David Giles

Bayesian Econometrics

5. Bayesian Computation

- Historically, the computational "cost" of Bayesian methods greatly limited their application.

- For instance, by Bayes' Theorem:

$$p(\boldsymbol{\theta} \,|\boldsymbol{y}) = p(\boldsymbol{\theta})p(\boldsymbol{y} \,|\boldsymbol{\theta})/p(\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y} \,|\boldsymbol{\theta})$$

- The proportionality constant is

$$p(\boldsymbol{y}) = \iiint_{-\infty}^{\infty} p(\boldsymbol{\theta})p(\boldsymbol{y} \,|\boldsymbol{\theta}) d\theta_1 \ldots d\theta_k$$

- Unless this integration can be performed analytically, it will have to be done numerically, or an approximation will have to be used.
- Natural-Conjugate priors are not always available, and not always appropriate.
- If $k > 3$ (or so) conventional numerical "quadrature" (*e.g.*, extensions of Simpson's rule), will be infeasible in terms of computational time.
- Same issue arises if we want to obtain $\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta} \mid \boldsymbol{y}]$, or if we want to marginalize the joint posterior p.d.f.:

$$p(\boldsymbol{\theta}_1 | \boldsymbol{y}) = \iiint_{-\infty}^{\infty} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \boldsymbol{y}) d\boldsymbol{\theta}_2$$

- Starting in the late 1970’s / early 1980's, several methods for dealing with this issue were considered.
- These involved approximating the required integrals.

  (i) Laplace integration (*analytic*)

  (ii) Monte Carlo *integration* ("importance sampling") (*simulation*)
- More recently, the big breakthroughs have come by not actually attempting to evaluate the integrals at all!
- Essentially simulate the densities that we're interested in - *e.g.*, a marginal posterior density.

- The family of methods that we'll explore is called Markov Chain Monte Carlo (MCMC; or $(MC)^2$) .
- We won't go into the mathematics of Markov Chains in any detail.
- Main group of MCMC methods we'll be concerned with is the so-called Metropolis-Hastings methodology.
- A special case of M-H is the so-called Gibbs Sampler.
- We'll start with the latter - it's easier to deal with.
- It can be applied to Bayesian problems of high dimension.
- However, may require some ingenuity, and may not be the most efficient method to use.

The Gibbs Sampler

- Why the name? Who was Gibbs?

Josiah Willard Gibbs (1839 – 1903)

Co-creator of statistical mechanics; creator of vector calculus; …..

- Name used by Geman & Geman, 1984: "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images".
- Let's illustrate the main steps for the Gibbs sampler.
- Remember, we want to obtain the marginal posterior densities for some parameters of interest.
- Once we have theses p.d.f.'s it will turn out to be a simple matter to use them to construct Bayes estimators, and BCI's, *etc*.
- Applying Bayes' Theorem, we have the *kernel* for the joint posterior p.d.f. for all of the parameters:

$$p(\boldsymbol{\theta} \,|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y} \,|\boldsymbol{\theta})$$

- For simplicity, suppose that k = 2. (In practice, k can be several thousands.)
- So, $p(\theta_1, \theta_2 \,|\boldsymbol{y}) \propto p(\theta_1, \theta_2)p(\boldsymbol{y}\,|\theta_1, \theta_2)$.
- Suppose that the two *conditional posterior densities*, $p(\theta_1|\theta_2, \boldsymbol{y})$ and $p(\theta_2|\theta_1, \boldsymbol{y})$ are of some (generally different) recognizable forms.
- (Actually, the requirements are even weaker than this, as we'll see.)
- Then we can take a random drawing from each of $p(\theta_1|\theta_2, \boldsymbol{y})$ and $p(\theta_2|\theta_1, \boldsymbol{y})$.
- The Gibbs Sampler then proceeds as follows:

(i) $\theta_1^{(1)} \leftarrow p(\theta_1 \,|\theta_2^{(0)}, \boldsymbol{y})$

(ii) $\theta_2^{(1)} \leftarrow p(\theta_2 \,|\theta_1^{(1)}, \boldsymbol{y})$

(iii) $\theta_1^{(2)} \leftarrow p(\theta_1 \,|\theta_2^{(1)}, \boldsymbol{y})$

(iv) $\theta_2^{(2)} \leftarrow p(\theta_2 \,|\theta_1^{(2)}, \boldsymbol{y})$

etc.............

- So, this gives us a string of thousands of drawings from the two conditional posterior p.d.f.'s for the 2 parameters.
- Continuing this process long enough, eventually the drawings will actually come from the *marginal posterior p.d.f.'s* for the parameters!

- We can then continue to keep drawing values from each distribution and we'll end up with thousands of simulated values.
- We'll need to discard lots of early values obtained by this process, as they'll actually be from the *conditional posterior* p.d.f.'s, and not from the *marginal posterior* p.d.f.'s
- This is referred to as the "Burn in".
- Various tools available to help us decide the length of the Burn in.
- Gibbs sampler lends itself to parallel processing - run many strings independently on different processors and then combine results.
- Exactly the same approach applies when we have more parameters.

- For instance, suppose that $k = 4$:

(i)	$\theta_1^{(1)}$	$\leftarrow$	$p(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \boldsymbol{y})$
(ii)	$\theta_2^{(1)}$	$\leftarrow$	$p(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \theta_4^{(0)}, \boldsymbol{y},)$
(iii)	$\theta_3^{(1)}$	$\leftarrow$	$p(\theta_3 \mid \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \boldsymbol{y})$
(iv)	$\theta_4^{(1)}$	$\leftarrow$	$p(\theta_4 \mid \theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(1)}, \boldsymbol{y})$
(v)	$\theta_1^{(2)}$	$\leftarrow$	$p(\theta_1 \mid \theta_2^{(1)}, \theta_3^{(1)}, \theta_4^{(1)}, \boldsymbol{y})$
(vi)	$\theta_2^{(2)}$	$\leftarrow$	$p(\theta_2 \mid \theta_1^{(2)}, \theta_3^{(1)}, \theta_4^{(1)}, \boldsymbol{y})$
(vii)	$\theta_3^{(2)}$	$\leftarrow$	$p(\theta_3 \mid \theta_1^{(2)}, \theta_2^{(2)}, \theta_4^{(1)}, \boldsymbol{y})$
(viii)	$\theta_4^{(2)}$	$\leftarrow$	$p(\theta_4 \mid \theta_1^{(2)}, \theta_2^{(2)}, \theta_3^{(2)}, \boldsymbol{y})$ *etc*.............

Example 1:

- Let's see if this works, by considering a situation where <u>we know the answer</u>.
- Note - *this won't be a Bayesian example*. The purpose is just to see how the Gibbs sampler moves from the conditional densities to the marginal densities.
- Suppose we have a random vector, $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma\right]$, where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

- It's easy to show that:

(i) $p(y_1 \mid y_2) \sim N\left[\left\{\mu_1 + \left(\frac{\rho\sigma_1}{\sigma_2}\right)(y_2 - \mu_2)\right\} \;,\; \sigma_1^2(1-\rho^2)\right]$

(ii) $p(y_2 \mid y_1) \sim N\left[\left\{\mu_2 + \left(\frac{\rho\sigma_2}{\sigma_1}\right)(y_1 - \mu_1)\right\} \;,\; \sigma_2^2(1-\rho^2)\right]$

(iii) $p(y_1) \sim N[\mu_1 \,, \sigma_1^2]$

(iv) $p(y_2) \sim N[\mu_2 \,, \sigma_2^2]$

- We'll consider the case where $\mu_1 = \mu_2 = 0$; $\sigma_1 = \sigma_2 = 1$.
- The Gibbs sampler will involve the following steps:

(i) Sample y_1 from $p(y_1 \mid y_2)$; (ii) Sample y_2 from $p(y_2 \mid y_1)$

(iii) Keep repeating steps (i) and (ii), lots of times.

