Bayesian Models 1

Charlie Qu

June 14, 2017

Outline

This is a first short writing about Conjugate Poisson-Gamma model, and comparing posteriors of two parameters via MC simulation. We are interested in estimating θ, mean number of eggs per nest.

1. Observe, Y_i, the number of eggs laid in nest i, for i=1,.,n.
2. Assume $Y_i|\theta$??? i.i.d. Poisson(θ)

We can show that $y =???_{i=1}^{n} Y_i$ n i=1???Poisson(nθ) We can use this example to consider: highest posterior density region;Sensitivity analysis. In both parts, we'll use a Gamma prior and take y=9 , n=5 .

Under a Gamma Prior

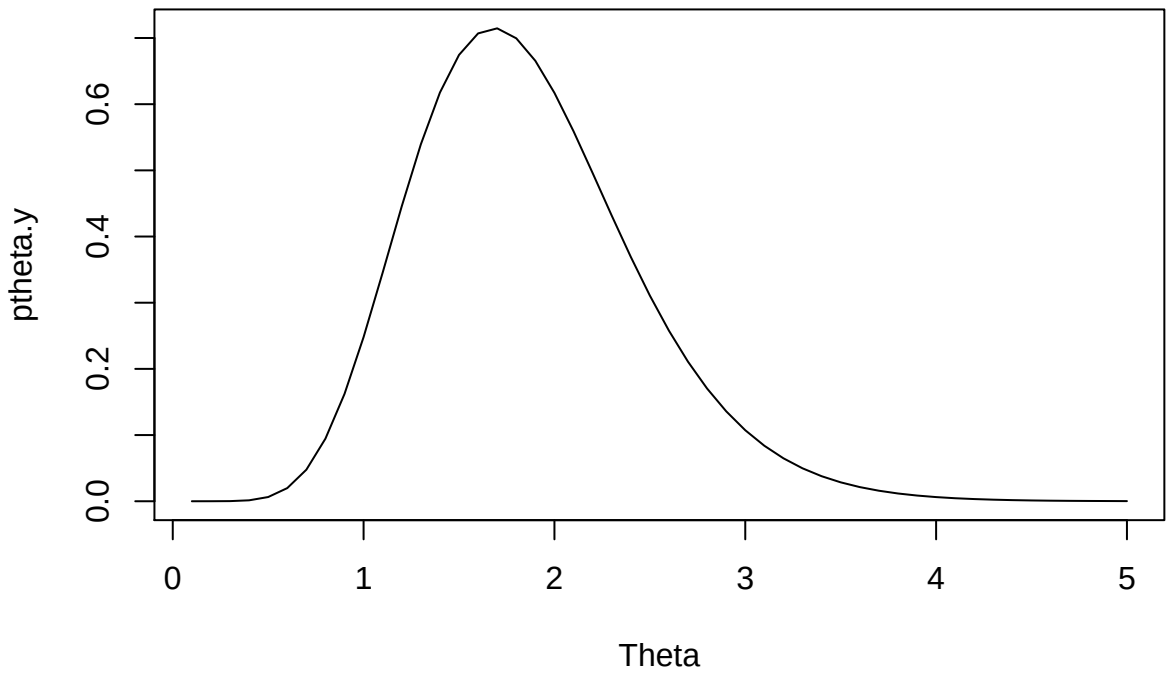
Assume θ???G(a=1.25,b=0.5)

(1) Under this prior, θ>0
(2) θ has prior mean E[θ]=2.5. Because the gamma distribution is conjugate for the Poisson sampling model, we know θ|y???G(a+y,b+n).

```
y <- 9; n <- 5
a <- 1.25; b <- 0.5
a.y <- a + y; b.y <- b + n
```

Posterior Under Gamma Prior

```
Theta <- seq(0.1, 5, by = 0.1)
ptheta.y <- dgamma(Theta, a.y, b.y)
plot(Theta, ptheta.y, type = "l")
```



95% CI Under Gamma Prior

We can use the qgamma function to compute an exact 95% confidence interval.

```
citheta.y <- qgamma(c(0.025, 0.975),
                    a.y, b.y)
citheta.y
```

```
## [1] 0.9032629 3.1659277
```

Some values of θ outside of the CI have a higher posterior density than those inside it, maybe we should calculate a 95% highest posterior density interval instead!

