Exact and MC-based Bayesian inference

Bernoulli model and two priors

Hedibert F. Lopes

2026-05-19

Contents

---

1 Bayesian basics

Recall the central identity of Bayesian inference, the Bayes theorem,

$$p(\theta \mid \text{data}) \;=\; \frac{p(\theta)\, p(\text{data} \mid \theta)}{p(\text{data})},$$

where the *data* is modeled via $p(\text{data} \mid \theta)$ and the prior for θ is represented by $p(\theta)$. The denominator of the above identity is commonly known as the **marginal likelihood**, also known as **normalising constant**, **integrated likelihood** or **prior predictive**, is nothing but the integral of the numerator with respect to θ:

$$p(\text{data}) \;=\; \int_\Theta p(\theta)\, p(\text{data} \mid \theta)\, d\theta.$$

For notational purpose, let uscall x_{obs} the set of observed *data* The **Posterior predictive** for new observations x_{new} given existing data is given by

$$p(x_{new} \mid x_{obs}) \;=\; \int_\Theta p(x_{new} \mid \theta, x_{obs})\, p(\theta \mid x_{obs})\, d\theta.$$

Notice that the posterior predictive and the prior predictive represent the likelihood of the data, existing or future, based on the different measures of uncertainty quantification, either $p(\theta)$ or $p(\theta \mid x_{obs})$.

2 Example: iid Bernoulli data and two priors

Let us consider the following illustrative example, where we assume that n observations are drawn from independent and identically distributed Bernoulli trials. More precisely,

$$x_1, \ldots, x_n \mid \theta \overset{\text{iid}}{\sim} \text{Ber}(\theta),$$

with the observation density explicitly given by

$$p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i} = \begin{cases} \theta & x_i = 1 \\ 1-\theta & x_i = 0 \end{cases},$$

for $i = 1, \ldots, n$.

2.1 Likelihood function

Combining the density of n observations, we arrive at the joint likelihood function

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \;=\; \theta^{\sum x_i}(1-\theta)^{n-\sum x_i},$$

which can be rewritten as a function of the sufficient statistics n (number of trials) and $y_n = \sum_{i=1}^{n} x_i$ (number of successes):

$$L(\theta \mid y_n) \;=\; \theta^{y_n}(1-\theta)^{n-y_n}.$$

2.2 Prior 1: Beta prior and conjugate posterior

We start with the standard conjugate prior, a Beta distribution with parameters α_0 and β_0, ie. $\theta \sim Beta(\alpha_0, \beta_0)$:

$$p(\theta) \propto \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1},$$

for $\alpha_0, \beta_0 > 0$. It is easy to check that

$$E(\theta) = \frac{\alpha_0}{\alpha_0 + \beta_0} \quad \text{and} \quad V(\theta) = \frac{\alpha_0\beta_0}{(\alpha_0+\beta_0)^2(\alpha_0+\beta_0+1)}.$$

Combining the likelihood and this prior, we arrive at the posterior for θ as

$$p(\theta \mid x_1, \ldots, x_n) = p(\theta|y_n) \;\propto\; \underbrace{\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}_{\text{prior}} \cdot \underbrace{\theta^{y_n}(1-\theta)^{n-y_n}}_{\text{likelihood}} \;=\; \theta^{(\alpha_0+y_n)-1}(1-\theta)^{(\beta_0+n-y_n)-1},$$

which is the kernel of $Beta(\alpha_1, \beta_1)$ with hyperparameters

$$\alpha_1 = \alpha_0 + y_n \quad \text{and} \quad \beta_1 = \beta_0 + (n - y_n),$$