- Eventually, $p(y_1 | y_2) \rightarrow p(y_1)$, and $p(y_2 | y_1) \rightarrow p(y_2)$,
- We'll then continue until we have a large sample of drawings from these two marginal p.d.f.'s.
- This will give us empirical p.d.f.'s of the form that we want, *without doing any integration of any sort*!
- We'll have to assign initial values, and decide on the length of the "Burn in".
- Recall - in this illustration we actually *know* what the marginal p.d.f.'s look like, so we'll know if the Gibbs sampler is really working.
- If you're convinced, then we can move to some real Bayesian examples.

- Code to do this using R:

```
library(tseries)


set.seed(123)
nrep<- 100000              # Total number of MC replications
nb<- 2000                  # Number of observations for the "Burn-in"
yy1<- array(,nrep)
yy2<- array(,nrep)


rho<- 0.5                  # Set the correlation between Y1 and Y2
sd<- sqrt(1-rho^2)
y2<- rnorm(1,0,sd)         # Initialize Y2
```

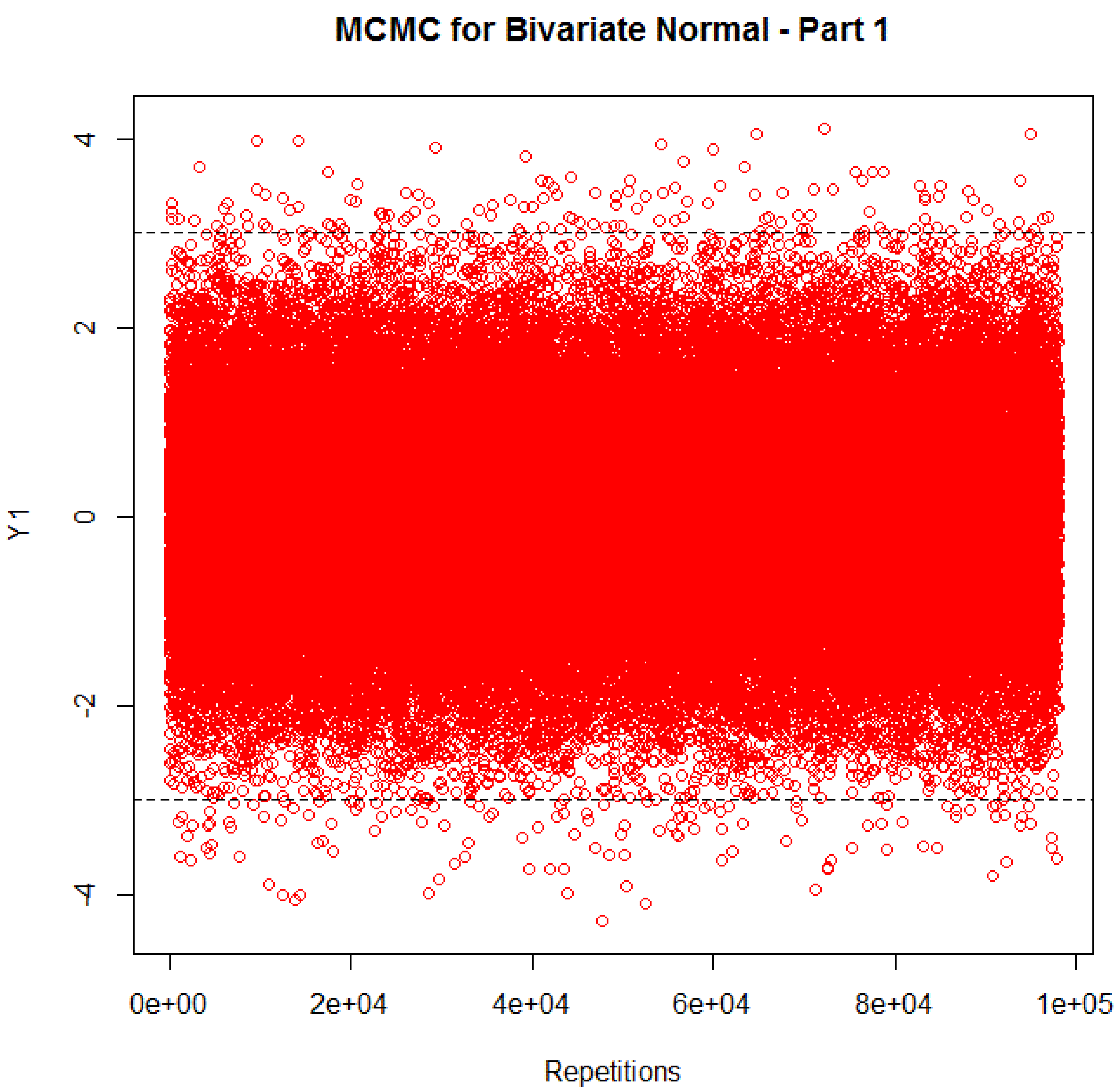
THE GIBBS SAMPLER

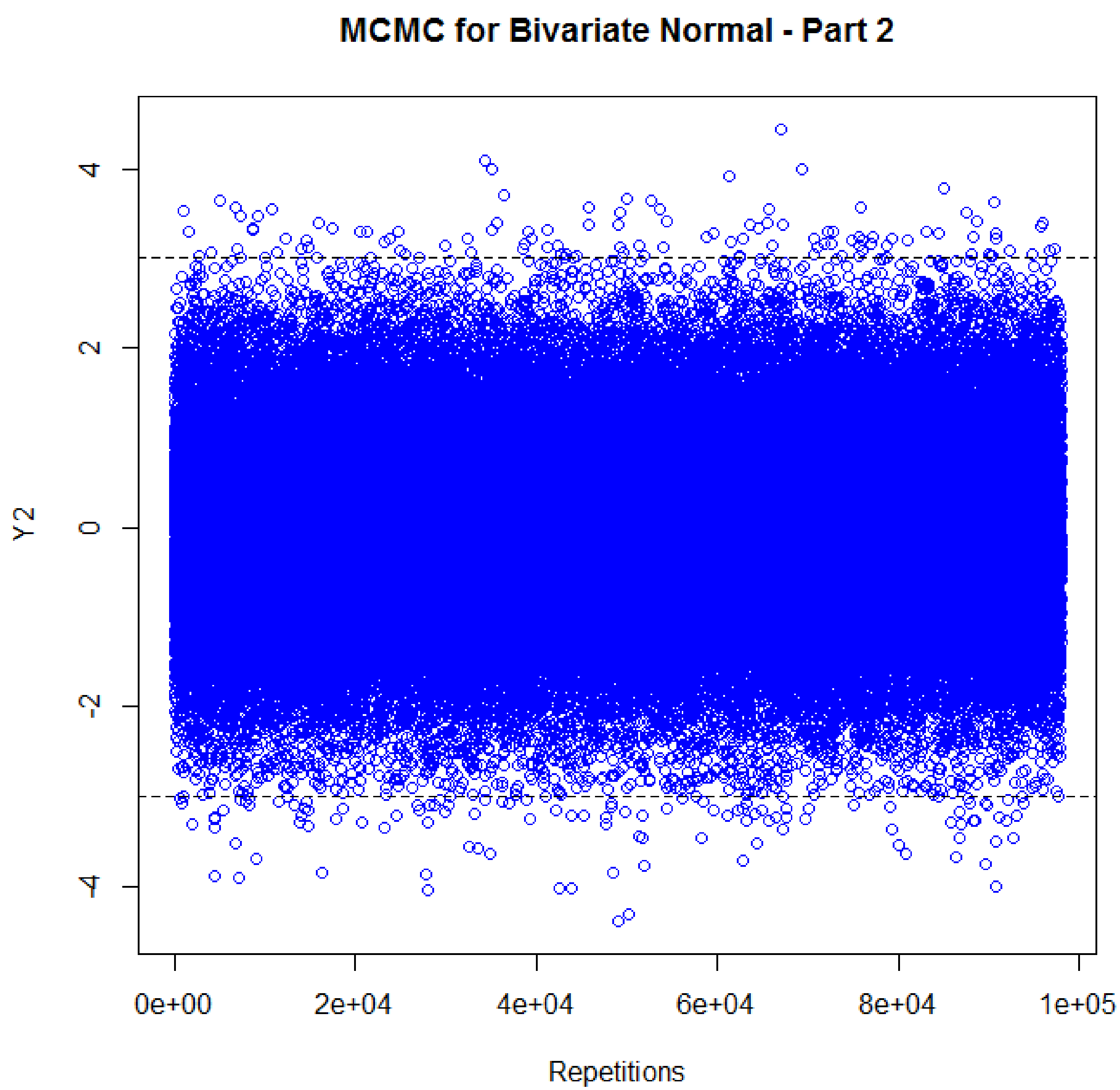
```
for (i in 1:nrep) {
y1<-  rnorm(1,0,sd)+rho*y2
y2<-  rnorm(1,0,sd)+rho*y1
