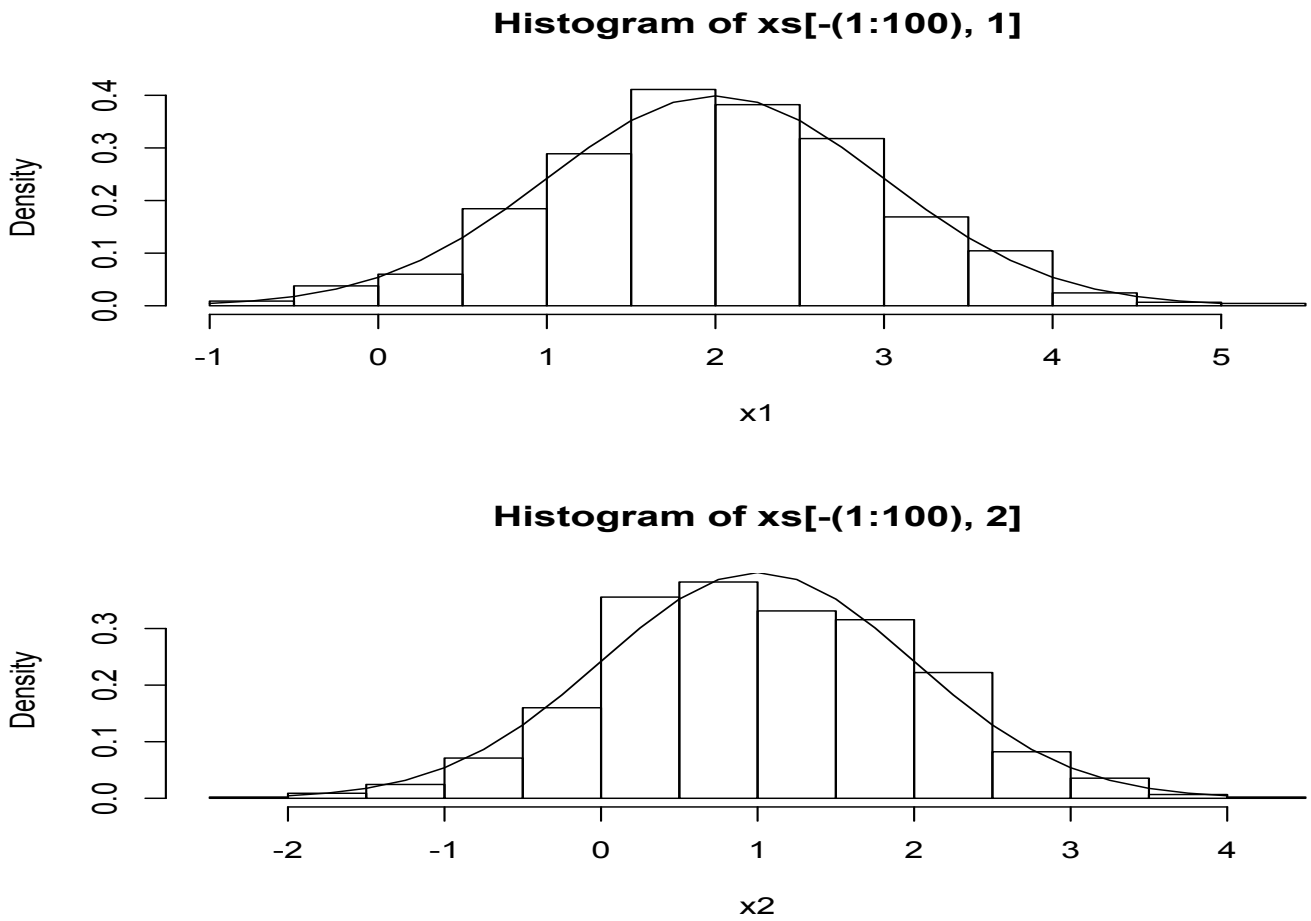
$$\begin{aligned} f(x_1|x_2) &= N(\mu_1 + (\sigma_{1,2}/\sigma_2^2)(x_2 - \mu_2), \sigma_1^2 - (\sigma_{1,2}/\sigma_2)^2) \\ f(x_2|x_1) &= N(\mu_2 + (\sigma_{1,2}/\sigma_1^2)(x_1 - \mu_1), \sigma_2^2 - (\sigma_{1,2}/\sigma_1)^2) \end{aligned}$$

- Example with $\mu = (2, 1)$, $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_{1,2} = 0.7$.