which depends on α_0, β_0, n and y_n. Hence

$$(\theta \mid y_n) \sim Beta(\alpha_1, \beta_1),$$

such that

$$E(\theta \mid y_n) = \frac{\alpha_1}{\alpha_1 + \beta_1}.$$

2.2.1 Posterior mean as a weighted average

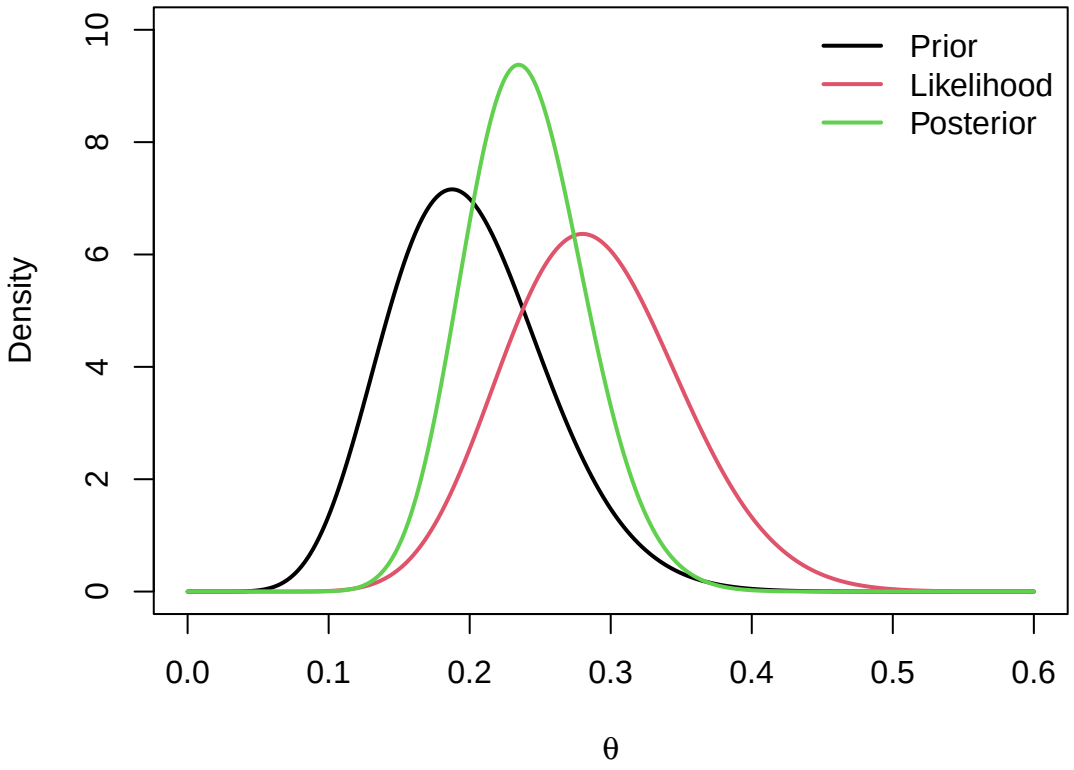
$$E(\theta \mid y_n) = \frac{\alpha_0 + y_n}{\alpha_0 + \beta_0 + n} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{\alpha_0}{n}}{\frac{\alpha_0+\beta_0}{n} + 1} \xrightarrow{n\to\infty} \frac{\sum x_i}{n} = \bar{x}_n.$$

As a **numerical example**, let us assume that $\alpha_0 = 10$, $\beta_0 = 40$, $y_n = 14$ and $n = 50$. Therefore, $(\theta \mid y_{50} = 14) \sim Beta(24, 86)$ and and

$$E(\theta \mid y_{50} = 14) = 0.24 \quad \text{and} \quad V(\theta \mid y_{50} = 14) = (0.043)^2.$$

The exact 95% posterior credibility interval for θ can be obtained from the quantiles of $Beta(24, 86)$. The interval is $(0.15, 0.30)$.

```r
alpha0 = 10
beta0  = 40
yn     = 14
n      = 50
alpha1 = alpha0 + yn
beta1  = beta0 + n - yn

thetas = seq(0,0.6,length=1000)
plot(thetas,dbeta(thetas,alpha0,beta0),type="l",xlab=expression(theta),
    ylab="Density",lwd=2,ylim=c(0,10))
lines(thetas,dbeta(thetas,yn+1,n-yn+1),col=2,lwd=2)
lines(thetas,dbeta(thetas,alpha1,beta1),col=3,lwd=2)
legend("topright",legend=c("Prior","Likelihood","Posterior"),col=1:3,lwd=2,bty="n")
```