yy1[i]<- y1
yy2[i]<- y2
}

            # Drop the first "nb" repetitions for the "Burn-in"
nb1<- nb+1
yy1b<-yy1[nb1:nrep]
yy2b<- yy2[nb1:nrep]
```

```
# Plot the "Trace" results for the 2 p.d.f.'s

plot(yy1b, col=2, main="MCMC for Bivariate Normal - Part 1", xlab="Repetitions",
ylab="Y1")

abline(h=3,lty=2)

abline(h=-3,lty=2)

plot(yy2b, col=4, main="MCMC for Bivariate Normal - Part 2", xlab="Repetitions",
ylab="Y2")

abline(h=3,lty=2)

abline(h=-3,lty=2)

# Determine the moments of the Marginal Posterior p.d.f.'s

summary(yy1b)

var(yy1b)

summary(yy2b)

var(yy2b)
```

Plot the histograms for the 2 marginal posterior p.d.f.'s

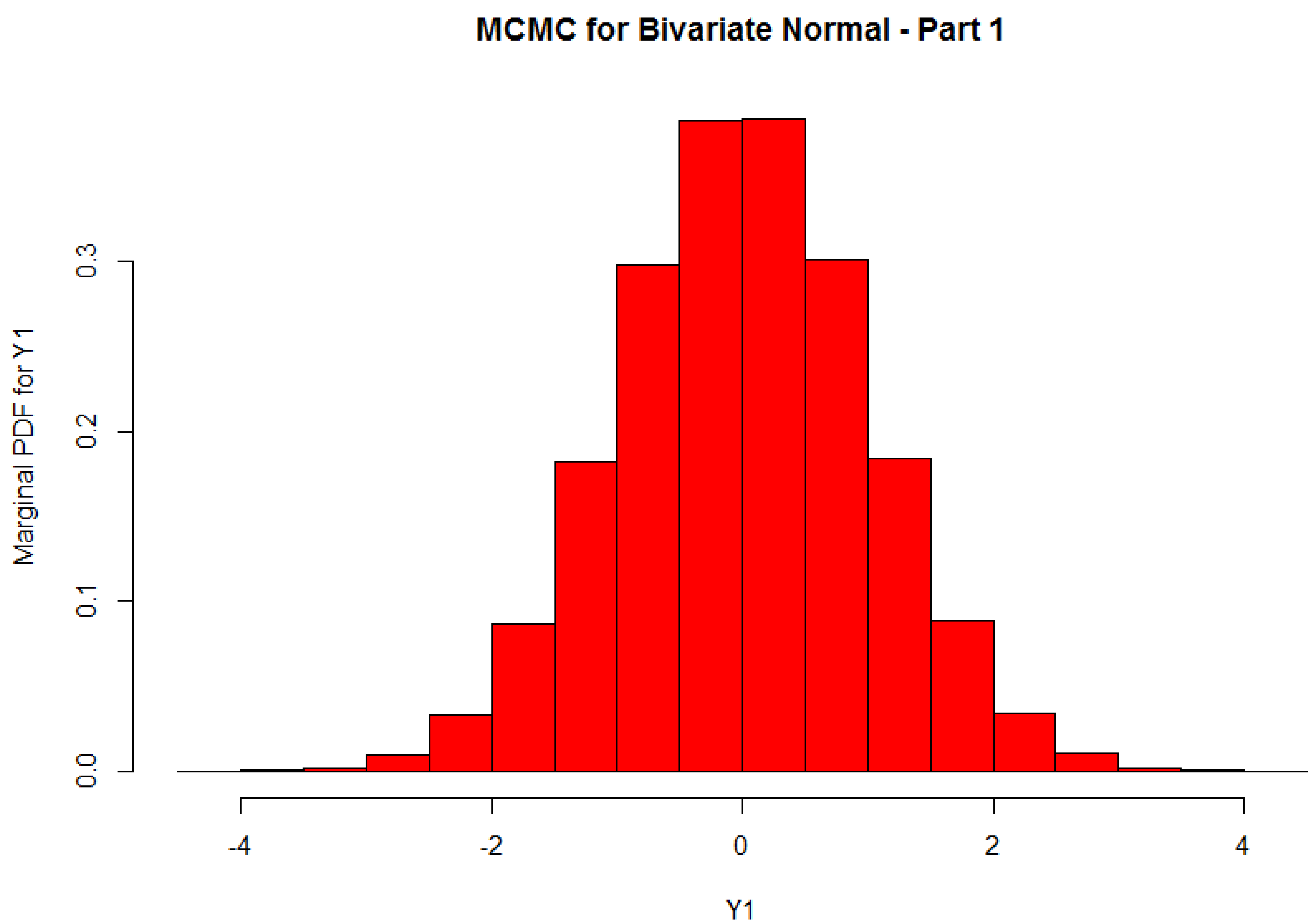
```
hist(yy1b, prob=T,col=2, main="MCMC for Bivariate Normal - Part 1", xlab="Y1",
ylab="Marginal PDF for Y1")

hist (yy2b,prob=T,col=4, main="MCMC for Bivariate Normal - Part 2", xlab="Y2",
ylab="Marginal PDF for Y2")

                # Check for Normality of the marginal posteriors

qqnorm(yy1b)                    # Q-Q Plots

qqline(yy1b,col=2)

qqnorm(yy2b)

qqline(yy2b,col=4)

jarque.bera.test(yy1b)  ;   jarque.bera.test(yy2b)
```

