MAST90104: Introduction to Statistical Learning

Solutions Week 12 Lab and Workshop

1 Lab

1. In practice, in the hatching eggs problem rather than using one observation x of the number of eggs in a region, there are likely to be m, say $\mathbf{x} = (x_1, \cdots, x_k)$ in k different parts of the one region. There are now different $N_i, i = 1, \cdots, k$ for the k different parts of the region and these random variables are assumed independent. Also rather than simulating from the full joint density, it is desired to simulate from the posterior distribution for $p, \mathbf{N} = (N_1, \cdots N_k)$ given the data $\mathbf{x}$.

   (a) What are the conditional densities than now need to be simulated at each step?
   Solution:
   The joint density now is:

$$f(\mathbf{N}, p, \mathbf{x}) = \left(\prod_{i=1}^{k} \binom{N_i}{x_i} p^{x_i} (1-p)^{N_i - x_i} \cdot \cdot \frac{\lambda^{N_i}}{N_i!} e^{-\lambda} \right) \cdot \frac{\Gamma(2+4)}{\Gamma(2)\Gamma(4)} p^{2-1} (1-p)^{4-1}.$$

   The joint density from which we wish to simulate is

$$f(\mathbf{N}, p|\mathbf{x}) \propto f(\mathbf{N}, p, \mathbf{x}).$$

   The required conditional densities are:

$$f(\mathbf{N}|p, \mathbf{x}), f(p|\mathbf{N}, \mathbf{x}) \propto f(\mathbf{N}, p, \mathbf{x}).$$

   The general trick from lectures says that if we can find probability densities that are proportional to the conditional densities, then these are the required conditional densities, because the constant of proportionality must then be 1.
   Considering only the terms in the joint density, $f(\mathbf{N}, p, \mathbf{x})$, which involve p shows that:

$$\begin{aligned} f(p|\mathbf{N}, \mathbf{x}) &\propto \left(\prod_{i=1}^{k} p^{x_i} (1-p)^{N_i - x_i} \cdot \right) \cdot p^{2-1} (1-p)^{4-1} \\ &= p^{x_\cdot + 2 - 1} (1-p)^{N_\cdot - x_\cdot + 4 - 1} \\ &\propto \text{Beta}(x_\cdot + 2, N_\cdot - x_\cdot + 4)(p), \end{aligned}$$

   where $x_\cdot = \sum_{i=1}^{k} x_i$, $N_\cdot = \sum_{i=1}^{k} N_i$ and $\text{Beta}(a, b)(p)$ is the Beta(a,b) density.
   The general trick says:

$$f(p|\mathbf{N}, \mathbf{x}) = \text{Beta}(x_\cdot + 2, N_\cdot - x_\cdot + 4)(p).$$

   Considering only the terms in the joint density, $f(\mathbf{N}, p, \mathbf{x})$, which involve $\mathbf{N}$ shows that:

$$\begin{aligned} f(\mathbf{N}|p, \mathbf{x}) &\propto \prod_{i=1}^{k} \binom{N_i}{x_i} (1-p)^{N_i - x_i} \cdot \frac{\lambda^{N_i}}{N_i!} \\ &\propto \prod_{i=1}^{k} \frac{N_i!}{(N_i - x_i)!} (1-p)^{N_i - x_i} \cdot \frac{\lambda^{N_i - x_i}}{N_i!} \\ &= \prod_{i=1}^{k} \frac{(\lambda(1-p))^{N_i - x_i}}{(N_i - x_i)!} \\ &\propto \prod_{i=1}^{k} \text{Poisson}(\lambda(1-p))(N_i - x_i) \end{aligned}$$

The general trick says the conditional distribution of $\mathbf{N}-\mathbf{x}$ given $(p, \mathbf{x})$ is that of k independent Poisson$(\lambda(1-p))$ random variables.

(b) Alter the code for the sampler from lectures

```
gibbs.f2 = function(x0, p0, N0, m, J){
# Genearates m samples of (x,p,N) values by the Gibbs sampler.
# In total J+m samples are generated but the first J are discarded.
# (x0,p0,N0) is the initial value.
x.seq <- p.seq <- N.seq <- rep(-1, J+m+1)
x.seq[1] <- x0; p.seq[1] <- p0; N.seq[1] <- N0
for(j in 2:(J+m+1)) {x.seq[j] <- rbinom(1, N.seq[j-1], p.seq[j-1])
p.seq[j] <- rbeta(1, (x.seq[j] + 2), (N.seq[j-1] - x.seq[j] + 4))
N.seq[j] <- rpois(1, 16 * (1 - p.seq[j])) + x.seq[j]}
result <- list(X=x.seq[(J+2):(J+m+1)],p=p.seq[(J+2):(J+m+1)],
N=N.seq[(J+2):(J+m+1)])
result
}
```

to simulate the posterior distribution for p, N given the data $x_1, \cdots, x_k$ where k is the number of regions.

Solution:

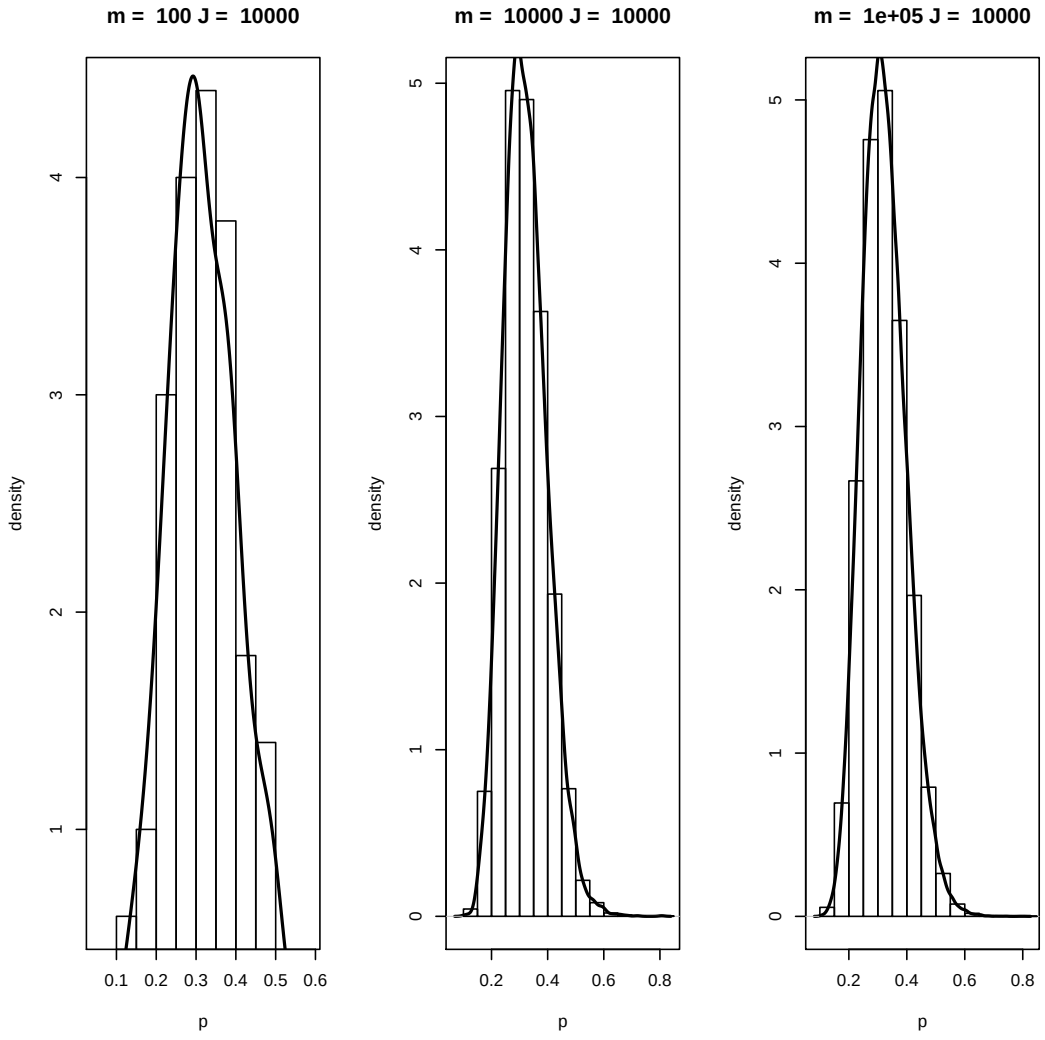
```
gibbs.f3 = function(x, p0, N0, m, J){
# Generates m samples of (p,N) values by the Gibbs sampler.
# In total J+m samples are generated
# but the first J are discarded.
# (p0,N0) is the initial value.
#
#  x is the data vector whose length is k
#  The length of N0 must be the same.
k <- length(x)
if(length(N0) != k){
print("The length of N0 must be equal to the length of x")
return
}
p.seq <-  rep(-1, J+m+1)
N.seq <- matrix(rep(-1,(k*(J+m+1))), nrow=(J+m+1),ncol=k)
p.seq[1] <- p0; N.seq[1,] <- N0
xdot <- sum(x)
for(j in 2:(J+m+1)) {
p.seq[j] <- rbeta(1, (xdot + 2),(sum(N.seq[j-1,]) - xdot + 4))
N.seq[j,] <- rpois(k, 16 * (1 - p.seq[j]))+x
+ xdot}
result <- list(p=p.seq[(J+2):(J+m+1)],N=N.seq[(J+2):(J+m+1),])
result
}
```

(c) Implement your code for data $x_1 = 5, x_2 = 4, x_3 = 6$ with a variety of burn in values and sample sizes after the burn in. Comment on the results using similar plots to lectures.

Solution:

There are no longer histograms or percentiles for different $x-$values, so just plot the histogram using three different parts of the output to get a feeling for dependence on sample size. Because the data is now fixed it is not necessary to simulate such a large number of values. Hence m has been taken to be 100000 rather than 1,000,000. The plots below show the histogram and density for the first 100, 1,000 and all 100,000 simulated values after the burn-in.

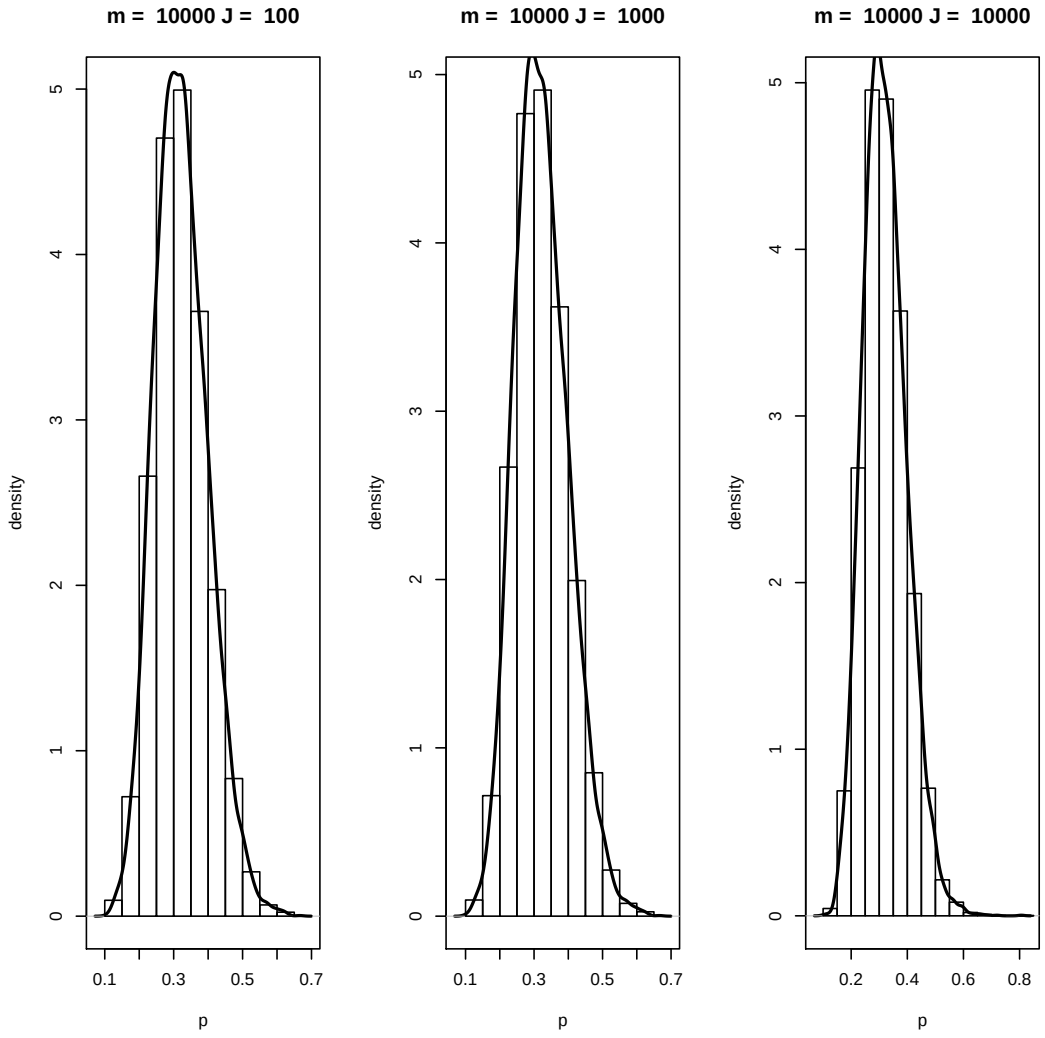
```
J=10000
m=100000
set.seed(456)
gibbsam3=gibbs.f3(c(5,4,6), 0.5, c(16,16,16), m, J)
par(mfrow=c(1,3))
# plot the three conditional densities with histograms
# selecting the right number of simulations
for (m in c(100,10000,100000))
{plot(density(gibbsam3$p[1:m]), xlab="p",
ylab="density", lwd=2,main=paste("m = ",m,"J = ",J),
ylim=range(hist(gibbsam3$p[1:m],plot=F)$density))
hist(gibbsam3$p[1:m], freq=F, add=T)}
```