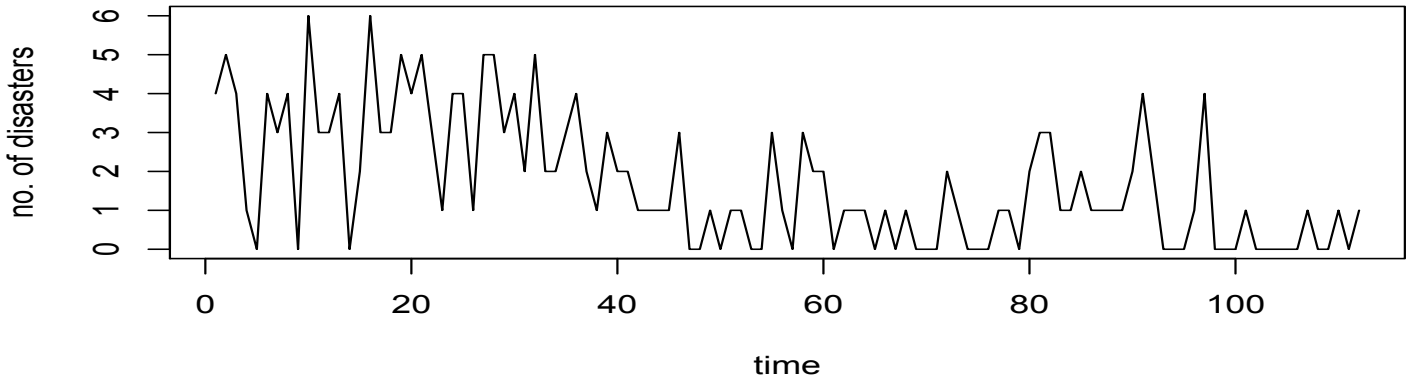
Trace plots for simulations of x_1 and x_2