2.2.2 Posterior Predictive

With $(\theta \mid y_{50} = 14) \sim \text{Beta}(24, 86)$, the posterior predictive probability that the next observation $x_{51} = 1$ is

$$\Pr(x_{51} = 1 \mid y_{50} = 14) = \int_0^1 Pr(x_{51} = 1 \mid \theta, y_{50} = 14)\, p(\theta \mid y_{50} = 14)\, d\theta.$$

Due to the conditional independence of the trails given θ,

$$Pr(x_{51} = 1 \mid \theta, y_{50} = 14) = Pr(x_{51} = 1 \mid \theta) = \theta.$$

Therefore,

$$\Pr(x_{51} = 1 \mid y_{50} = 14) = \int_0^1 \theta p(\theta \mid y_{50} = 14)\, d\theta = E(\theta | y_{50} = 14) = \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\alpha_0 + y_{50}}{\alpha_0 + \beta_0 + n}.$$

2.3 Prior 2: Logit-Normal Prior

Let us now assume that we have prior information for γ where

$$\gamma = \text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right),$$

so that $\theta = \dfrac{1}{1+e^{-\gamma}}$.

To derive the prior for θ as a function of the prior for γ, recall that derivative of γ with respect to θ is given by

$$\frac{d\gamma}{d\theta} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

We place a normal prior on γ:

$$\gamma \sim N(\mu_0, \sigma_0^2),$$

for (μ_0, σ_0^2) known hyperparameters. By the change-of-variables formula, the induced prior on θ is

$$p_\theta(\theta) = p_\gamma\left(\log\left(\frac{\theta}{1-\theta}\right)\right)|J(\gamma,\theta)| = \phi_{\mu_0,\sigma_0^2}\left(\log\left(\frac{\theta}{1-\theta}\right)\right)\frac{1}{\theta(1-\theta)},$$

where $\phi_{\mu_0,\sigma_0^2}(\gamma)$ denotes the $N(\mu_0,\sigma_0^2)$ density evaluated at γ and

$$J(\gamma,\theta) = \frac{d\log\left(\frac{\theta}{1-\theta}\right)}{d\theta},$$

is the Jacobian of the transformation. Therefore,

$$p(\theta) \;\propto\; \exp\left\{-\frac{0.5}{\sigma_0^2}\left(\log\left(\tfrac{\theta}{1-\theta}\right)-\mu_0\right)^2\right\}\frac{1}{\theta(1-\theta)}.$$

2.3.1 Posterior Kernel under Logit-Normal Prior

$$p(\theta \mid y_n, n) \;\propto\; \theta^{y_n}(1-\theta)^{n-y_n}p(\theta) \;\propto\; \theta^{y_n}(1-\theta)^{n-y_n}\phi_{\mu_0,\sigma_0^2}\left(\log\left(\frac{\theta}{1-\theta}\right)\right)\frac{1}{\theta(1-\theta)}.$$

Notice that the likelihood has a Beta kernel which be combined with the final part of the prior:

$$p(\theta \mid y_n, n) \;\propto\; \theta^{y_n-1}(1-\theta)^{n-y_n-1}p(\theta) \;\propto\; \theta^{y_n}(1-\theta)^{n-y_n}\phi_{\mu_0,\sigma_0^2}\left(\log\left(\frac{\theta}{1-\theta}\right)\right).$$

There, one can think of the posterior for θ as a function of the $Beta(y_n, n-y_n)$ multiplied by a nonlinear function of θ, say

$$g(\theta) = \phi_{\mu_0,\sigma_0^2}\left(\log\left(\frac{\theta}{1-\theta}\right)\right),$$