It seems that the histograms and densities for 10,000 and 100,000 simulations are very similar. Hence for the experiments on burn-in only 10,000 simulations will be done. Three different burn-in periods of 100, 1000 and 10,000 will be taken.

```
m=10000
par(mfrow=c(1,3))
for (J in c(100,1000,10000))
{
set.seed(456)
gibbsam3=gibbs.f3(c(5,4,6), 0.5, c(16,16,16), m, J)
# plot the three conditional densities with histograms
```

```
# selecting the right number of simulations
   plot(density(gibbsam3$p), xlab="p",
ylab="density", lwd=2,main=paste("m = ",m,"J = ",J),
ylim=range(hist(gibbsam3$p,plot=F)$density))
hist(gibbsam3$p, freq=F, add=T)
}
```



It seems that the burn in of 10000 gives a slightly smoother density curve.

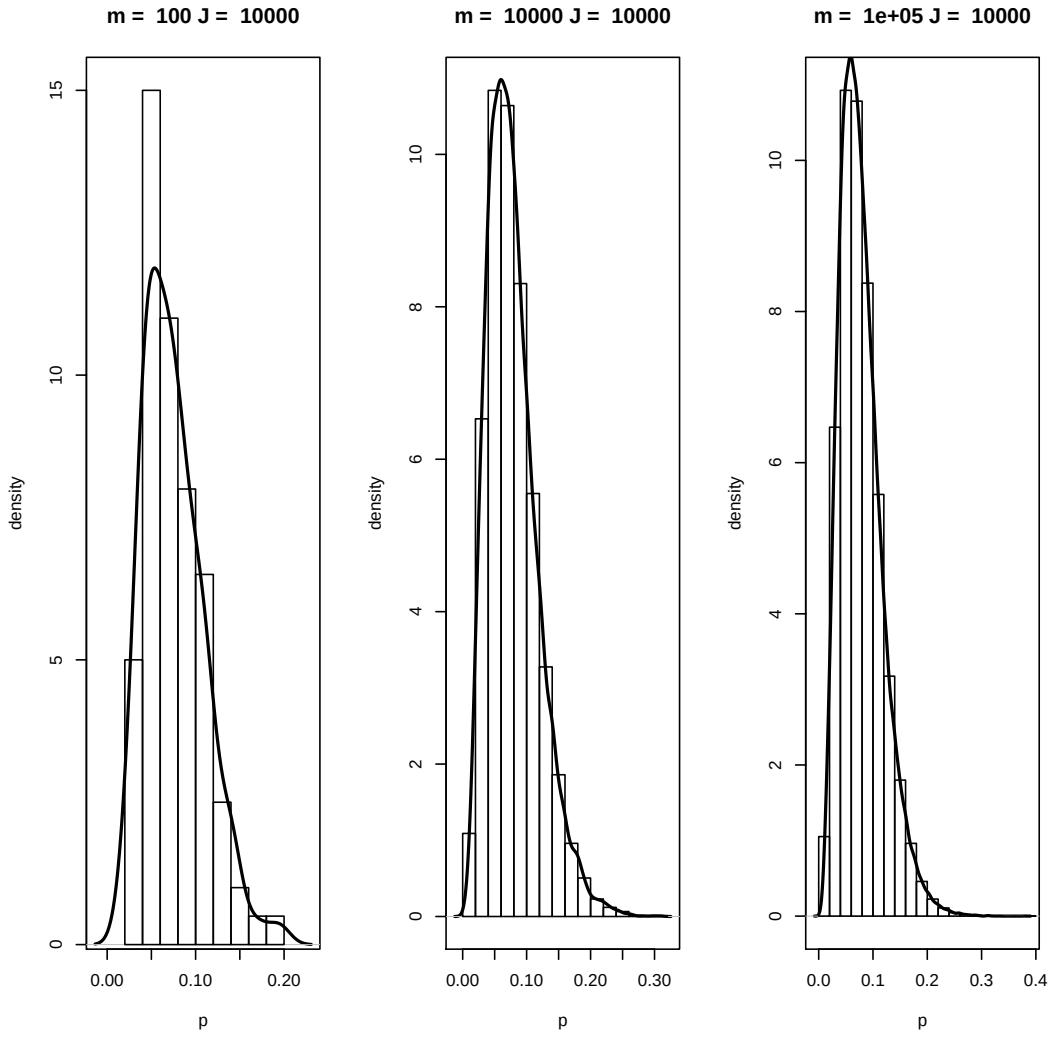
(d) Implement your code with 3 different x-values, say $x_1 = 1, x_2 = 0, x_3 = 1$ and $x_1 = 20, x_2 = 25, x_3 = 15$. How does this effect the answer to the previous question and the resulting plots. (Note that with $x_1 = 20, x_2 = 25, x_3 = 15$ you may need to alter the prior distribution for $\mathbf{N}$ because the joint density is only positive if $N_. > x_.$.)