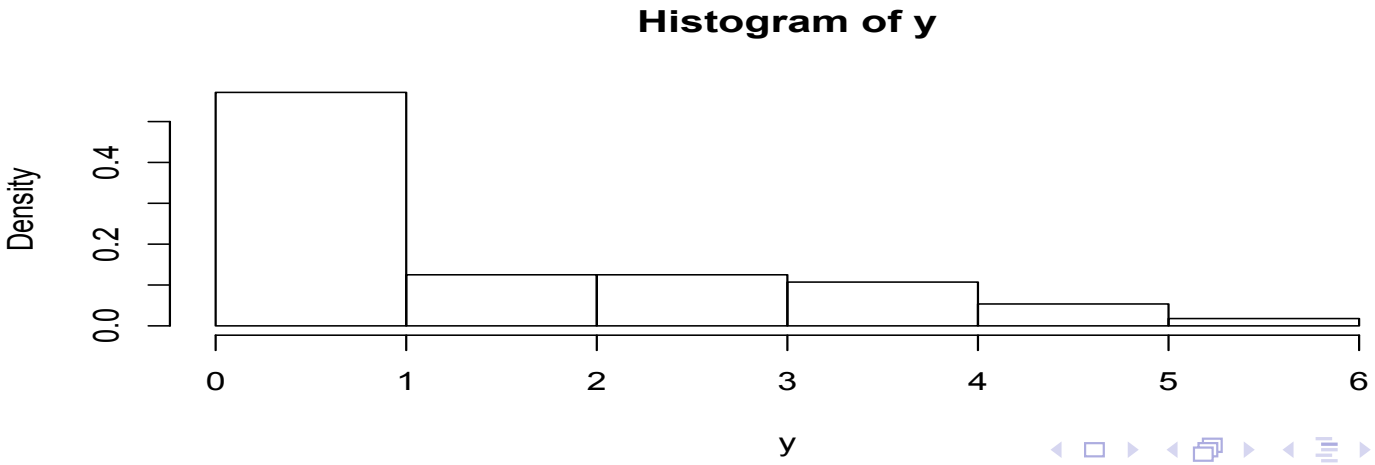


Histograms for simulations of x_1 and x_2



British coal mining disasters data





Example: Poisson process with a Change Point

- *Hierarchical* model proposed in Carlin, Gelfand and Smith (1992) (also see Section 3.10).
- *First stage:*

$$\begin{aligned} X_i &\sim Poisson(\mu), i = 1, 2, \ldots, k. \\ X_i &\sim Poisson(\lambda), i = k+1, k+2, \ldots, m \end{aligned}$$

  so X_i represents the observations and the parameter of interest are (μ, λ, k).
- Data consists of coal-mining disasters in the U.K.
- *Second stage:* Independent priors on (μ, λ, k).
  - k discrete uniform on $\{1, 2, \ldots, m\}$, m=sample size.
  - $\mu \sim Ga(a_1, b_1)$ and $\lambda \sim Ga(a_2, b_2)$
  - a_1, b_1, a_2, b_2 are fixed.

- The joint posterior distribution for (μ, λ, k) has the form:

$$\pi(\mu, \lambda, k|\underline{X}) \propto f(\underline{X}|\mu, \lambda, k)\pi(\mu)\pi(\lambda)\pi(k)$$

- The likelihood function is:

$$\begin{aligned} f(\underline{X}|\theta, \lambda, k) &= \prod_{i=1}^{k} f(x_i|\mu, k) \prod_{i=k+1}^{m} f(x_i|\lambda, k) \\ &= \prod_{i=1}^{k} \frac{\mu^{x_i} e^{-\mu}}{x_i!} \prod_{i=k+1}^{m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \end{aligned}$$

