In-Class Activity #12: Computational Bayesian Techniques

Solutions

Instructions

1. Download this `.Rmd` file from Moodle and open it in RStudio.
2. Update the author name and date.
3. Answer the questions below, using LaTeX within R Markdown to type up your answers.
4. When you finish, `Knit to PDF` and turn in that final PDF on Moodle.
5. Submit your final **PDF** to Moodle.

Note: Most of this activity will involve working in R. However, for any questions involving mathematical derivations may write your answers to those questions on the whiteboard and submit a picture of your written work.

Direct Sampling

Suppose we have data from a normal distribution with unknown mean μ but known variance σ^2,

$$Y \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right),$$

and we put the following prior on the mean μ:

$$\mu \sim N(m, v).$$

1. What is the posterior distribution for μ? *Hint: look at Activity 10.*

ANSWER: As we proved in class last week, the posterior for μ is also Normal:

$$\mu \mid y \sim N\left(\frac{yv + m\sigma^2}{\sigma^2 + v}, \left(\frac{1}{\sigma^2} + \frac{1}{v}\right)^{-1}\right)$$

2. Since, as you should have found in Question 1, the posterior distribution for μ is recognizably Normally distributed, we can find closed-form expressions for the posterior mean, median, and mode. What are they?

ANSWER: Since a Normal distribution is symmetric, the posterior mean, median, and mode are all the same:

$$\frac{yv + m\sigma^2}{\sigma^2 + v}$$

3. We could also use a computer to generate samples $\mu^{(0)}, \mu^{(1)}, \ldots, \mu^{(N)}$ from our posterior distribution, and use those samples to estimate the posterior mean (e.g., $\hat{E}(\mu \mid y) = \frac{1}{N}\sum_{i=1}^{N} \mu^{(i)}$), median, and mode. Do this now using $N = 1000$ samples. Assume that $m = 0, v = 1, \sigma^2 = 1$ and $y = 2$. How do your estimates compare to what you derived in Question 2?

ANSWER:

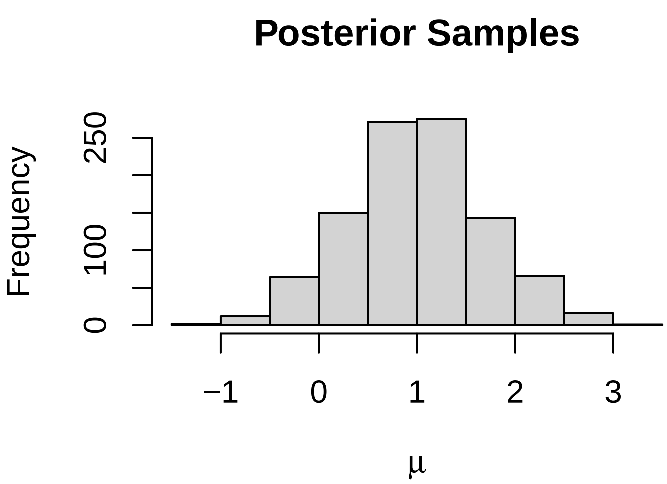
```
### draw 1000 samples from the posterior
### hint: use the rnorm function
# set up
N <- 1000
m <- 0
v <- 1
sig2 <- 1
y <- 2

# sample from the posterior
set.seed(455)
mus <- rnorm(n = N, mean = (y*v + m*sig2)/(v + sig2), sd = (1/sig2 + 1/v)^(-1/2))

### estimate the mean, median, and mode of the posterior
# estimate posterior mean
mean(mus)
```

```
## [1] 1.006331
```

```
# estimate posterior median
median(mus)
```

```
## [1] 1.001694
```

```
# plot a histogram of the samples to look at mode
hist(mus, xlab = expression(mu), main = 'Posterior Samples')
```



We can also tell by looking at the histogram that the posterior mode is near 1. Note that these estimates are all consistent with what we derived above:

$$\frac{yv + m\sigma^2}{\sigma^2 + v} = \frac{2 \cdot 1 + 0 \cdot 1}{1 + 1} = \frac{2}{2} = 1$$

4. We can also use these samples from our posterior to get a 90% credible interval for μ (i.e,. an interval within which there is a 90% chance μ falls), and to estimate the probability that $\mu > 2$. Note that Bayesian credible intervals and posterior probabilities actually have the type of probability interpretation that many people wish they could make with confidence intervals and p-values!