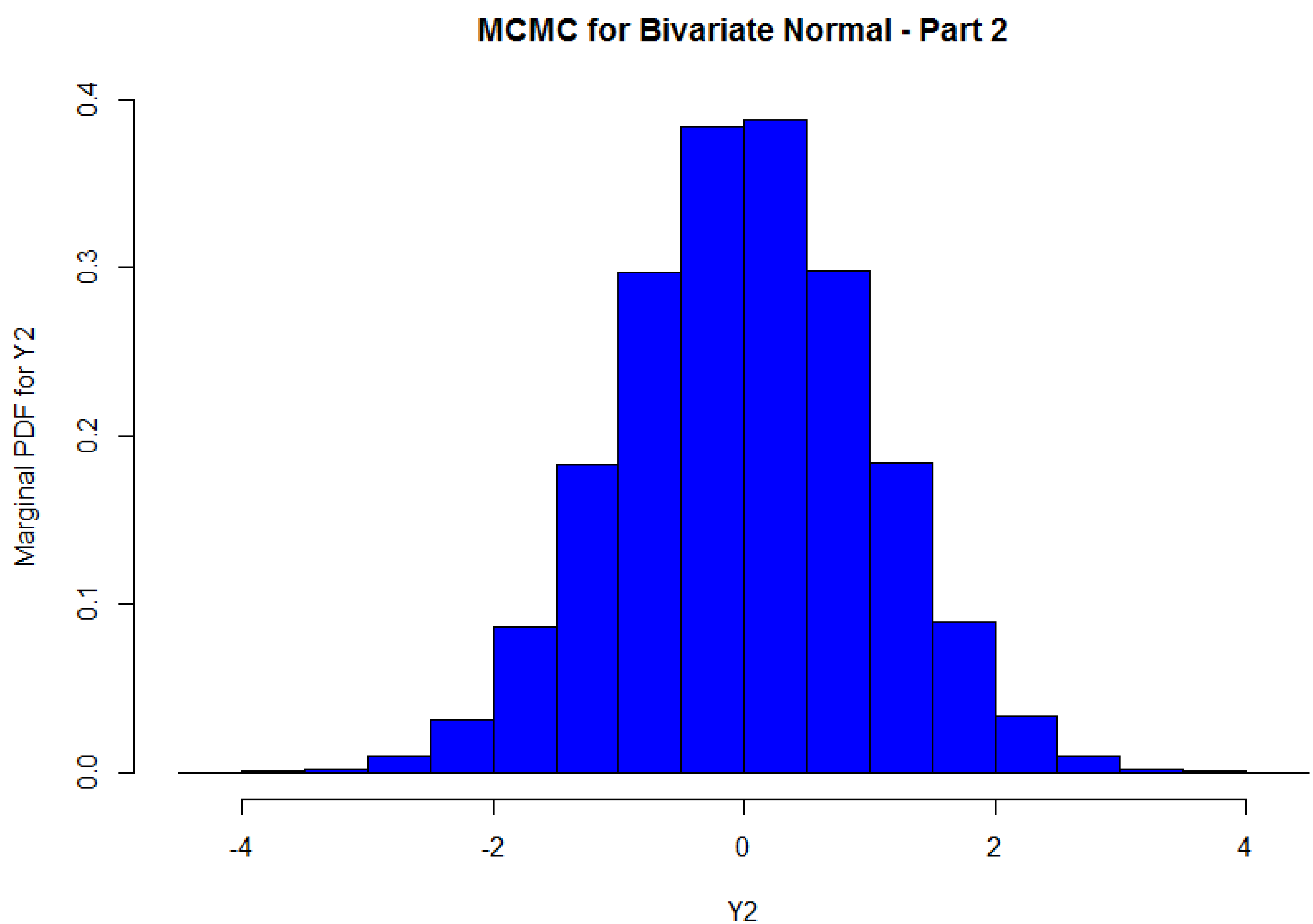

MCMC for Bivariate Normal - Part 1
Y1
4
2
0
-2
-4
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Repetitions