so that

$$p(\theta \mid y_n, n) \propto g(\theta)p_{Beta}(\theta; y_n, n-y_n).$$

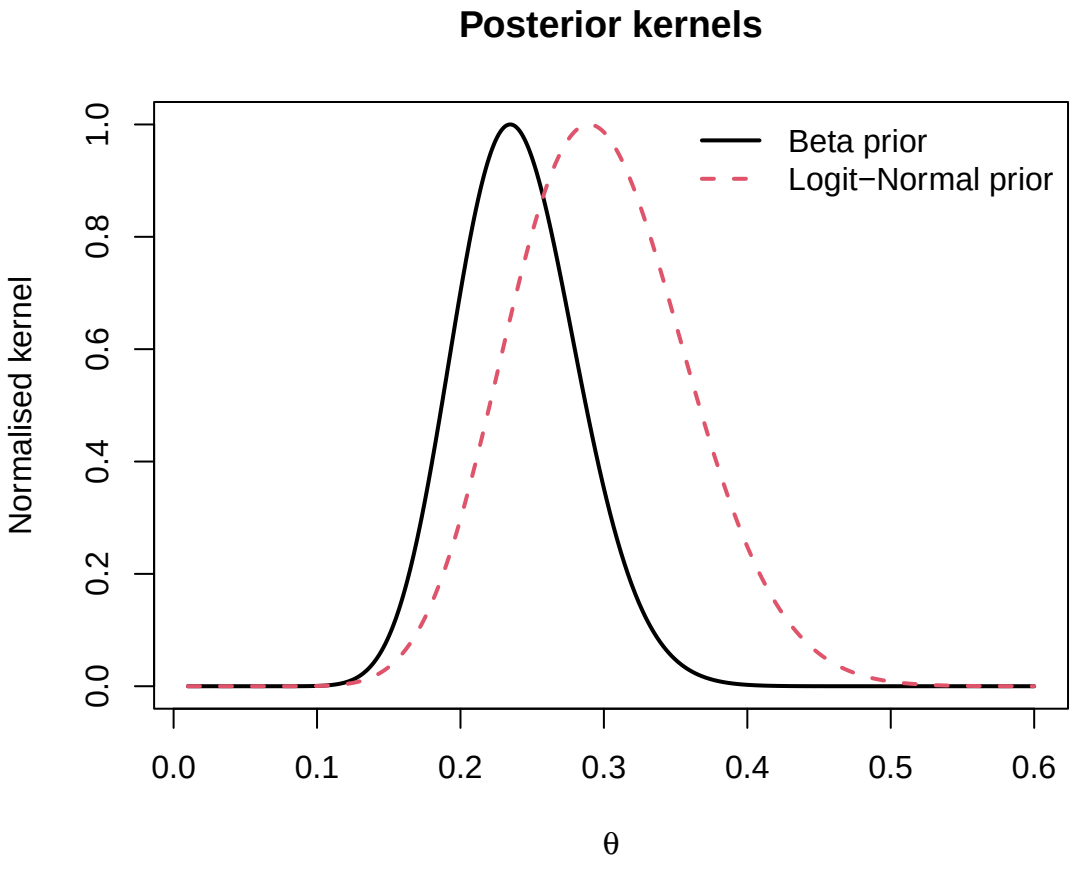
This posterior has **no closed form**, motivating numerical methods, such as rejection sampling, sampling importance re-sampling (SIR), random-walk Metropolis-Hastings, independent Metropolis-Hastings, Gibbs sampler, hamiltonean Monte Carlo, etc.

```
mu0    = 0
sigma0 = 1
# Prior 1: Beta kernel (unnormalised)
post1 = function(theta){
  dbeta(theta,alpha0,beta0)*dbeta(theta,yn+1,n-yn+1)
```

```r
}

# Prior 2: Logit-Normal kernel (unnormalised)
post2 = function(theta){
  theta^(yn-1)*(1-theta)^(n-yn-1)*dnorm(log(theta/(1-theta)),mu0,sigma0)
}
```

2.4 Kernel Comparison Plot

```r
thetas       = seq(0.01, 0.6, length = 1000)
post.kernel1 = post1(thetas)
post.kernel2 = post2(thetas)

plot(thetas, post.kernel1  / max(post.kernel1),
     type = "l", lwd = 2,
     xlab = expression(theta), ylab = "Normalised kernel",
     main = "Posterior kernels")
lines(thetas, post.kernel2 / max(post.kernel2),
      col = 2, lwd = 2, lty = 2)
legend("topright",
       legend = c("Beta prior", "Logit-Normal prior"),
       col = c(1, 2), lty = c(1, 2), lwd = 2, bty = "n")
```



2.5 Monte Carlo-based posterior inference

We will now perform posterior inference via Monte Carlo schemes for illustration. We start with a SIR scheme and then try two MH ones.