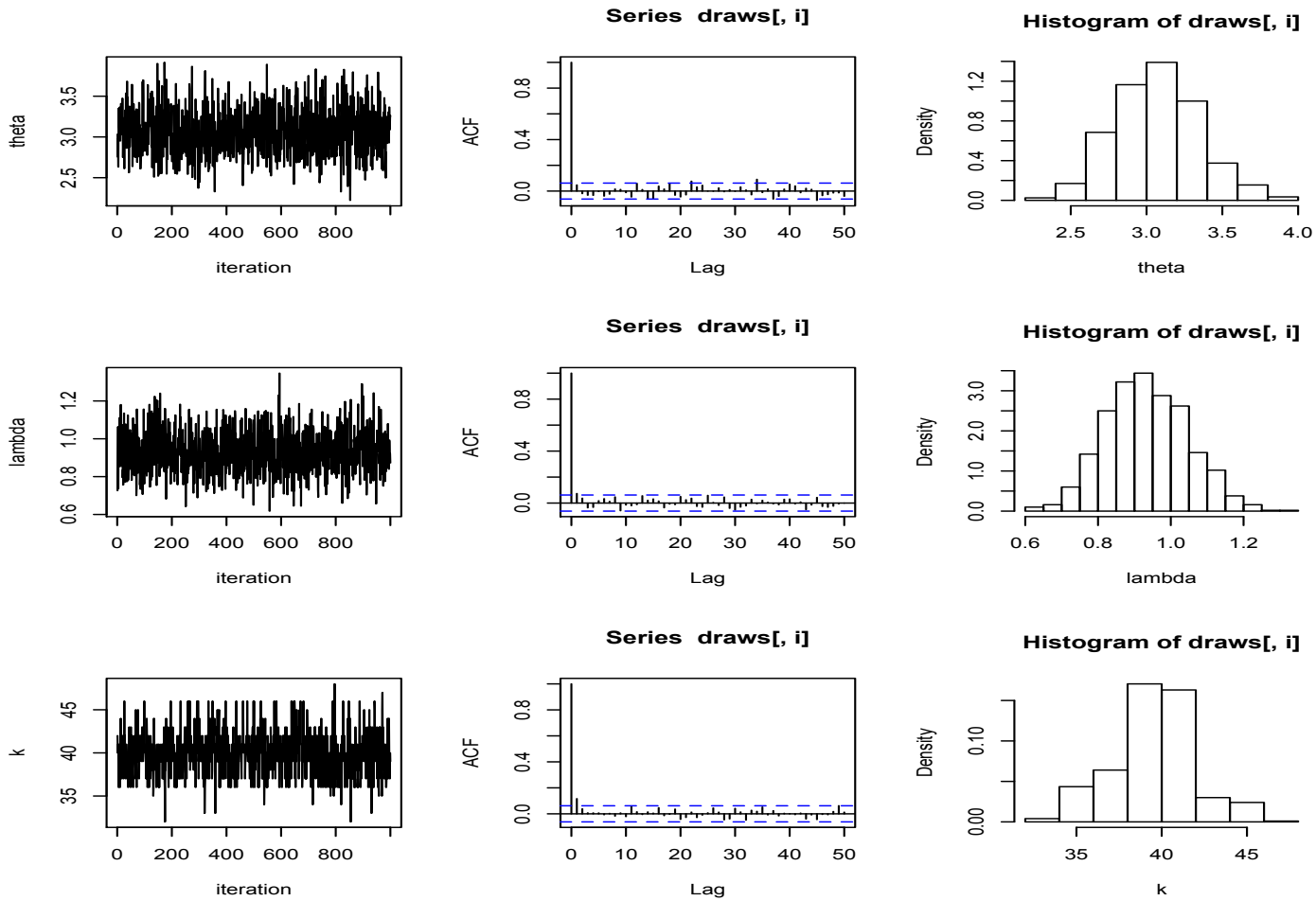
- Therefore,

$$\pi(\mu, \lambda, k|\underline{X}) \propto \prod_{i=1}^{k} \frac{\mu^{x_i} e^{-\mu}}{x_i!} \prod_{i=k+1}^{m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} (\mu^{a_1-1} e^{-\mu b_1})(\lambda^{a_2-1} e^{-\lambda b_2}) \left(\frac{1}{m}\right)$$

- Or equivalently,

$$\pi(\mu, \lambda, k|\underline{X}) \propto \mu^{a_1+\sum_{i=1}^{k} x_i-1} e^{-\mu(k+b_1)} \lambda^{a_2+\sum_{i=k+1}^{m} x_i-1} e^{-\lambda(m-k+b_2)}$$

- Full conditional distributions,
  1. $\pi(\mu|\lambda, k, \underline{X}) \propto \mu^{a_1+\sum_{i=1}^{k} x_i-1} e^{-\mu(k+b_1)} =$ $Ga(a_1 + \sum_{i=1}^{k} x_i, k + b_1)$
  2. $\pi(\lambda|\mu, k, \underline{X}) \propto \lambda^{a_2+\sum_{i=k+1}^{m} x_i-1} e^{-\lambda(m-k+b_2)} =$ $Ga(a_2 + \sum_{i=k+1}^{m} x_i, m - k + b_2)$
  3. $\pi(k|\mu, \lambda, \underline{X}) \propto \mu^{\sum_{i=1}^{k} x_i} \lambda^{\sum_{i=k+1}^{m} x_i} e^{-k(\mu-\lambda)} \equiv L(\underline{X}|\mu, \lambda, k)$
     - So, $p(k|\mu, \lambda, \underline{X}) = \frac{L(\underline{X}|\mu,\lambda,k)}{\sum_{k=1}^{m} L(\underline{X}|\mu,\lambda,k)}, k = 1, 2, \ldots, m$
- Some results with the British Coal mining disaster data follow.