95% HPDI Under Gamma Prior

We can iteratively find the 95% HPDI by discretizing θ.

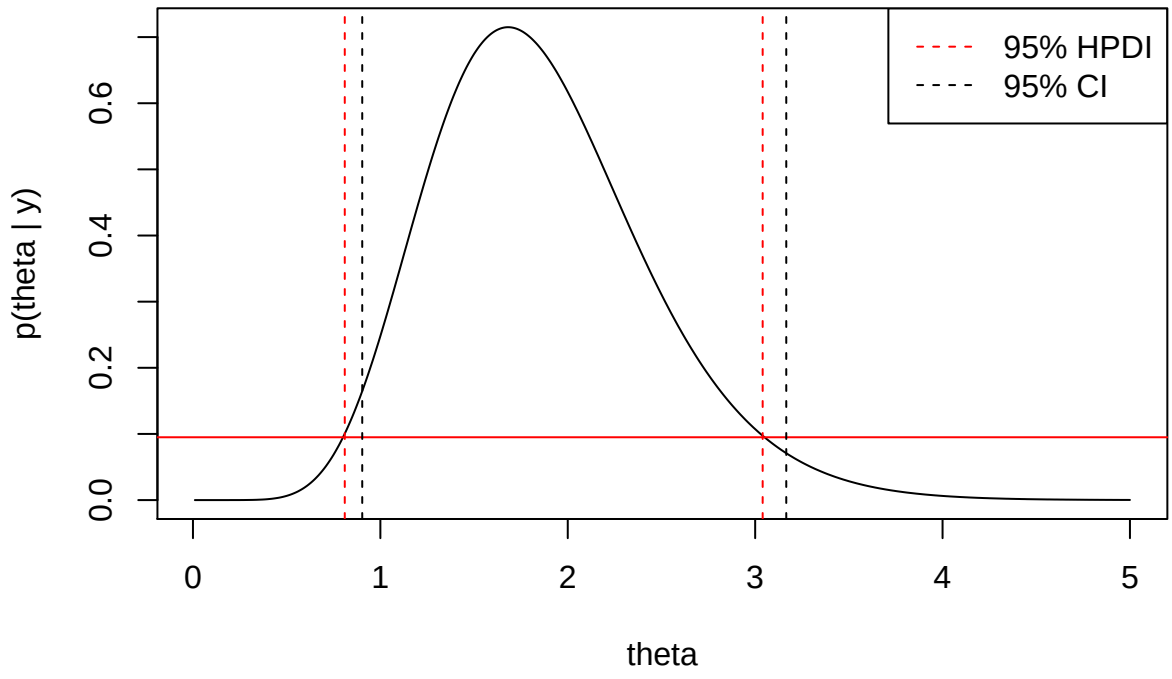
```
# Discretize theta
Theta <- seq(0.01, 5, by = 0.01)
ptheta.y <- dgamma(Theta, a.y, b.y)
# Start with a horizontal line on top of posterior
hpd.cutoff <- max(ptheta.y)
# Find intersection of first line w/posterior (will be one point)
hptheta.y <- range(Theta[ptheta.y >= hpd.cutoff])
# Find area between line and posterior (will be 0)
hpd.p <- pgamma(hptheta.y[2], a.y, b.y) - pgamma(hptheta.y[1], a.y, b.y)
```

```
while(hpd.p <= 0.95 & hpd.cutoff > 0) {
  hpd.cutoff <- hpd.cutoff - 0.005 # Move hline down
  # Find intersection of line w/posterior
  hptheta.y <- range(Theta[ptheta.y >= hpd.cutoff])
  # Compute area between line and posterior
  hpd.p <- pgamma(hptheta.y[2], a.y, b.y) - pgamma(hptheta.y[1], a.y, b.y)
}
hptheta.y
```

```
## [1] 0.81 3.04
```