Solution:

Repeating with data $x_1 = 1, x_2 = 0, x_3 = 1$

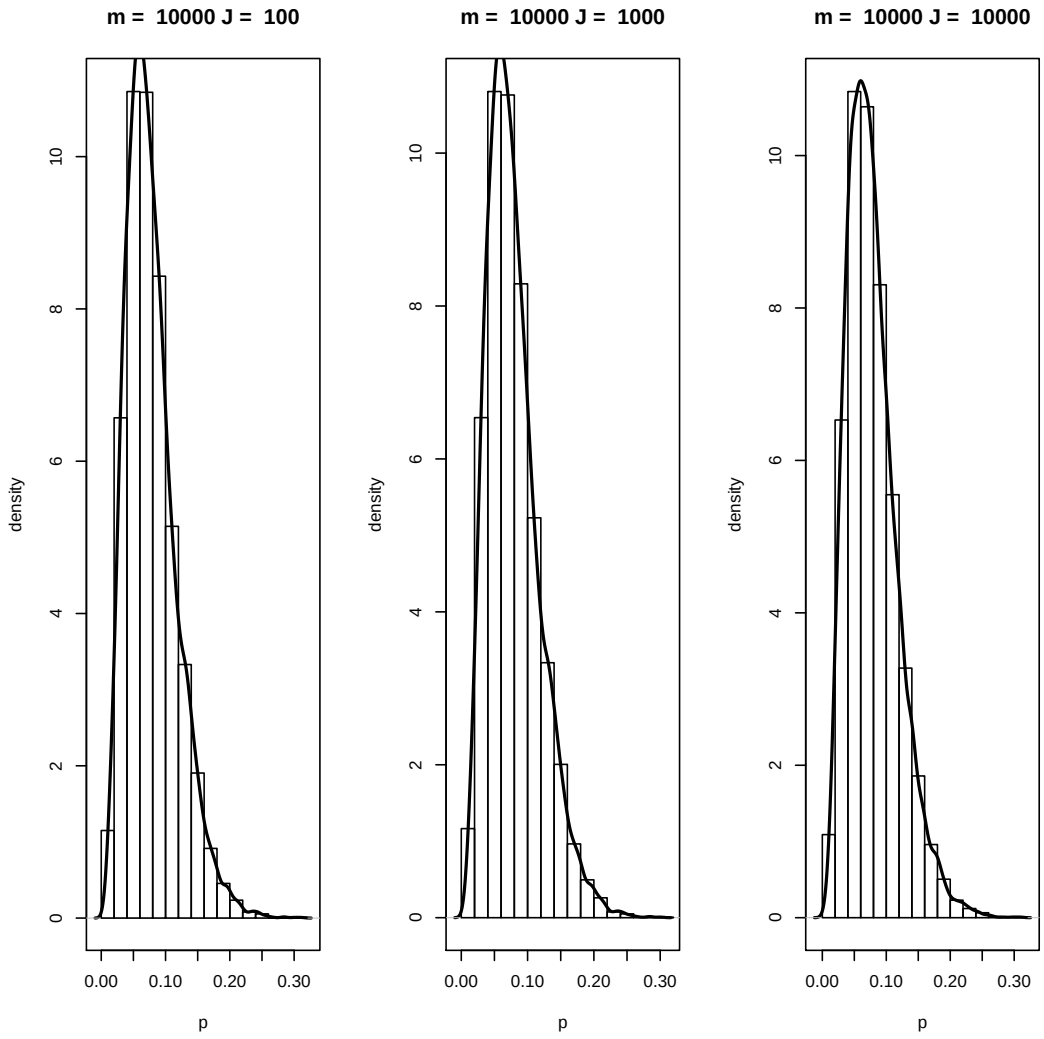
```
J=10000
m=100000
set.seed(456)
gibbsam3=gibbs.f3(c(1,0,1), 0.5, c(16,16,16), m, J)
par(mfrow=c(1,3))
# plot the three conditional densities with histograms
# selecting the right number of simulations
for (m in c(100,10000,100000))
```

```
{plot(density(gibbsam3$p[1:m]), xlab="p",
ylab="density", lwd=2,main=paste("m = ",m,"J = ",J),
ylim=range(hist(gibbsam3$p[1:m],plot=F)$density))
hist(gibbsam3$p[1:m], freq=F, add=T)}
```



The densities and histograms for p are now centred below 0.1. Again, it seems that the histograms and densities for 10,000 and 100,000 simulations are very similar. Again for the experiments on burn-in only 10,000 simulations will be done. Three different burn-in periods of 100, 1000 and 10,000 will be taken.