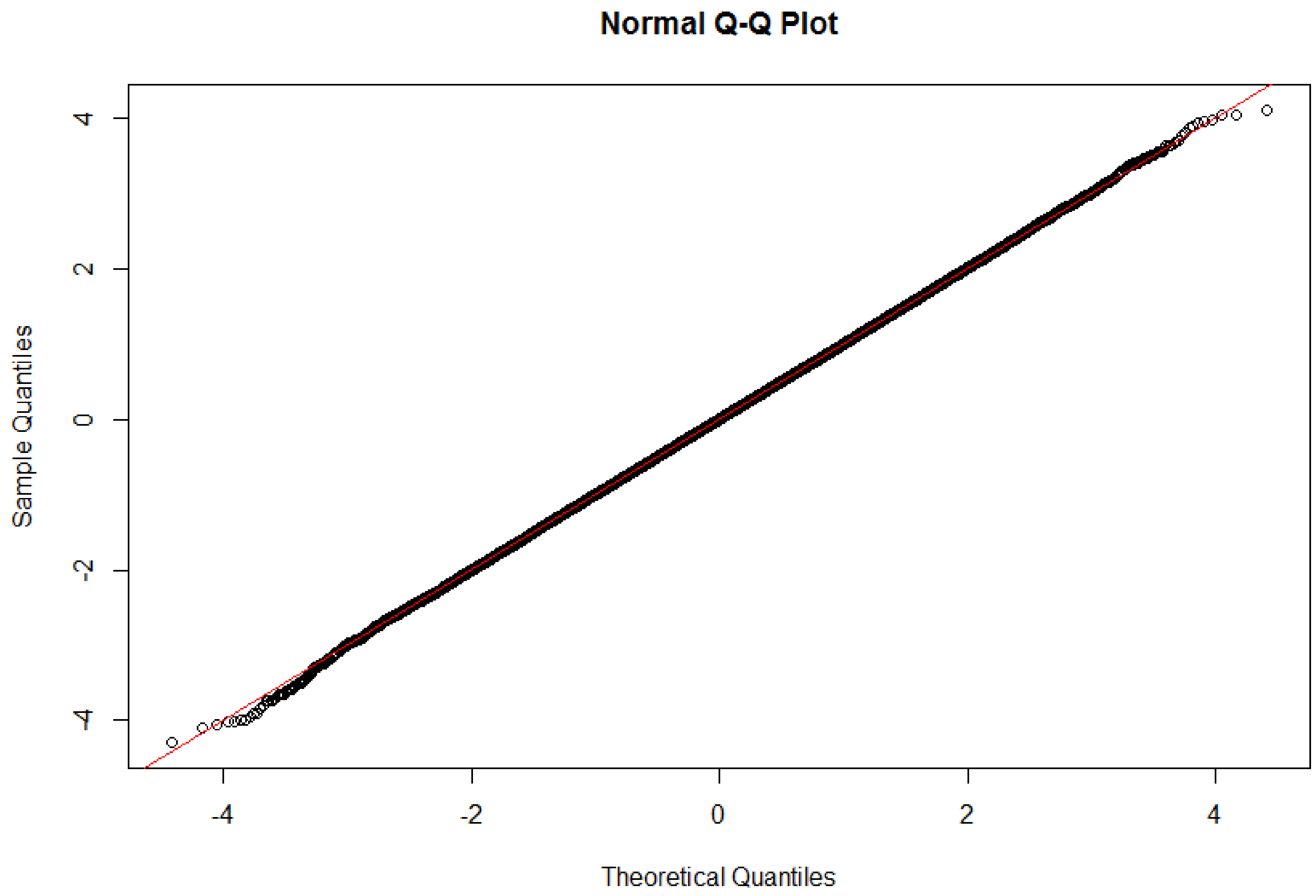

MCMC for Bivariate Normal - Part 2
Y2
4
2
0
-2
-4
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Repetitions

```
> summary(yy1b)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.285000 -0.668500  0.006683  0.006148  0.679600  4.111000
> var(yy1b)
[1] 1.003738
> summary(yy2b)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.396000 -0.671000  0.006226  0.004281  0.675800  4.448000
> var(yy2b)
[1] 0.9953644
```

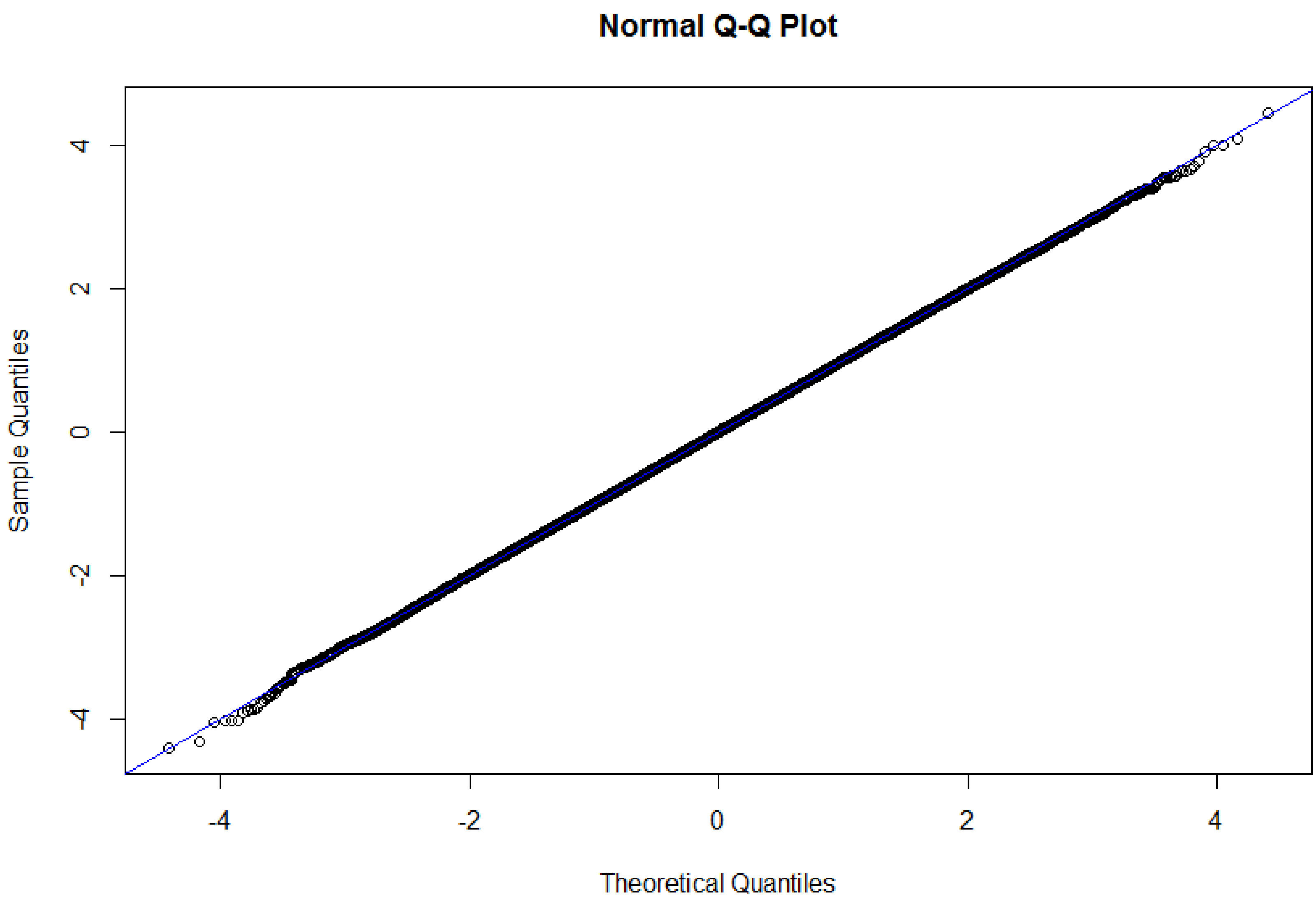
```
> jarque.bera.test(yy1b)  ;   jarque.bera.test(yy2b)

        Jarque Bera Test

data:  yy1b
X-squared = 1.4732, df = 2, p-value = 0.4787


        Jarque Bera Test

data:  yy2b
X-squared = 0.7989, df = 2, p-value = 0.6707
```


MCMC for Bivariate Normal - Part 1
Marginal PDF for Y1
0.0
0.1
0.2
0.3
-4
-2
0
2
4
Y1


MCMC for Bivariate Normal - Part 2
Marginal PDF for Y2
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Y2



Q-Q Plot for $p(y_1)$



Q-Q Plot for $p(y_2)$

Example 2:

- Let's consider another example where we know the answer.
- However, *this one is a Bayesian example*.
- We want to estimate the 2 unknown parameters of a Normal population - the mean, μ, and the precision, τ (= 1 / σ^2).
- Diffuse (Jeffrey's) prior p.d.f.: $p(\mu, \tau) = p(\mu)\, p(\tau) \propto 1/\tau$
- Likelihood function:

$$p(\boldsymbol{y}\,|\mu, \tau) \propto \tau^{n/2} exp\left\{-\tau/2 \sum_{i=1}^{n} (y_i - \mu)^2\right\}$$

- Bayes' Theorem:

$$p(\mu, \tau \mid \boldsymbol{y}) \propto \tau^{\frac{n}{2}-1} exp\left\{-(\frac{\tau}{2}) \sum_{i=1}^{n} (y_i - \mu)^2\right\}$$

- Consider the *conditional posterior* densities.

- $p(\mu \mid \tau, \boldsymbol{y}) \propto exp\left\{-(\frac{\tau}{2}) \sum_{i=1}^{n} (y_i - \mu)^2\right\}$

$$\propto exp\left\{-(\frac{\tau}{2})[vs^2 + n(\bar{y} - \mu)^2]\right\}$$

$$\propto exp\left\{-(\frac{n\tau}{2})(\mu - \bar{y})^2\right\}$$

- This is the kernel of a $N[\bar{y}\,, (n\tau)^{-1}]$ density.