The 95% HPD interval is shifted a little bit to the left, but otherwise similar to the 95% confidence interval.

```
plot(Theta, ptheta.y, type = "l",
     xlab = "theta", ylab = "p(theta | y)")
abline(h = hpd.cutoff, col = "red")
abline(v = citheta.y, lty = 2)
abline(v = hptheta.y, lty = 2, col = "red")
legend("topright", lty = c(2, 2), col = c("red", "black"),
       legend = c("95% HPDI", "95% CI"))
```



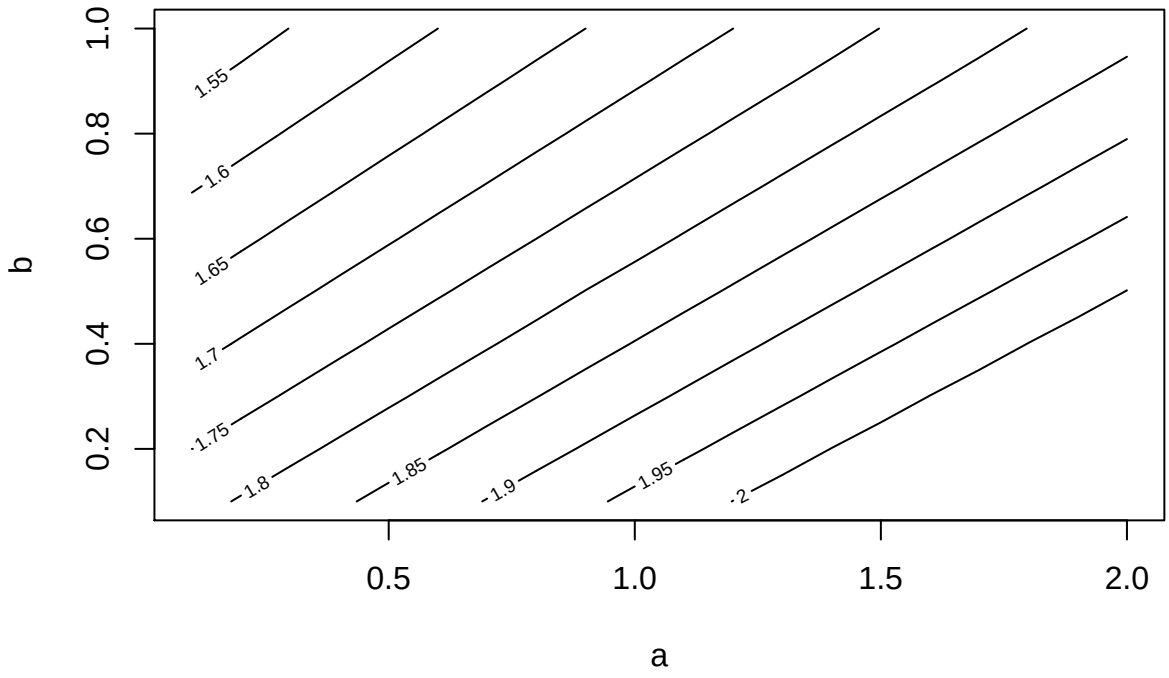
Sensitivity Analysis

Suppose we chose different parameters aa and bb for our gamma prior distribution.