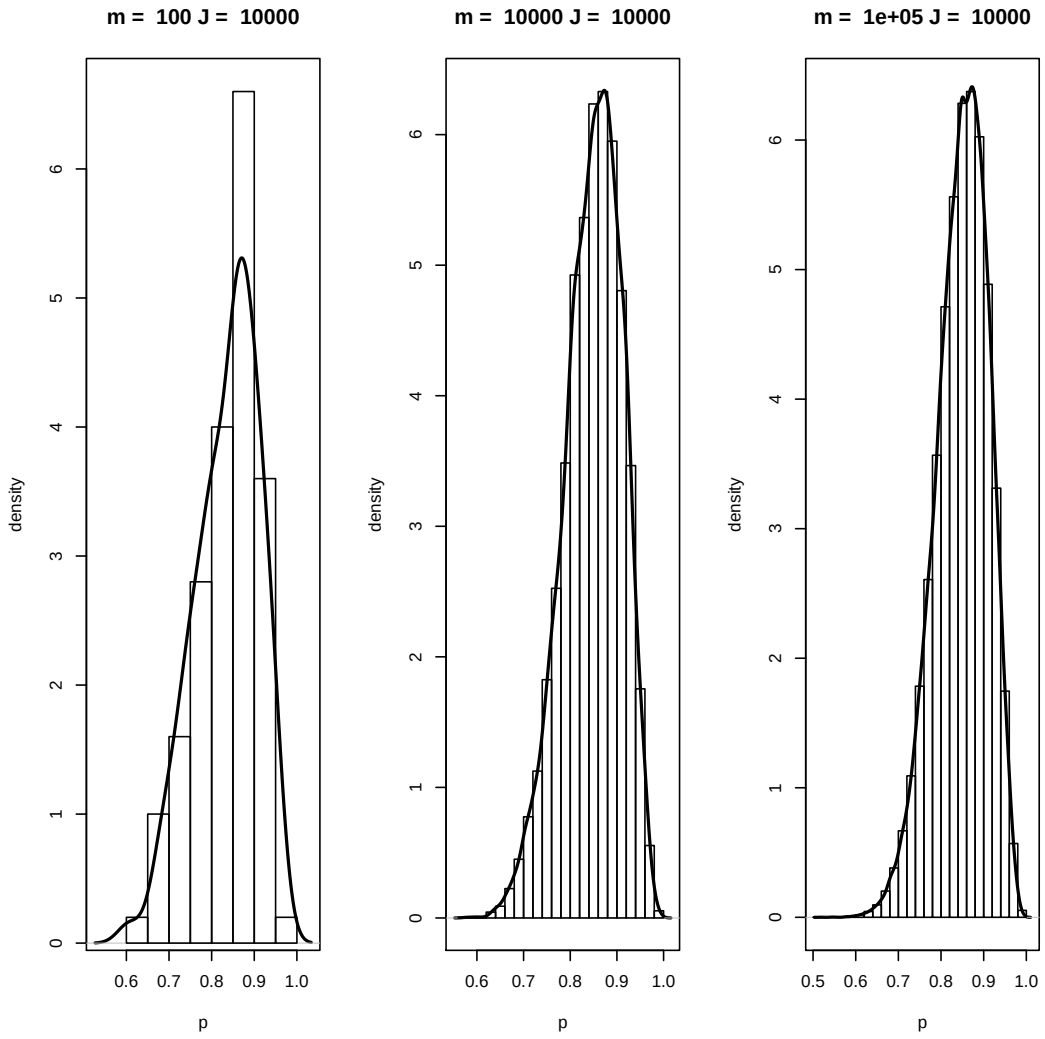
```
m=10000
par(mfrow=c(1,3))
for (J in c(100,1000,10000))
{
set.seed(456)
gibbsam3=gibbs.f3(c(1,0,1), 0.5, c(16,16,16), m, J)
# plot the three conditional densities with histograms
# selecting the right number of simulations
plot(density(gibbsam3$p), xlab="p",
ylab="density", lwd=2,main=paste("m = ",m,"J = ",J),
ylim=range(hist(gibbsam3$p,plot=F)$density))
hist(gibbsam3$p, freq=F, add=T)
}
```



Again it seems that the burn in of 10000 gives a slightly smoother density curve.

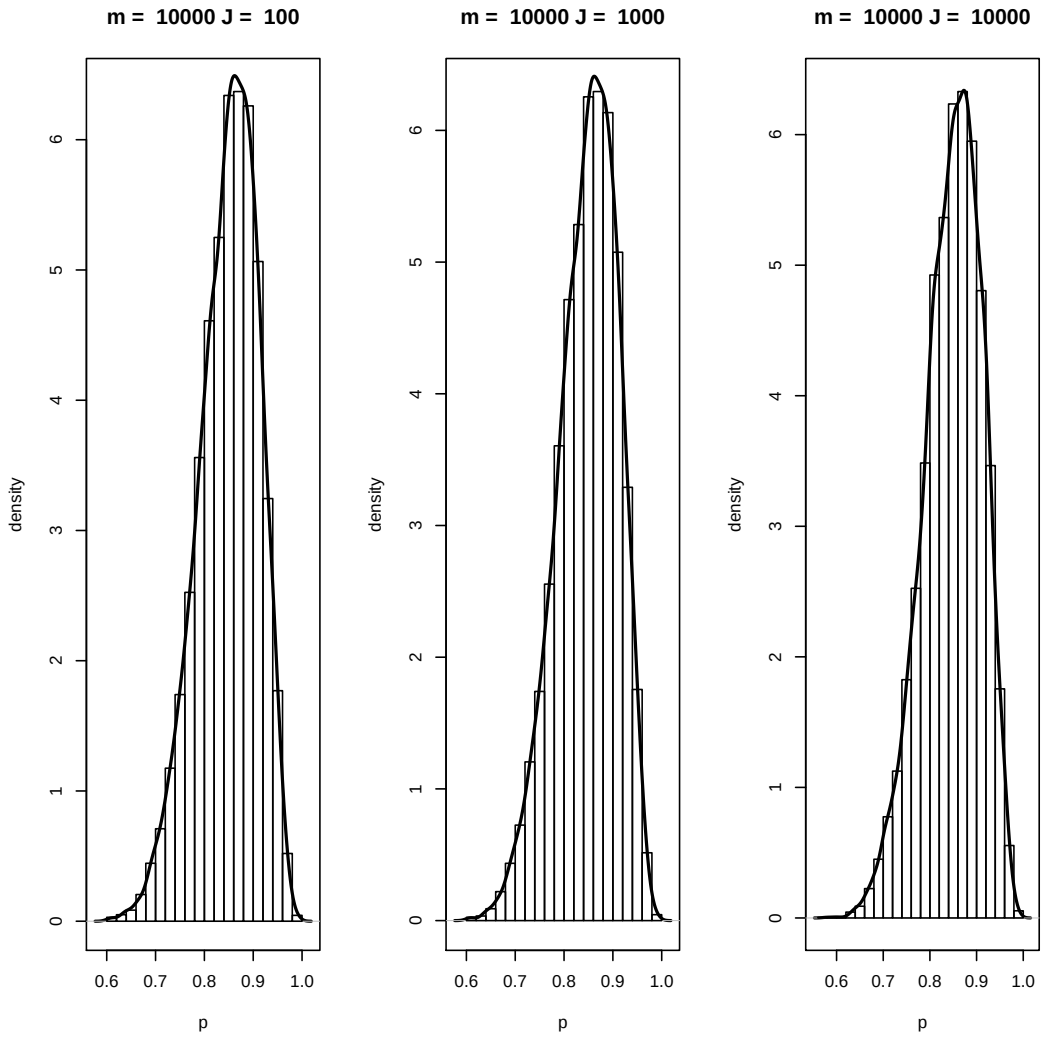
Repeating with data $x_1 = 20, x_2 = 25, x_3 = 15$ requires reconsideration of the Poisson parameters since $N.$ must be greater than $x.$ for the Beta simulation to be possible (and the joint density be positive). A Poisson with mean 50 is taken instead.

```
J=10000
m=100000
set.seed(456)
gibbsam3=gibbs.f3(c(20,25,15), 0.5, c(50,50,50), m, J)
par(mfrow=c(1,3))
# plot the three conditional densities with histograms
# selecting the right number of simulations
for (m in c(100,10000,100000))
{plot(density(gibbsam3$p[1:m]), xlab="p",
ylab="density", lwd=2,main=paste("m = ",m,"J = ",J),
ylim=range(hist(gibbsam3$p[1:m],plot=F)$density))
hist(gibbsam3$p[1:m], freq=F, add=T)}
```