theta
iteration
Series draws[, i]
ACF
Lag
Histogram of draws[, i]
Density
theta
lambda
k

Metropolis-Hastings Algorithm (continuous case)

- Sample (indirectly) a value X from a pdf $f(x)$.
- In Bayesian problems $f(x)$ is the posterior distribution.
- Need a *trial* or *proposal* density $q(x, y) \geq 0$.
  1. Set an initial value $X^{(0)}$. For $n = 1, 2, \ldots,$
  2. Given $X^{(n)} = x$, generate $Y = y$, from $q(x, y)$.
  3. Compute,
$$\alpha = \min \left\{ 1, \frac{f(y)q(x, y)}{f(x)q(x, y)} \right\}$$
  4. $U \sim U(0, 1)$. If $U \leq \alpha$, then $X^{(n+1)} = y$. Otherwise, $X^{(n+1)} = x$
  5. Change n to $n + 1$ and go back to 2.

- Common to run algorithm say 1000 iterations and delete the first, say 100 (burn-in).
- Under a symmetric proposal (Metropolis), $q(x, y) = q(y, x)$ and

$$\alpha = \min\left\{1, \frac{f(y)}{f(x)}\right\}$$

- *Random walk* proposal,

$$y = x + z; z \sim N(0, h^2).$$

- h is a scaling parameter.
- h is specified so that $40\% - 60\%$ *acceptance rate* is achieved.

Example with Bivariate Normal Distribution

- $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and S is a 2×2 covariance matrix with diagonal entries $s_{1,1}$, $s_{2,2}$ and off-diagonals $s_{1,2} = s_{2,1}$. The pdf is

$$\pi(\boldsymbol{\theta}) \propto |S|^{-1/2} exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^t S^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- Marginal distributions: $\theta_1 \sim N(\mu_1, s_{1,1})$ and $\theta_2 \sim N(\mu_2, s_{2,2})$.
- Example with $\boldsymbol{\mu} = (2, 1)$, $s_{1,1} = s_{2,2} = 1$ and $s_{1,2} = 0.7$.

- *Trial* or *candidate* density $q(\theta, \theta')$, also a bivariate normal so

$$\theta' = \theta + \boldsymbol{z}; \boldsymbol{z} \sim N(\boldsymbol{0}, \Sigma)$$