2.5.1 Sampling Importance Resampling (SIR)

Proposal: $\theta^{(m)} \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$, density $g(\theta) = 1$.

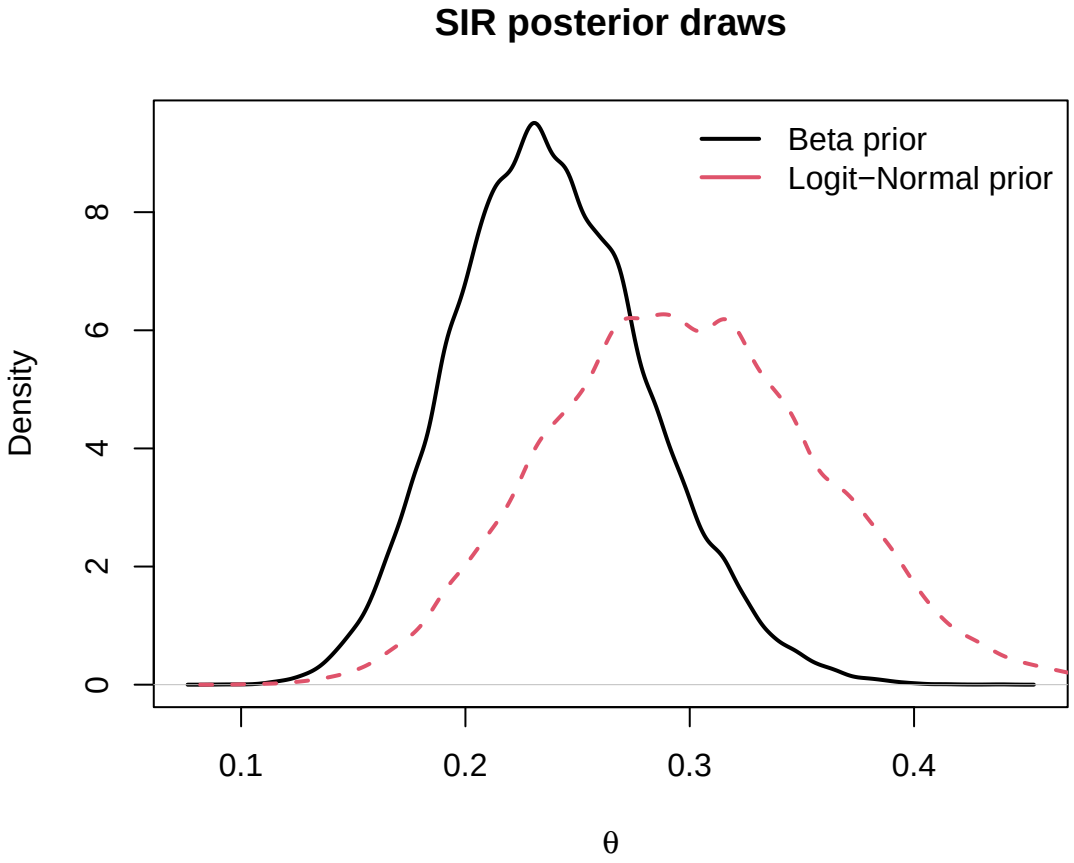
Unnormalised importance weights:

$$w^{(m)} = \frac{p(\theta^{(m)} \mid \text{data})}{g(\theta^{(m)})} \propto p(\theta^{(m)} \mid \text{data}).$$

Resample M draws with replacement proportional to $\{w^{(m)}\}$.

```
M           = 50000
theta.draw = runif(M)

# Importance weights (Uniform proposal => w = posterior kernel)
weight1 = post1(theta.draw)
weight2 = post2(theta.draw)

# Resample
theta.post1 = sample(theta.draw, size = M, replace = TRUE, prob = weight1)
theta.post2 = sample(theta.draw, size = M, replace = TRUE, prob = weight2)
```

```
par(mfrow = c(1, 1))
plot(density(theta.post1),type = "l", lwd = 2,xlab = expression(theta),
     ylab = "Density", main = "SIR posterior draws")
lines(density(theta.post2), col = 2, lwd = 2, lty = 2)
legend("topright",col =1:2, lty=1, lwd = 2, bty = "n",
       legend = c("Beta prior", "Logit-Normal prior"))
```




```
# Posterior mean comparison
pm = c(mean(theta.post1), mean(theta.post2))
names(pm) = c("SIR: Beta prior", "SIR: Logit-Normal")
print(round(pm, 5))
```