The histograms and densities are now centred around 0.85. Again, it seems that the histograms and densities for 10,000 and 100,000 simulations are very similar. Again for the experiments on burn-in only 10,000 simulations will be done. Three different burn-in periods of 100, 1000 and 10,000 will be taken.

```
m=10000
par(mfrow=c(1,3))
for (J in c(100,1000,10000))
{
set.seed(456)
gibbsam3=gibbs.f3(c(20,25,15), 0.5, c(50,50,50), m, J)
# plot the three conditional densities with histograms
# selecting the right number of simulations
plot(density(gibbsam3$p), xlab="p",
ylab="density", lwd=2,main=paste("m = ",m,"J = ",J),
ylim=range(hist(gibbsam3$p,plot=F)$density))
hist(gibbsam3$p, freq=F, add=T)
}
```



Again it seems that the burn in of 10000 gives a slightly smoother density curve.

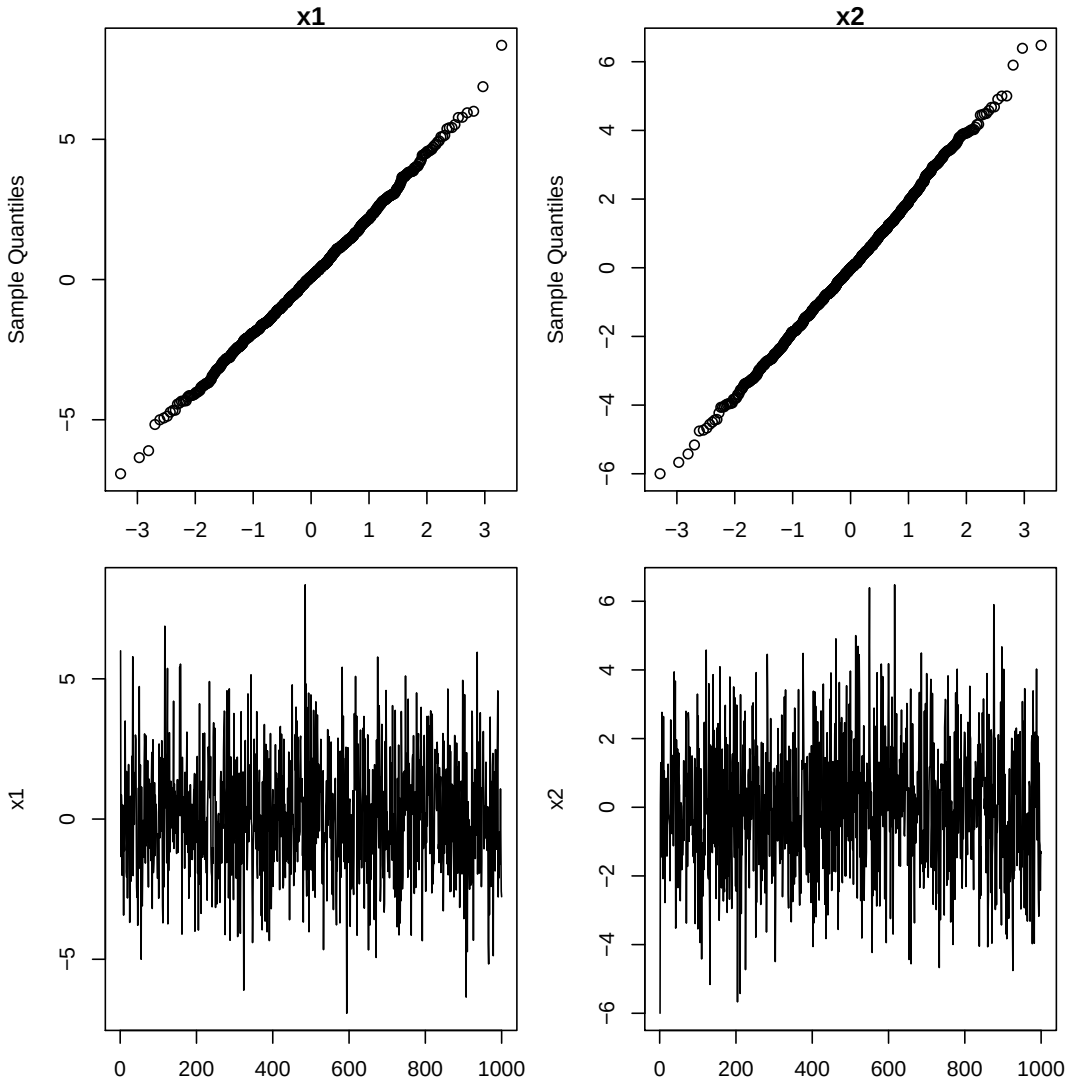
2. Suppose that $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$.

   (a) What is the conditional distribution of $X_1|X_2 = x_2$?
   Solution: Normal with mean $\mu_1 + (x_2 - \mu_2)\sigma_{12}/\sigma_2^2$ and variance $\sigma_1^2 - \sigma_{12}/\sigma_2^2$.

   (b) Write an R function that uses the Gibbs sampler to generate a sample of size $n = 1000$ from the $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\right)$ distribution.
   Plot traces of X_1 and X_2.
   Solution: To test the simulator we do a normal probability plot for each marginal, and both look good. The traces show pretty good mixing.

```r
set.seed(200)

# params
mu1 <- 0
mu2 <- 0
s11 <- 4
s12 <- 1
s22 <- 4
```

```r
# initial values
x1 <- 6
x2 <- -6

# sample size
nreps <- 1000
Gsamples <- matrix(nrow=nreps, ncol=2)
Gsamples[1,] <- c(x1, x2)

# main loop
for (i in 2:nreps) {
x1 <- rnorm(1, mu1 + (x2 - mu2)*s12/s22, sqrt(s11 - s12/s22))
x2 <- rnorm(1, mu2 + (x1 - mu1)*s12/s11, sqrt(s22 - s12/s11))
Gsamples[i,] <- c(x1, x2)
}

# output
par(mfrow=c(2,2), mar=c(2,4,1,1))
qqnorm(Gsamples[,1], main="x1")
qqnorm(Gsamples[,2], main="x2")
plot(Gsamples[,1], type="l", xlab="iteration", ylab="x1")
plot(Gsamples[,2], type="l", xlab="iteration", ylab="x2")
```



(c) Use your simulator to estimate $\mathbb{P}(X_1 \geq 0, X_2 \geq 0)$. To get a feel for the convergence rate, calculate the estimate using samples $\{1, \ldots, k\}$, for $k = 1, \ldots, n$, and then plot the estimates against n.