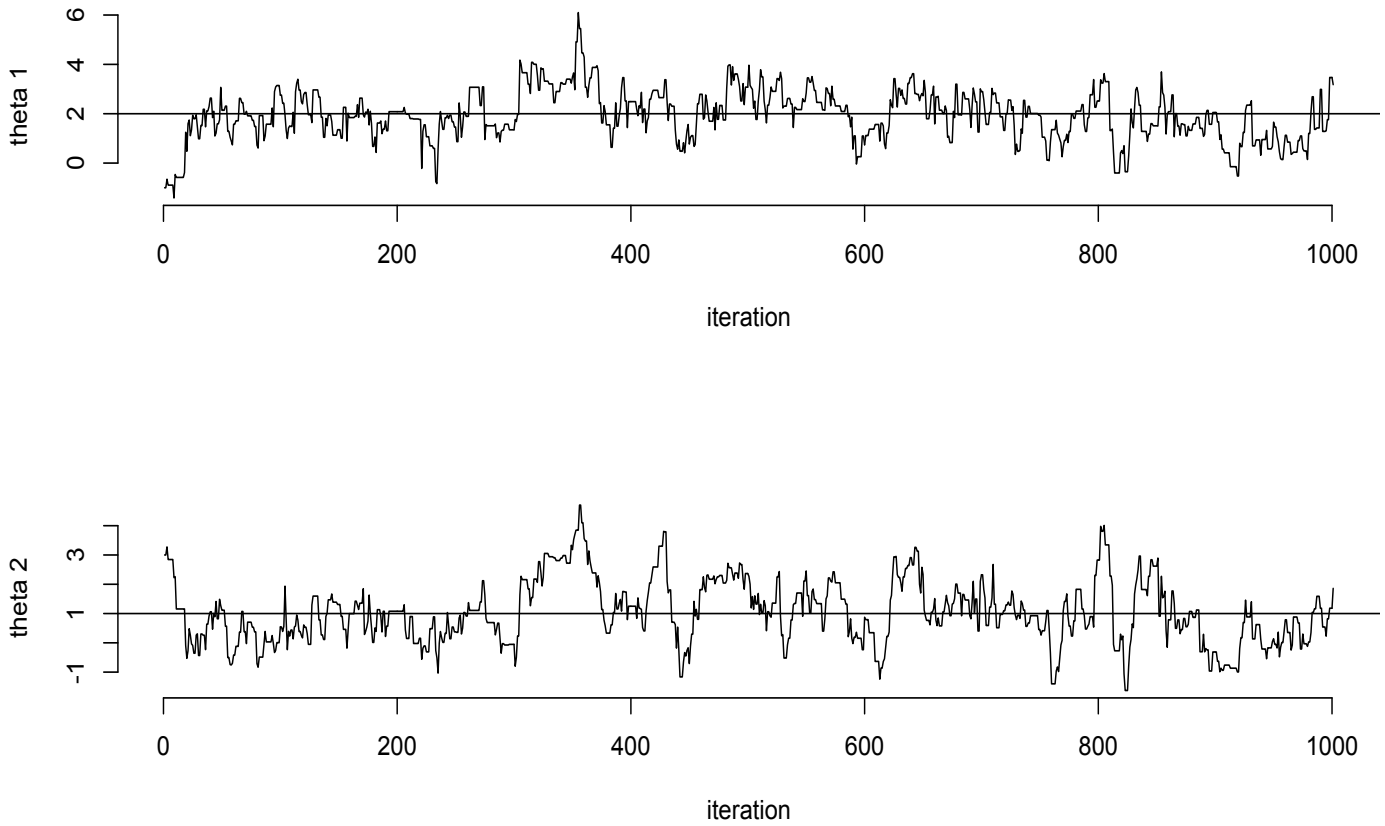
- with Σ a diagonal matrix with diagonal entries 0.6, 0.4.
- $z_1 \sim N(0, 0.6)$ and $z_2 \sim N(0, 0.4)$.
- Random walk proposal so $q(\theta, \theta') = q(\theta', \theta)$.
- The Metropolis ratio is:

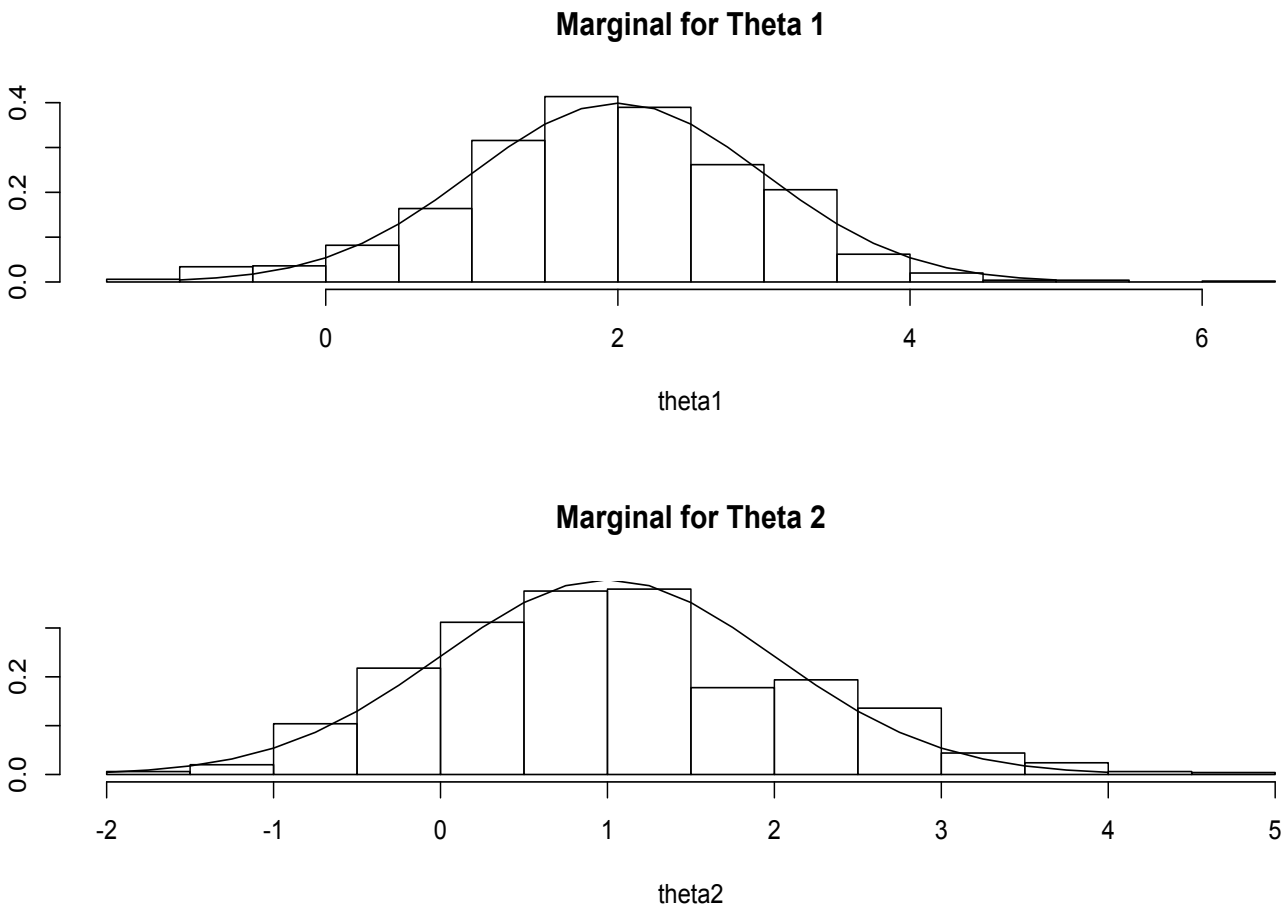
$$\begin{aligned} \alpha(\theta, \theta') &= \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\} = \\ &= \min\left\{1, \frac{exp\left(-\frac{1}{2}(\theta' - \mu)^t S^{-1}(\theta' - \mu)\right)}{exp\left(-\frac{1}{2}(\theta - \mu)^t S^{-1}(\theta - \mu)\right)}\right\} \end{aligned}$$

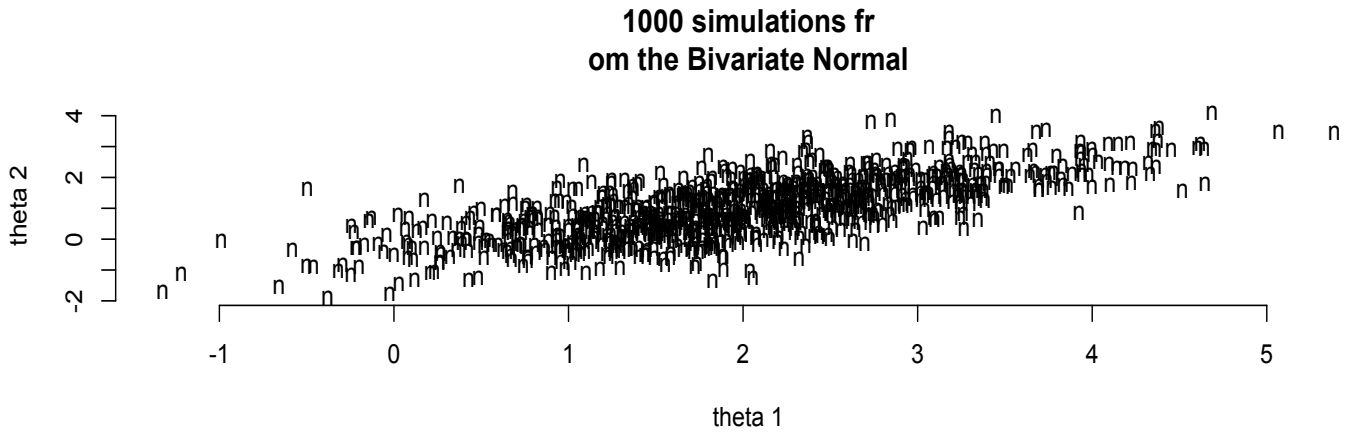
- From an initial point $\boldsymbol{\theta}^{(0)}$. For $i = 1, 2, \ldots,$