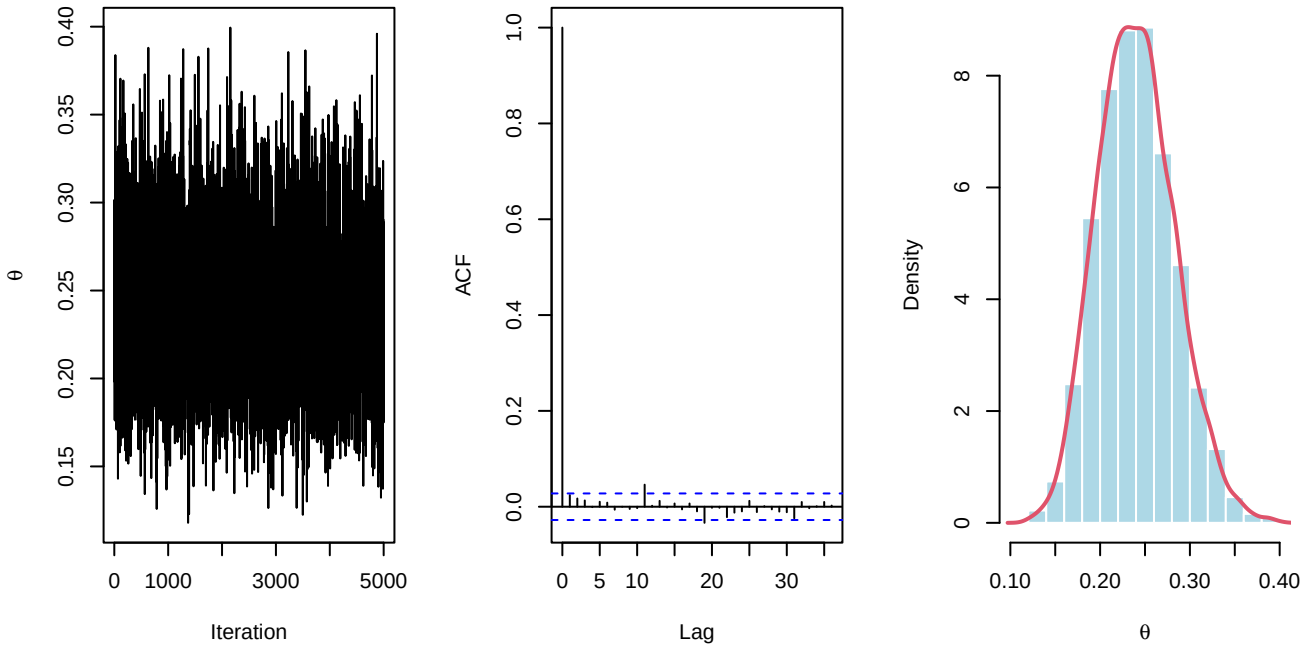
```
##   SIR: Beta prior SIR: Logit-Normal
##           0.23890           0.29817
```

2.5.2 Random walk Metropolis-Hastings

```
par(mfrow = c(1, 3))

thetas.draw1 = NULL
theta       = 0.5
burnin      = 1000
lag         = 10
M_mcmc      = 5000
niter       = burnin + lag * M_mcmc

for (iter in 1:niter) {
  theta.star = rnorm(1, theta, 0.2)
  if (theta.star > 0 && theta.star < 1) {
    alpha_acc = min(1, post1(theta.star) / post1(theta))
    if (runif(1) < alpha_acc) theta = theta.star
  }
  thetas.draw1 = c(thetas.draw1, theta)
```