We can assess how sensitive the posterior mean,$(a+y)/(b+n)$ is to the choice of aa and bb by computing

the posterior mean for a={0.1,.,2}a={0.1,.,2} and b={0.1,.,1}b={0.1,.,1}.

```
a <- seq(0.1, 2, by = 0.1)
b <- seq(0.1, 1, by = 0.1)
etheta.y <- matrix(nrow = length(a), ncol = length(b))
for (i in 1:length(a)) {
  for (j in 1:length(b)) {
    etheta.y[i, j] <- (y + a[i])/(n + b[j])
  }
}
contour(a, b, etheta.y,
        levels = seq(1, 2, by = 0.05),
        xlab = "a", ylab = "b")
```



For comparison, we can also perform sensitivity analysis for another scenario, where n=20n=20.

```
y.1 <- 36
n.1 <- 20
etheta.y.1 <- matrix(nrow = length(a), ncol = length(b))
for (i in 1:length(a)) {
  for (j in 1:length(b)) {
    etheta.y.1[i, j] <- (y.1 + a[i])/(n.1 + b[j])
    etheta.y[i, j] <- (y + a[i])/(n + b[j])
  }
}
```

We can see that the posterior mean is less sensitive to changes in aa and bb when the sample size is larger.

We'll be taking random draws from various distributions in this lab.

To get the same answers each time you run the code, you need to set a seed.

```
set.seed(1)
```

Note that we can compute the mode and any other feature of the Beta distribution that can be written as a function of its parameters.

Problem

(1) Interested in estimating θ, mean number of eggs per nest;
(2) Observe Y_i, the number of eggs laid in nest i, for i=1,.,n;
(3) Assume $Y_i|\theta \sim Poisson(\theta)$;
(4) Assume θ???Θ=$\{0.1,.,5.0\}$, and p(θ)=150, for each θ in Θ.

We can show $y = \Sigma_{i=1}^{n} Y_i \sim Poisson(n\theta)$

Suppose we have n=20, y=36.

Goal: compute posterior means, variances and 95% intervals using Monte Carlo approximation.

First, we need the posterior distribution, p(θ|y).

Inference Under Another Continuous Prior

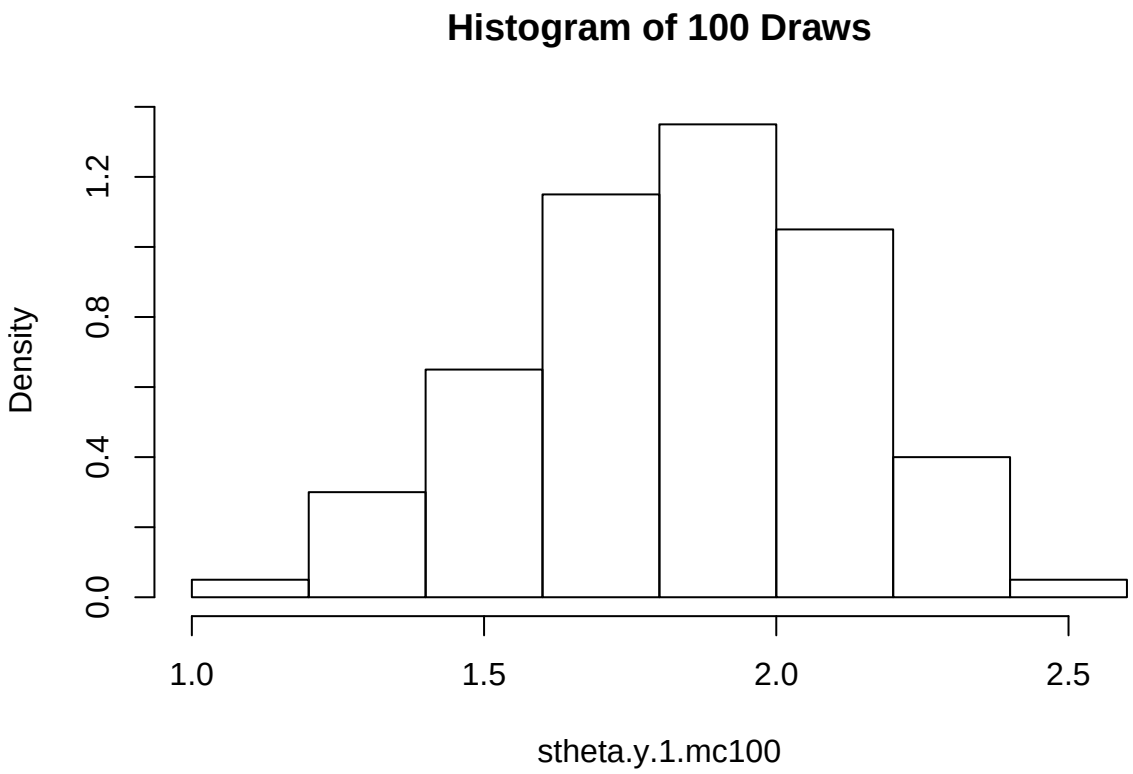
```
y.1 <- 36
n.1 <- 20

Theta <- seq(0.1, 5, by = 0.1)
py.theta.1 <- dpois(y.1, n.1*Theta)
ptheta <- rep(1/length(Theta), length(Theta))
pytheta.1 <- py.theta.1*ptheta
ptheta.y.1 <- pytheta.1/sum(pytheta.1)
```