Solution: The plot appears after part (d).

```
par(mfrow=c(1,1))
success <- apply(Gsamples, 1, function(x) (x[1] > 0)&(x[2] > 0))
mean(success)
```

```
## [1] 0.296
```

```
plot(1:nreps, cumsum(success)/(1:nreps), type="l", xlab="k", ylab="prob", ylim=c(0,1))
```

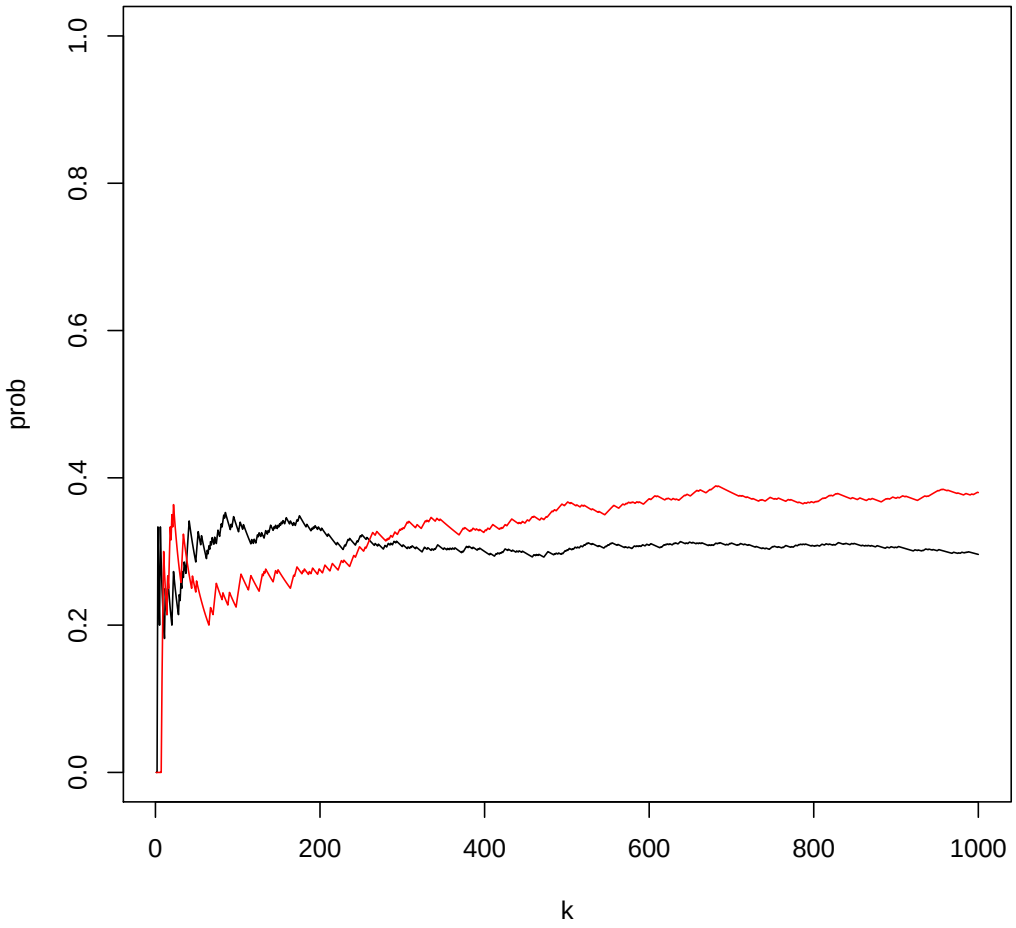
(d) Now change $\mathbf{\Sigma}$ to $\begin{pmatrix} 4 & 2.8 \\ 2.8 & 4 \end{pmatrix}$ and generate another sample of size 1000.

What do the traces/estimates look like now?

Solution: We put `s12 <- 2.8` then re-run the code above, getting a different `Gsamples`. We plot the cumlative estimates on top of the previous graph using `lines`. The cumulative estimates are more volatile in the second case, reflecting the stronger autocorrelation in the Markov chain, caused by the stronger correlation between X_1 and X_2.

```
success <- apply(Gsamples, 1, function(x) (x[1] > 0)&(x[2] > 0))
mean(success)
lines(1:nreps, cumsum(success)/(1:nreps), col="red")
```

```
## [1] 0.38
```



2 Workshop

3. Suppose $g(u)$ is a decreasing function of u. Let a random number X be generated by the following algorithm:

   1° Generate U from Unif$(0,1)$ and V from Unif$(0,1)$ independently.

   2° If $U+V<1$, then deliver $X=g(U)$; otherwise, go to 1°.

   Show that the distribution function of X is given by

$$F(x)=P(X\le x)=(1-g^{-1}(x))^2,\quad g(1)\le x\le g(0).$$

Solution: According to the algorithm, the cdf of X is

$$\begin{aligned} F(x) &= P(X \le x) \\ &= P(g(U) \le x | U+V<1) \\ &= P(U \ge g^{-1}(x) | U+V<1) \quad \text{(because } g(u) \text{ is decreasing)} \\ &= \frac{P(U \ge g^{-1}(x), U < 1-V)}{P(U+V<1)} \\ &= \frac{\int_0^{1-g^{-1}(x)} \int_{g^{-1}(x)}^{1-v} 1 du dv}{\int_0^1 \int_0^{1-v} 1 du dv} \\ &= \frac{\frac{1}{2}(1-g^{-1}(x))^2}{\frac{1}{2}} \;=\; (1-g^{-1}(x))^2. \end{aligned}$$

4. Consider a random sample $X_1, \cdots, X_n$ satisfying $X_i \sim \text{pois}(\theta)$. We assume a gamma prior pdf $\theta \sim \text{gamma}(\beta, \kappa)$ with known β and κ, i.e. $p(\theta) = \frac{1}{\beta^\kappa \Gamma(\kappa)} \theta^{\kappa-1} e^{-\theta/\beta}$; $\theta > 0$. Find the posterior distribution of θ and its mean.