- Similarly, we can get the conditional posterior for τ :

$$p(\tau \mid \mu, \boldsymbol{y}) \propto \tau^{\frac{n}{2}-1} \exp\left\{-\tau\left((\frac{1}{2})\sum_{i=1}^{n}(y_i-\mu)^2\right)\right\}$$

- This is the kernel of a Gamma density, $\Gamma(r, \lambda)$, with shape & scale

  parameters, $r = n/2$; $\lambda = \left[(\frac{1}{2})\sum_{i=1}^{n}(y_i-\mu)^2\right]^{-1}$.

- Now, in fact we know that for this problem, the *marginal posterior* for μ is

  Student –t, centered at $\bar{y}$; and the *marginal posterior* for τ is Gamma.

- Suppose that we don't know this, and we decide to use the Gibbs sampler.

- Let's see what we get, with n = 10.

- Here is the R code:

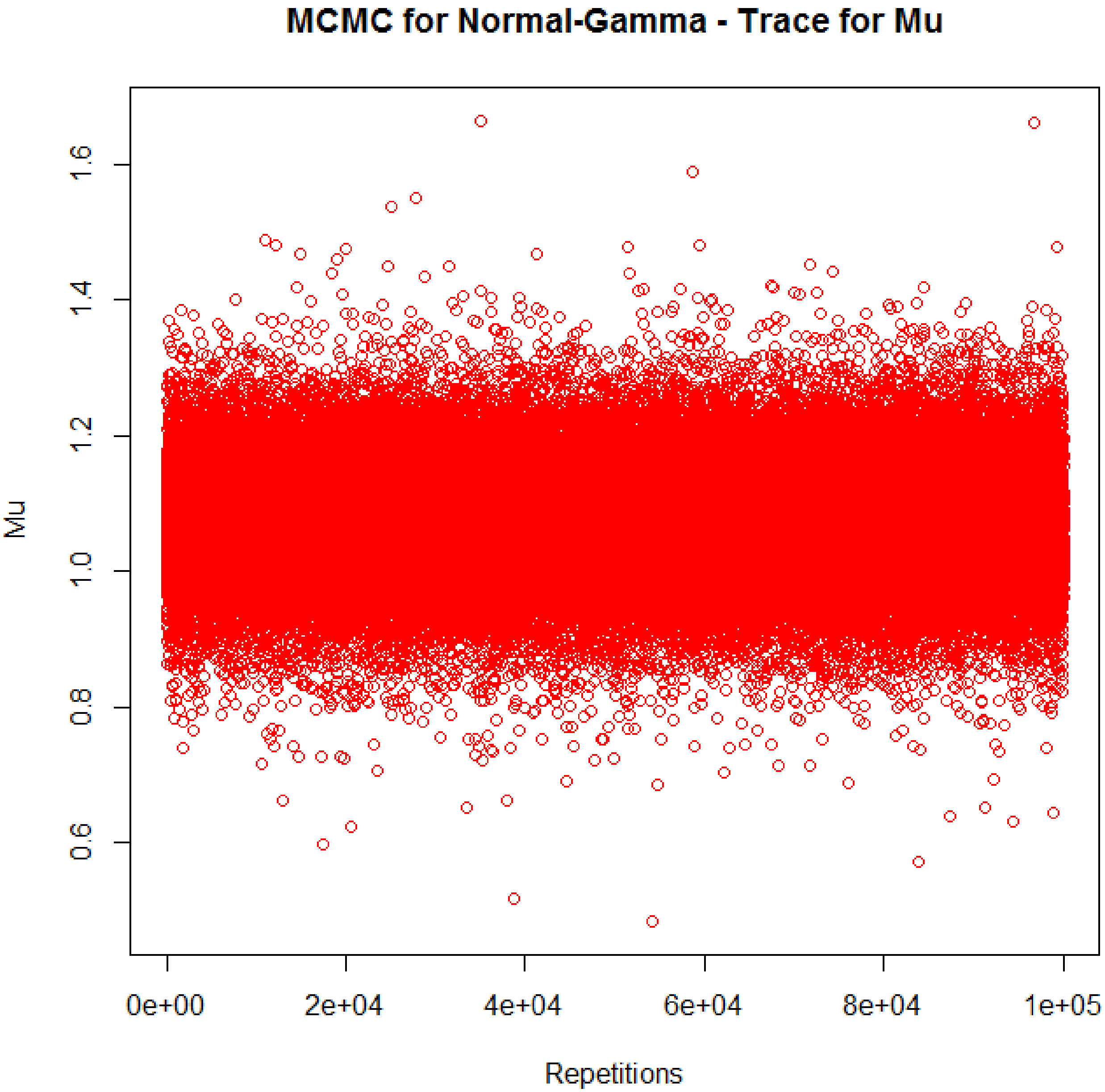
```
library(moments)

set.seed(123)

nrep<- 105000                          # Total number of MC replications

nb<- 5000                              # Number of observations for the "Burn-in"

n<- 10                                 # Sample size

tau<- array(,nrep)                     # Set up vectors for storing results

mu<- array(,nrep)


y<- rnorm(n,mean=1,sd=1)               #  Create a sample of data:   N[1,1]

                 # True values of Mu and Tau are each 1

ybar<- mean(y)

yy<- sum(y^2)

lambda<- 1/(0.5*n*var(y))
```