Sampling from the Posterior Distribution

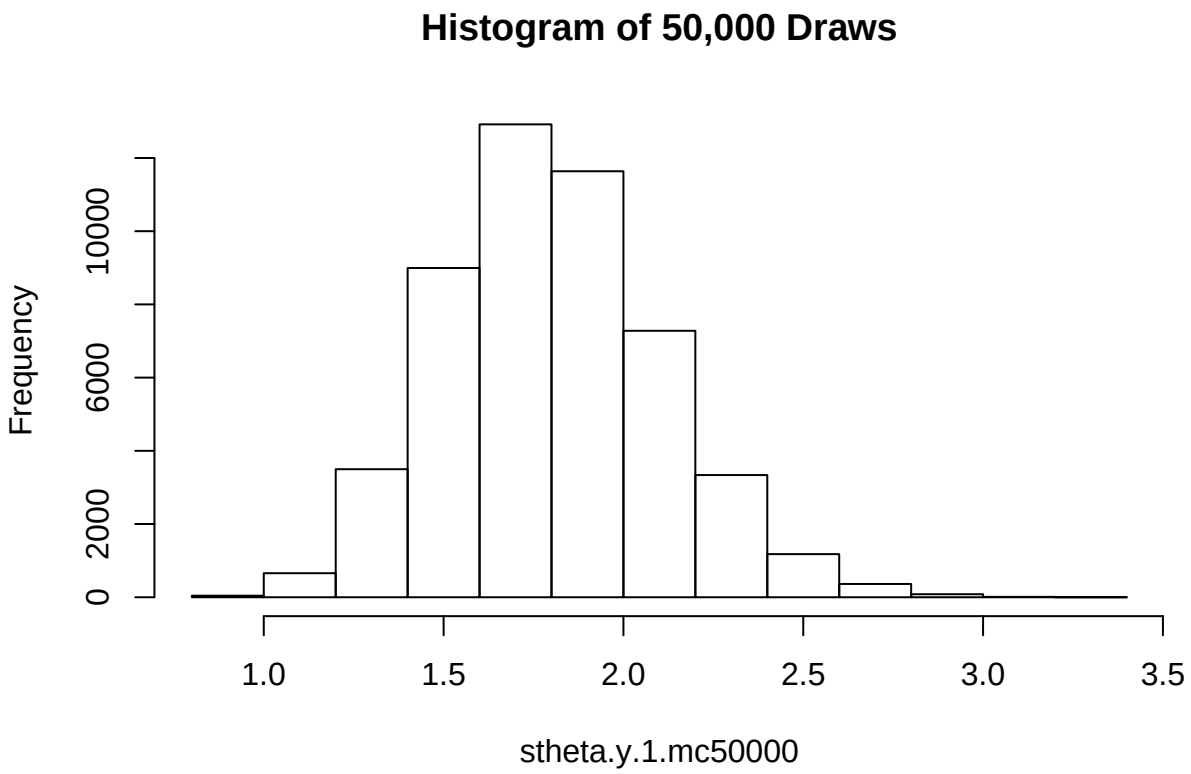
We can use the sample function to draw 100100 values from the posterior distribution.

```
stheta.y.1.mc100 <- sample(x = Theta,
                           size = 100,
                           replace = TRUE,
                           prob = ptheta.y.1)
hist(stheta.y.1.mc100, freq = FALSE,
     main = "Histogram of 100 Draws")
```