   Solution:

   Note that in this question and the next one the gamma distribution has been parameterised using *scale* and shape parameters (rather than *rate* and shape). The joint pdf of $\mathbf{X} = (X_1, \cdots, X_n)$ and θ, for non-negative integers $\mathbf{x} = (x_1, \cdots, x_n)$, is

$$f(\mathbf{x}, \theta) = f(\mathbf{x}|\theta) p(\theta) \propto (\prod_{i=1}^n \theta^{x_i} e^{-\theta}) \theta^{\kappa-1} e^{-\theta/\beta} = \theta^{\sum_{i=1}^n x_i + \kappa - 1} \exp\{-(\beta^{-1}+n)\theta\}.$$

   The posterior pdf of θ given $\mathbf{x}$ is proportional to this as a function of θ, which on comparison with the Gamma density, is in turn proportional to a Gamma density with parameters $(\beta^{-1}+n)^{-1}$ and $\sum_{i=1}^n x_i + \kappa$. Hence, using the argument from Lectures, this is the posterior distribution. The mean of a Gamma distribution is the product of the two parameters, so the posterior mean is

$$\frac{\sum_{i=1}^n x_i + \kappa}{\beta^{-1} + \sum_{i=1}^n x_i} = \frac{\beta^{-1}}{\beta^{-1}+n} \kappa\beta + \frac{n}{\beta^{-1}+n} \bar{x}.$$

   This in turn is a weighted mean of the prior mean $\kappa\beta$ and the data mean $\bar{x}$ with weights proportional to the sample size and β^{-1}.

5. The *precision* of a normal distribution is the reciprocal of its variance. Show that the gamma distribution is conjugate for the normal distribution with a known mean and unknown precision. Identify the posterior parameters in terms of the data and the prior parameters.

   Solution:

   Suppose the precision $\tau = \sigma^{-2}$ has a gamma distribution with shape parameter α and rate parameter β so that the density, for $x > 0$, is:

$$f(\tau) = \frac{\beta^\alpha \exp[-\beta\tau] \tau^{\alpha-1}}{\Gamma(\alpha)}.$$

   Suppose the known mean of the normal distribution is μ. For data $x_1, \cdots, x_n$, the likelihood as a function of τ is proportional to

$$\tau^{n/2} \exp[-\tau/2 \sum_{i=1}^n (x_i - \mu)^2].$$

   Hence the posterior as a function of τ is proportional to

$$\tau^{\alpha+n/2-1} \exp[-(\beta + 1/2 \sum_{i=1}^n (x_i-\mu)^2)\tau]$$

   which in turn is proportional to a gamma distribution with shape parameter $\alpha + n/2$ and rate parameter $\beta + 1/2 \sum_{i=1}^n (x_i - \mu)^2$, showing that this is the posterior distribution (using the general trick).

6. The *normal-gamma* family of distributions for a random vector (M, R) has a marginal density for R which is gamma and the conditional density of M given $R = r$ as normal fixed mean and precision proportional to r. Show that the *normal-gamma* family is conjugate for normal data with unknown mean and precision. Identify the posterior parameters in terms of the data and the prior parameters. Interestingly, the marginal distribution the normal-gamma family is a shifted and scaled t-distribution whose degrees of freedom depend on the shape parameter of the gamma prior and the sample size, so the t-distribution is relevant to Bayesian as well as frequentist inference - but this is not part of this exercise.

Solution:

Suppose that given the precision τ and mean μ, the data are normally distributed with this mean and precision. Suppose that the prior distribution of τ is a gamma distribution with shape parameter α and rate parameter β, as in question 5. Further now suppose that given τ, the conditional prior density of μ is normal with mean θ and precision $r\tau$, for some θ and $r > 0$.

The joint prior density is therefore proportional to

$$\tau^{1/2} \exp[-r\tau/2(\mu - \theta)^2]\tau^{\alpha-1} \exp[-\beta\tau].$$

The analysis of variance identity says that the data $x_1, \cdots, x_n$ satisfies:

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n((\bar{x} - \mu)^2 + RSS/n), RSS = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Hence the likelihood is proportional, as a joint function of τ and μ, to

$$\tau^{n/2} \exp[-nr\tau/2(\bar{x} - \mu)^2].$$

The algebra for normal data with known precision and unknown mean can be made more precise by showing:

$$nr\tau(\bar{x} - \mu)^2 + r\tau(\theta - \mu)^2 = \tau\left((n + r)(\mu - \theta')^2 + \frac{nr(\bar{x} - \theta)^2}{n + r}\right),$$

where

$$\theta' = \frac{r\theta + n\bar{x}}{n + r}.$$

This gives the posterior density, as a joint function of τ and μ, is proportional to:

$$\left(\tau^{1/2} \exp[-(r + n)\tau/2(\mu - \theta')^2]\right)\left(\tau^{\alpha+n/2-1} \exp[-\beta'\tau]\right),$$

where

$$\beta' = \beta + RSS/2 + \frac{nr(\bar{x} - \theta)^2}{2(n + r)}.$$

This is of the required form for a normal-gamma family with the updated gamma parameters for τ as $\alpha + n/2, \beta'$ and the conditional density for μ given τ being normal with mean θ' and precision $(r + n)\tau$.

7. Let $X_1, \cdots, X_n$ be a random sample from an exponential distribution with pdf $f(x|\theta) = \theta e^{-\theta x}$, $x > 0$. (So the mean of X_i is $\frac{1}{\theta}$ instead of θ.) Assume the prior pdf of θ is $p(\theta) = \beta e^{-\beta\theta}$ where β is known. Find the posterior distribution for θ and its mean.

Solution:

The joint pdf of $\mathbf{X} = (X_1, \cdots, X_n)$ and θ is

$$f(\mathbf{x}, \theta) = f(\mathbf{x}|\theta)p(\theta) = (\prod_{i=1}^{n} \theta e^{-\theta x_i})\beta e^{-\beta\theta} = \beta\theta^n \exp\{-(\beta + \sum_{i=1}^{n} x_i)\theta\}.$$

The posterior pdf of θ given $\mathbf{x}$ is proportional to this as a function of θ, which on comparison with the Gamma density (given in the previous question), is in turn proportional to a Gamma density with parameters $(\beta + \sum_{i=1}^{n} x_i)^{-1}$ and $n + 1$. Hence, using the argument from Lectures, this is the posterior distribution. The mean of a Gamma distribution is the product of the two parameters, so the posterior mean is

$$\frac{n + 1}{\beta + \sum_{i=1}^{n} x_i} = \frac{1}{\frac{1}{n+1}\beta + \frac{n}{n+1}\bar{x}}.$$