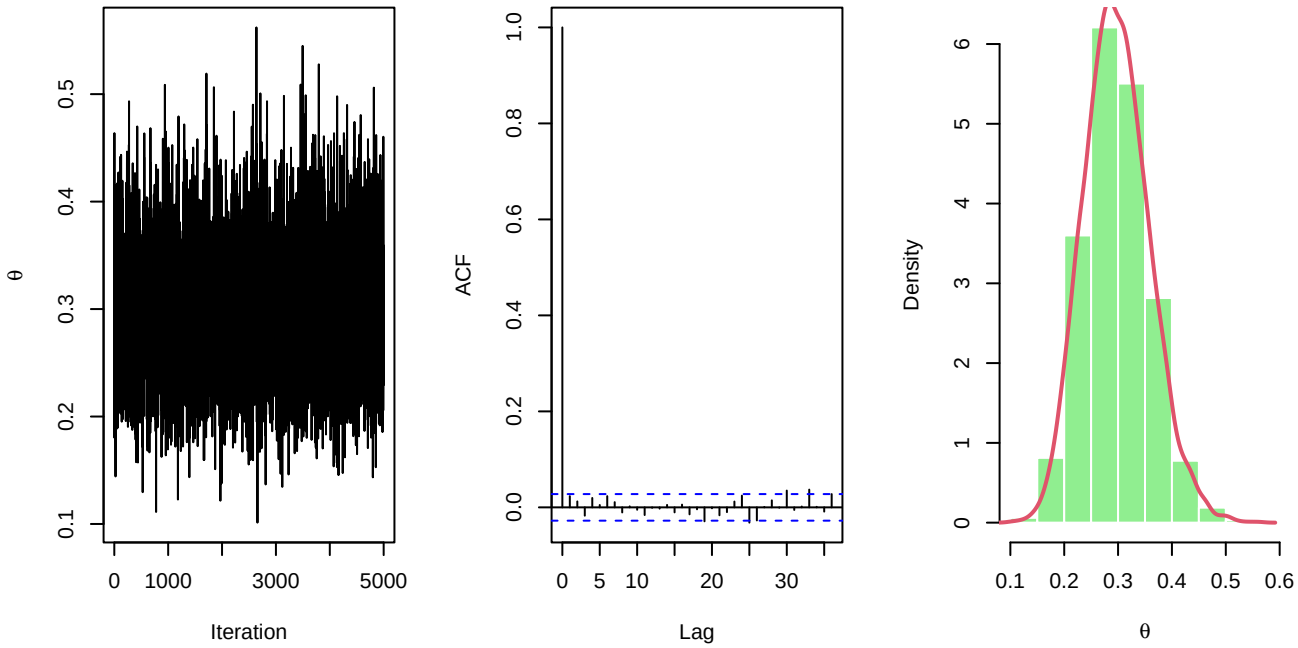
```
}

thetas.draw1 = thetas.draw1[seq(burnin + 1, niter, by = lag)]

ts.plot(thetas.draw1,xlab = "Iteration", ylab = expression(theta),main = "")
acf(thetas.draw1,   main = "")
hist(thetas.draw1,freq=FALSE,col="lightblue",border="white",xlab=expression(theta),main="")
lines(density(thetas.draw1),col = 2, lwd = 2)
```