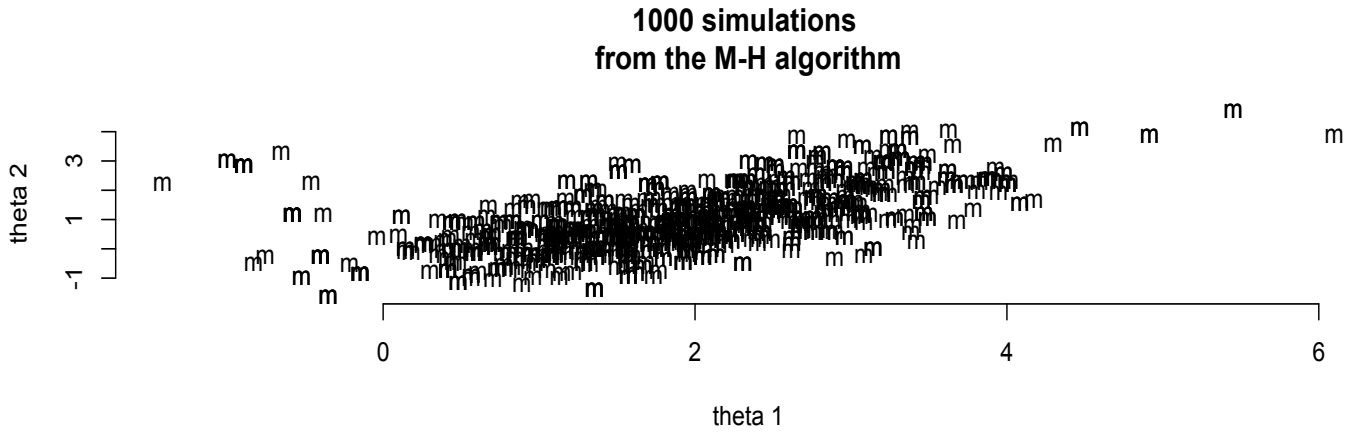
$$\boldsymbol{\theta}' = \boldsymbol{\theta}^{(i-1)} + \boldsymbol{z}$$

- Then compute the acceptance probability, $\alpha(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}')$.
- Make $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}'$ if $U < \alpha(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}')$.
- Otherwise, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$.
- From $\theta^{(0)} = (-1, 3)$, we'll look at:
  - Time series plots of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ values.
  - Histograms of *marginal* samples for θ_1 and θ_2.
  - Bivariate plots of samples.


theta 1
0
2
4
6
0
200
400
600
800
1000
iteration
theta 2
-1
1
3
0
200
400
600
800
1000
iteration


Marginal for Theta 1
0.0
0.2
0.4
0
2
4
6
theta1
Marginal for Theta 2
0.0
0.2
-2
-1
0
1
2
3
4
5
theta2


1000 simulations fr
om the Bivariate Normal
theta 2
-2
0
2
4
-1
0
1
2
3
4
5
theta 1


1000 simulations
from the M-H algorithm
theta 2
-1
1
3
0
2
4
6
theta 1