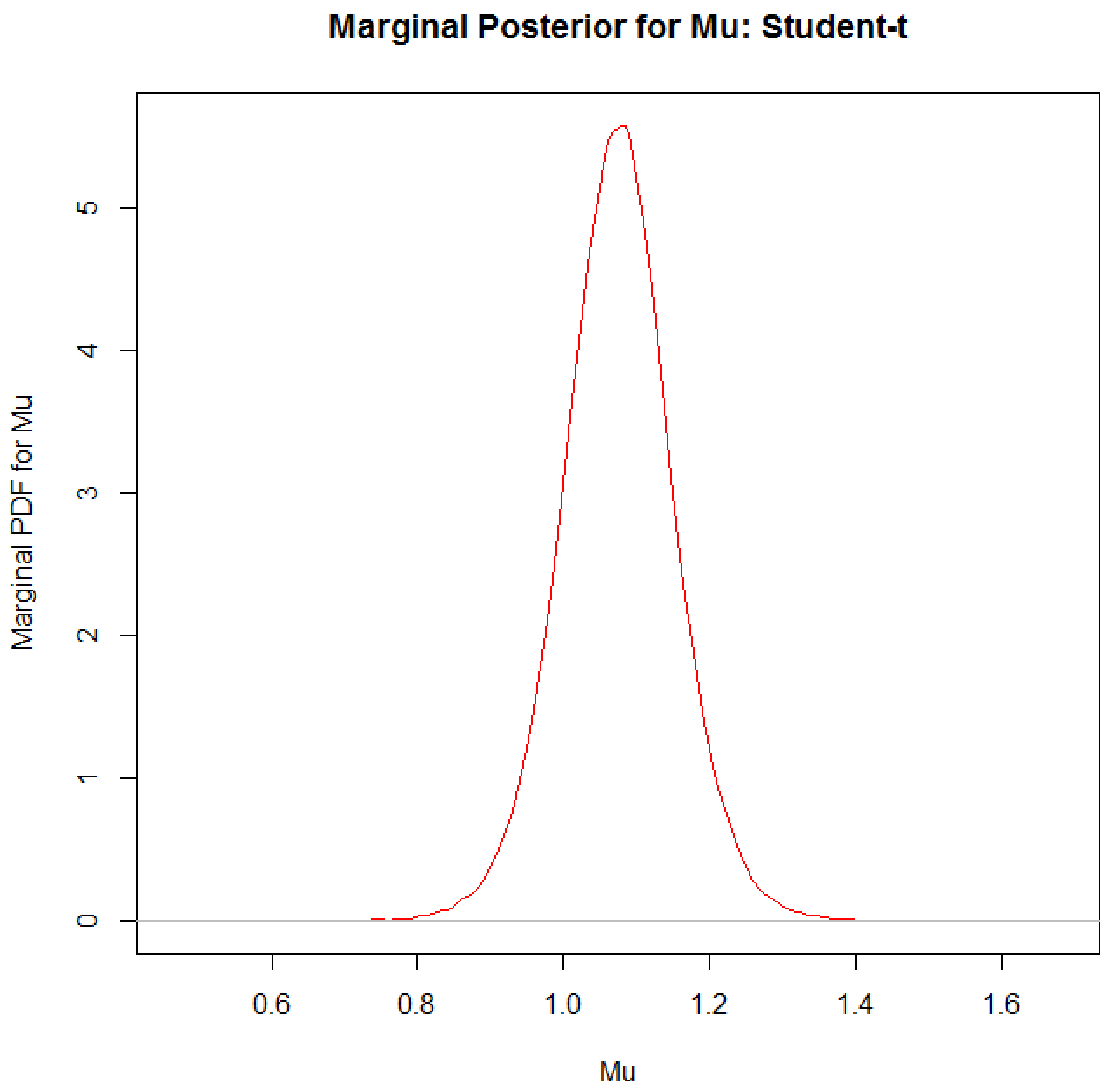
```
ttau<- rgamma(1, shape = n/2, scale = lambda)      #initialize Tau

                    #START OF THE MCMC LOOP:

for (i in 1:nrep) {

mmu<-rnorm(1,mean = ybar,sd = 1/sqrt(n*ttau))

scal<- 1 / (0.5*(yy+n*mmu^2-2*n*mmu*ybar))

ttau<- rgamma(1, shape=n/2, scale=scal)

tau[i]<- ttau

mu[i]<- mmu

}

                    # END OF THE MCMC LOOP
```

```
# Drop the first "nb" repetitions for the "Burn-in"

# We have 100,000 values for the marginal posteriors

# Let's see if the results seem to be accurate:


nb1<-nb+1

taub<-tau[nb1:nrep]

mub<- mu[nb1:nrep]


# Plot the traces for the marginal p.d.f.'s

plot(mub, col=2, main="MCMC for Normal-Gamma - Trace for Mu", xlab="Repetitions", ylab="Mu")

plot(taub, col=4, main="MCMC for Normal-Gamma - Trace for Tau", xlab="Repetitions", ylab="Tau")
```

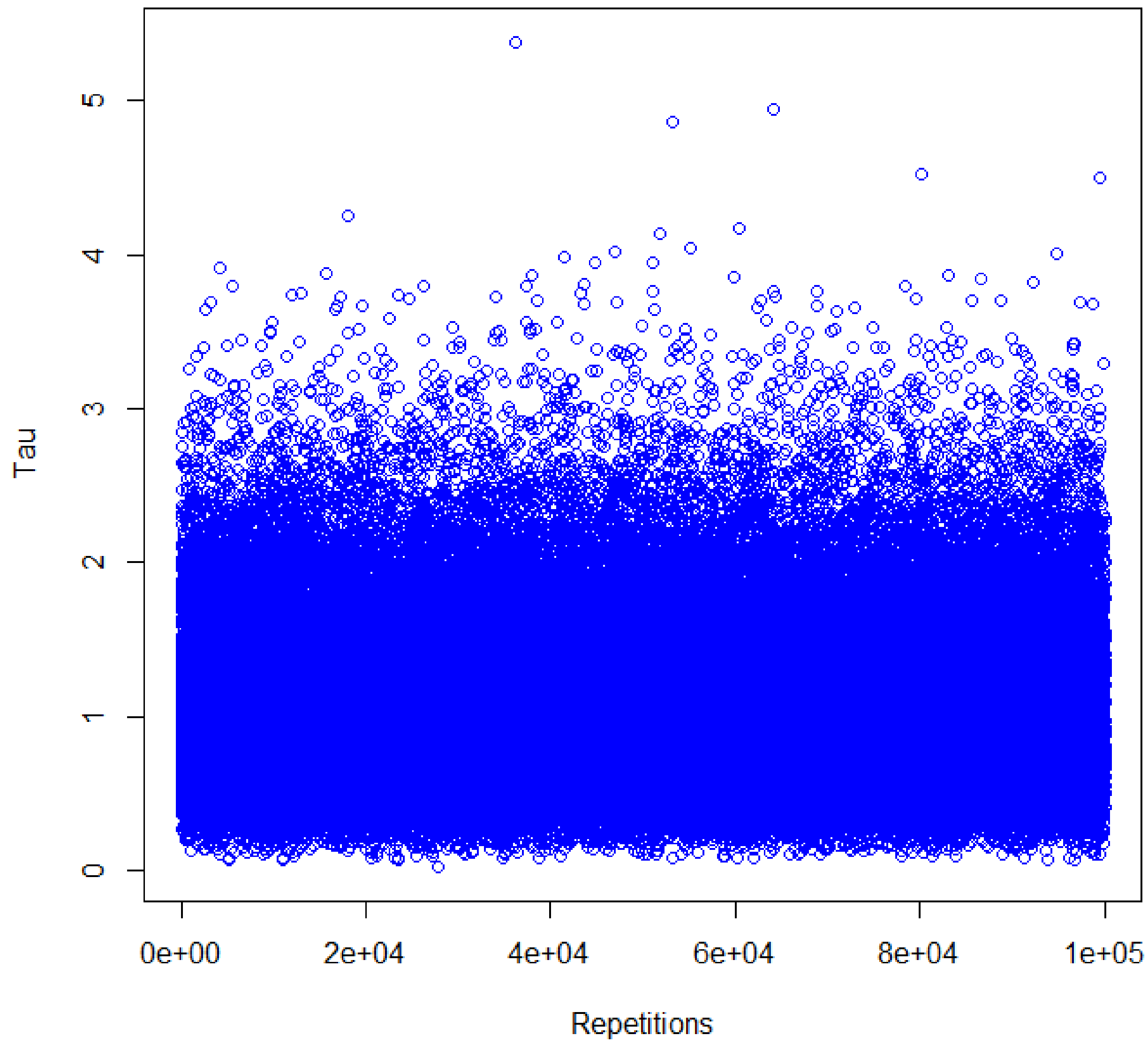
```
# The marginal posteriors for Mu and Tau should be Student-t (n-1), and Gamma,
respectively

summary(mub)  ; var(mub)
ybar       # The mean of the marginal posterior for Mu should be ybar ( = 1.0746)

skewness(mub)        #  the skewness of Student-t is zero
kurtosis(mub)
                  # The EXCESS kurtosis for Student-t (n-1) is 6/(n-5)=1.2; so kurtosis = 4.2
summary(taub)  ; var(taub)
skewness(taub)      #  the skewness of Gamma is (2/sqrt(shape)) =  (2/sqrt(n/2) = 0.8944
kurtosis(taub)      # excess kurtosis for Gamma is  (6/shape) = 6/(n/2) = 1.2
```