2.5.2.1 Beta Prior



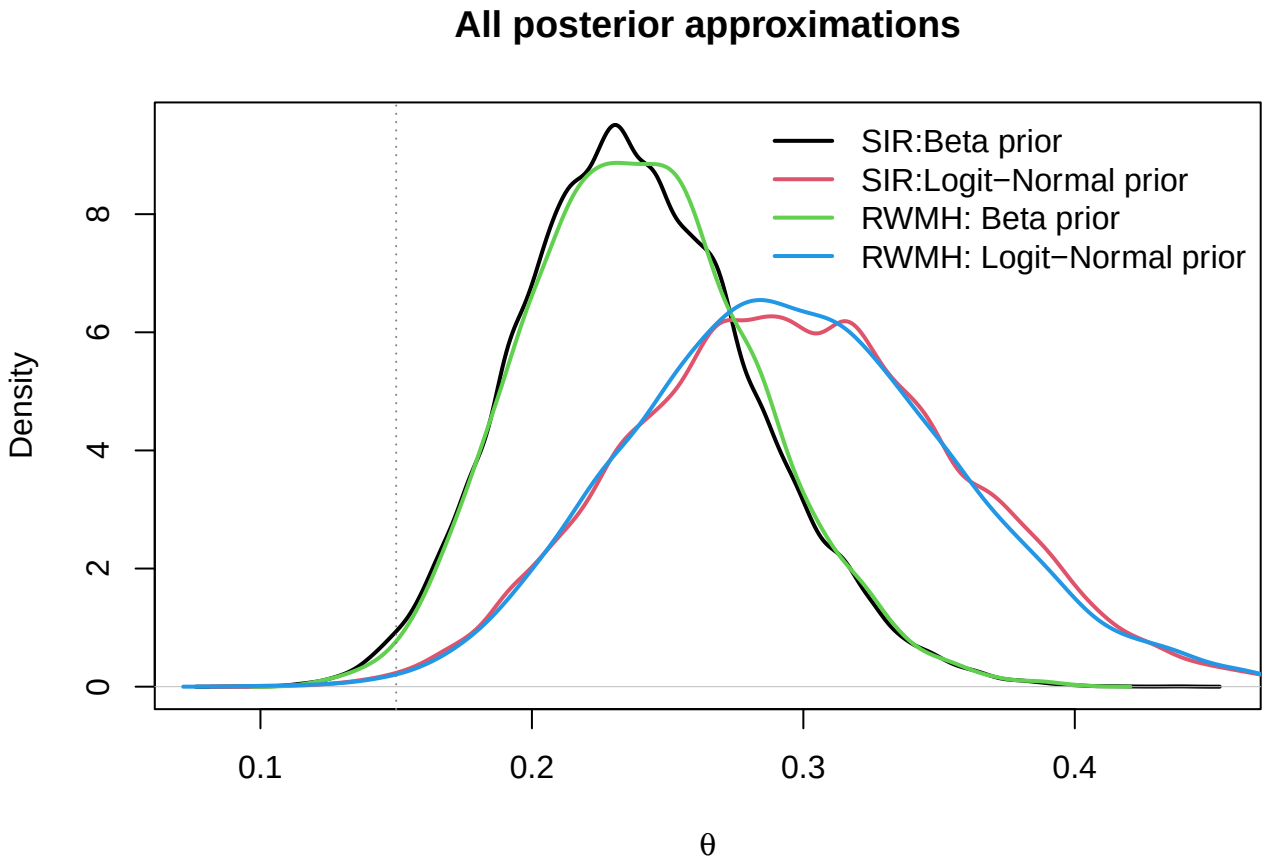

```
par(mfrow = c(1, 3))
thetas.draw2 = NULL
theta        = 0.5

for (iter in 1:niter) {
  theta.star = rnorm(1, theta, 0.25)
  if (theta.star > 0 && theta.star < 1) {
    alpha_acc = min(1, post2(theta.star) / post2(theta))
    if (runif(1) < alpha_acc) theta = theta.star
  }
  thetas.draw2 = c(thetas.draw2, theta)
}

thetas.draw2 = thetas.draw2[seq(burnin + 1, niter, by = lag)]

ts.plot(thetas.draw2,xlab = "Iteration", ylab = expression(theta),main = "")
acf(thetas.draw2,main = "")
hist(thetas.draw2,freq=FALSE,col="lightgreen",border="white",xlab=expression(theta),main="")
lines(density(thetas.draw2),col=2,lwd=2)
```

2.5.2.2 Logit-Normal Prior



2.6 All Posterior Approximations

```
par(mfrow = c(1, 1))
plot(density(theta.post1),type = "l", lwd = 2, col = 1,xlab = expression(theta),
     ylab = "Density",main = "All posterior approximations")
lines(density(theta.post2),  col = 2, lwd = 2)
lines(density(thetas.draw1),  col = 3, lwd = 2)
lines(density(thetas.draw2), col = 4, lwd = 2)
abline(v = 0.15, col = "grey50", lty = 3)
legend("topright",col=1:4,lty=1,lwd=2,bty="n",legend=c("SIR:Beta prior",
       "SIR:Logit-Normal prior","RWMH: Beta prior","RWMH: Logit-Normal prior"))
```


All posterior approximations
SIR:Beta prior
SIR:Logit−Normal prior
RWMH: Beta prior
RWMH: Logit−Normal prior
Density
0
2
4
6
8
0.1
0.2
0.3
0.4
θ