We can use the sample function to draw 50,00050,000 values from the posterior distribution.

```
stheta.y.1.mc50000 <- sample(x = Theta,
                             size = 50000,
                             replace = TRUE,
                             prob = ptheta.y.1)
hist(stheta.y.1.mc50000, freq = TRUE,
     main = "Histogram of 50,000 Draws")
```



As we take more samples from the posterior, the distribution of our samples looks more like p(θ|y).

MC Approximation of the Posterior Mean

The actual posterior mean is calculated to be E[θ|y]=1.85. Alternatively, we could get a Monte Carlo approximation for this quantity.

```
mean(stheta.y.1.mc100)
```

```
## [1] 1.887
```

```
mean(stheta.y.1.mc50000)
```

```
## [1] 1.85084
```

With a large enough sample from our posterior, we can get a very good approximation to the true E[θ|y]!

MC Approximation of the Posterior Variance

The true posterior variance is calculated to be V[θ|y]=0.0925. Alternatively, we could get a Monte Carlo approximation for this quantity.

```
var(stheta.y.1.mc100)
```

```
## [1] 0.08073838
```

```
var(stheta.y.1.mc50000)
```

```
## [1] 0.09278315
```

With a large enough sample from our posterior, we can get a very good approximation to the true V[θ|y]!

MC Approximation of the 95% CI

The true posterior variance is calculated to be V[θ|y]=0.0925. Alternatively, we could get a Monte Carlo approximation for this quantity.

```
quantile(stheta.y.1.mc100, c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 1.3475 2.3525
```

```
quantile(stheta.y.1.mc50000, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##   1.3   2.5
```

With a large enough sample from our posterior, we can get an approximate 95% confidence interval more easily than we could otherwise!

MC Approx. of Functions of Two Params.

Suppose we also compute the posterior for a second larger sample with n=40n=40 and y=72. What if we want to ask if θ_2 from this sample is greater than θ_1 from the first sample, with n=20 and y=36?

We'll assume that θ_1 and θ_2 are conditionally independent given the data.

First we need to compute the posterior distribution of θ_2.

```
n.2 <- 40
y.2 <- 72

py.theta.2 <- dpois(y.2, n.2*Theta)
ptheta <- rep(1/length(Theta), length(Theta))
pytheta.2 <- py.theta.2*ptheta
ptheta.y.2 <- pytheta.2/sum(pytheta.2)
stheta.y.2.mc50000 <- sample(x = Theta,
                             size = 50000,
                             replace = TRUE,
                             prob = ptheta.y.2)
```

Now we need to get a sample from the posterior of ??2??2 given the data.

```
stheta.y.2.mc50000 <- sample(x = Theta,
                             size = 50000,
                             replace = TRUE,
                             prob = ptheta.y.2)
```

Recall that we already got a posterior sample of θ_1 given the data with the same number of draws. We can obtain these posterior samples separately from each other because we have assumed that θ_1 and θ_2 are conditionally independent given the data.

First, we can get an approximation of $\Pr(\theta_2 \leq \theta_1 | y_1, y_2)$.

```
mean(stheta.y.2.mc50000 <= stheta.y.1.mc50000)
```

```
## [1] 0.57548
```

Given that both sampled data sets had the same sample mean and were modeled using the same prior, it's not surprising that $\Pr(\theta_2 \leq \theta_1 | y_1, y_2)$ is close to 0.5.

We can get an approximation of $E[\theta_2/\theta_1 | y_1, y_2]$

```
hist(stheta.y.2.mc50000/stheta.y.1.mc50000, main = "Hist. of theta2 / theta1")
abline(v = mean(stheta.y.2.mc50000/stheta.y.1.mc50000), col = "red")
```