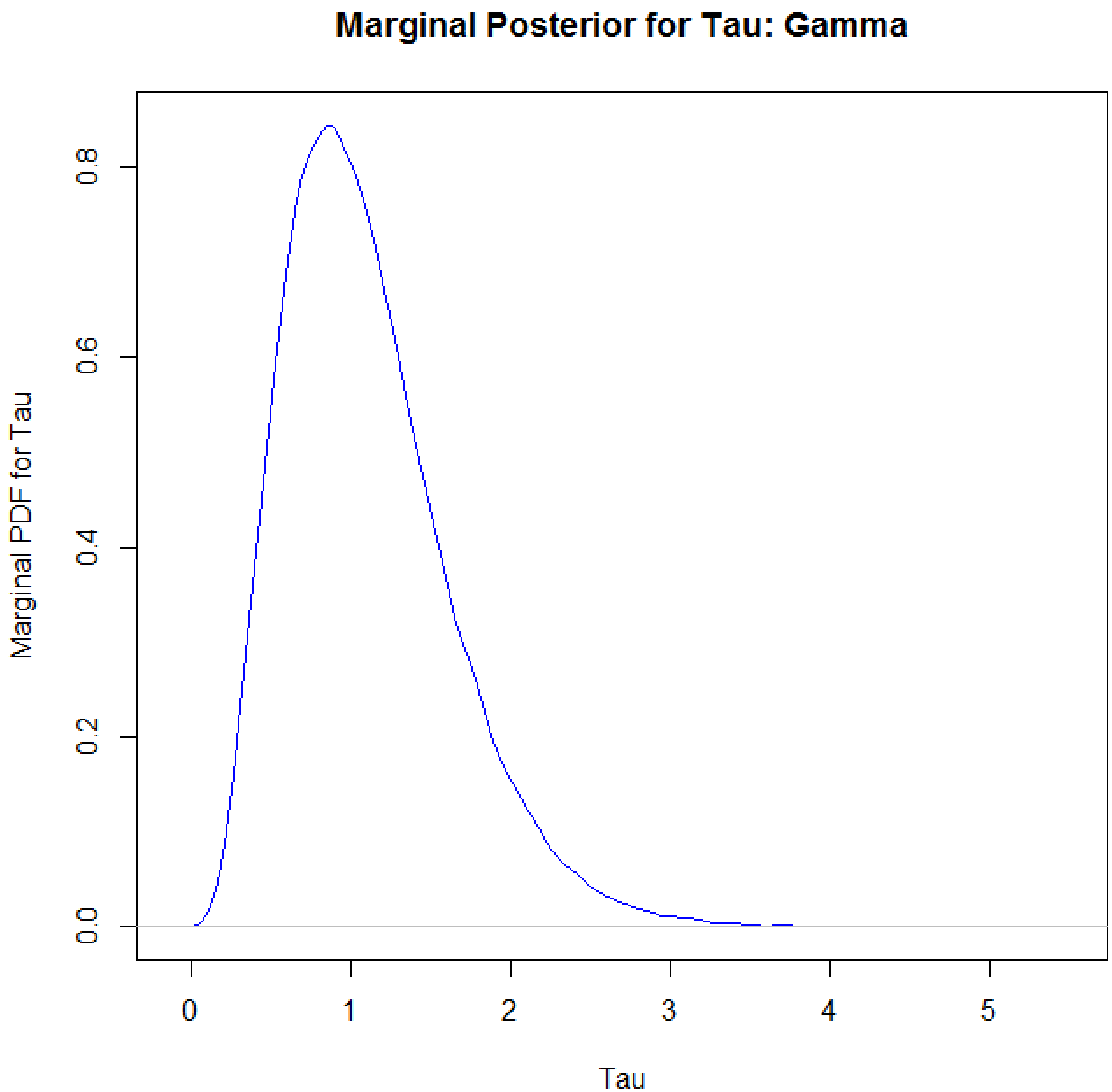
```
# Plot the marginal posterior p.d.f.'s, using nonparametric smoothing

plot(density(mub), col=2,main="Marginal Posterior for Mu: Student-t", xlab="Mu",
ylab="Marginal PDF for Mu")


plot(density(taub), col=4, main="Marginal Posterior for Tau: Gamma", xlab="Tau",
ylab="Marginal PDF for Tau")
```


MCMC for Normal-Gamma - Trace for Mu
Mu
1.6
1.4
1.2
1.0
0.8
0.6
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Repetitions


Marginal Posterior for Mu: Student-t
Marginal PDF for Mu
0
1
2
3
4
5
0.6
0.8
1.0
1.2
1.4
1.6
Mu

MCMC for Normal-Gamma - Trace for Tau




Marginal Posterior for Tau: Gamma
Marginal PDF for Tau
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
Tau

```
> summary(mub)  ; var(mub)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.7570  0.8629  1.0760  1.0750  1.2850  3.9620
[1] 0.1160567
> ybar        # The mean of the poterior for Mu should be ybar ( = 1.0746)
[1] 1.074626
>
> skewness(mub)  #  the skewness of Student-t is zero
[1] -0.007609531
> kurtosis(mub)  # The EXCESS kurtosis for Student-t (n-1) is 6/(n-5)=1.2; so kurtosis = 4.2
[1] 4.28897
```

Bayes estimate of μ is 1.075, if we have a Quadratic loss function, or if we have an Absolute-error loss function.

A 50% BCI (& HPD interval) for μ is [0.8629 ; 1.2850]

```
> summary(taub)   ; var(taub)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02333 0.71940 1.02200 1.10200 1.39600 5.38100
[1] 0.2717406
> skewness(taub)  #  the skewness of Gamma is (2/sqrt(shape))= (2/sqrt(n/2)=0.8944
[1] 0.9472144
> kurtosis(taub)  # excess kurtosis for Gamma is  (6/shape) = 6/(n/2)=1.2
[1] 4.34717
```

Bayes estimate of τ is 1.102, if we have a Quadratic loss function, and 1.022 if we have an Absolute-error loss function.

A 50% BCI for τ is [0.7194 ; 1.3960]

- Get other quantiles of the marginal posteriors so we can create BCI's:

$$\hat{\mu} = 1.075$$

```
> quantile(mub,  probs = c(1, 2.5, 5, 10, 90, 95, 97.5, 99)/100)
       1%      2.5%        5%       10%       90%       95%     97.5%       99% 
0.2310106 0.3975249 0.5232828 0.6583863 1.4928146 1.6281092 1.7532289 1.9194482 
>
> quantile(taub,  probs = c(1, 2.5, 5, 10, 90, 95, 97.5, 99)/100)
       1%      2.5%        5%       10%       90%       95%     97.5%       99% 
0.2590156 0.3307797 0.4048896 0.5089736 1.7996453 2.0729340 2.3293200 2.6657315 
> 1/var(y)
```

$$\hat{\tau} = 1.10$$

- Next, we'll look at some examples involving the Gibbs sampler in situations where we don't know the forms of the marginal posterior p.d.f.'s.

- That is, there will be a *genuine need* for the G.S.