ANSWER:

```
### get 90% credible interval
### hint: use the quantile function
quantile(mus, probs = c(0.05, 0.95))
```

```
##         5%        95%
## -0.2021053  2.1704267
```

```
### estimate P(mu > 2 | y)
mean(mus > 2)
```

```
## [1] 0.083
```

5. Suppose we wanted to know the posterior distribution for μ^2 instead of μ. This would be hard to derive by hand, but is very easy to assess using the posterior samples we generated for μ above:

ANSWER:

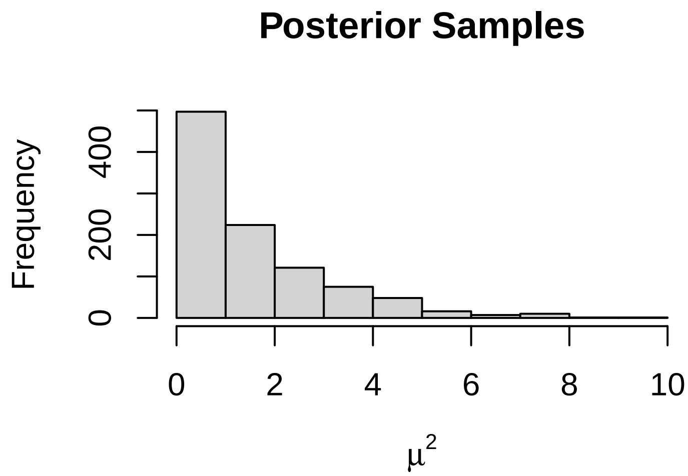
```
### get posterior samples for mu^2
### hint: use the samples for mu
musq <- mus^2

### estimate the mean, median, and mode of the posterior for mu^2
# posterior mean
mean(musq)
```

```
## [1] 1.50819
```

```
# posterior median
median(musq)
```

```
## [1] 1.004854
```

```
# histogram to look at mode
hist(musq, xlab = expression(mu^2), main = 'Posterior Samples')
```



Gibbs Sampling

Now suppose we have data from a normal distribution where both the mean **and** variance are unknown. For convenience, we'll parameterize this model in terms of the *precision* $\gamma = \frac{1}{\sigma^2}$ instead of the variance σ^2.

$$Y \mid \mu, \gamma \sim N\left(\mu, \frac{1}{\gamma}\right)$$

Suppose we put the following *independent* priors on the mean μ and precision γ:

$$\mu \sim N(m, v)$$

$$\gamma \sim \text{Gamma}(a, b)$$

1. Write down the joint posterior distribution for μ, γ. Does this look like a recognizable probability distribution?

ANSWER: No, this is not a recognizable posterior:

$$\begin{aligned}
g(\mu, \gamma \mid y) &\propto f(y \mid \mu, \gamma) f(\mu, \gamma) \\
&= f(y \mid \mu, \gamma) f(\mu) f(\gamma), \text{ since } \mu, \gamma \text{ independent} \\
&= \left[(2\pi)^{-\frac{1}{2}} \gamma^{\frac{1}{2}} e^{-\frac{1}{2}\gamma(y-\mu)^2}\right] \left[(2\pi v)^{-\frac{1}{2}} e^{-\frac{1}{2v}(\mu-m)^2}\right] \left[\frac{b^a}{\Gamma(a)} \gamma^{a-1} e^{-b\gamma}\right] \\
&\propto \gamma^{\frac{1}{2}} e^{-\frac{1}{2}\gamma(y-\mu)^2} e^{-\frac{1}{2v}(\mu-m)^2} \gamma^{a-1} e^{-b\gamma} \\
&= \gamma^{\frac{1}{2}+a-1} e^{-\frac{1}{2}\gamma(y-\mu)^2 + -\frac{1}{2v}(\mu-m)^2 - b\gamma} \\
&= \gamma^{\frac{1}{2}+a-1} e^{-\frac{1}{2}\left[\gamma(y-\mu)^2 + \frac{1}{v}(\mu-m)^2 + 2b\gamma\right]} \\
&= \gamma^{\frac{1}{2}+a-1} e^{-\frac{1}{2}\left[\gamma y^2 - 2\mu y\gamma + \gamma\mu^2 + \mu^2/v - 2m\mu/v + m^2/v + 2b\gamma\right]} \\
&= \gamma^{\frac{1}{2}+a-1} e^{-\frac{1}{2}\left[\gamma(y^2+2b) - 2\mu(y\gamma + m/v) + \mu^2(\gamma + 1/v) + m^2/v\right]} \\
&\propto \gamma^{\frac{1}{2}+a-1} e^{-\frac{1}{2}\left[\gamma(y^2+2b) - 2\mu(y\gamma + \frac{m}{v}) + \mu^2(\frac{1}{v} + \gamma)\right]}
\end{aligned}$$

You should have answered "no" to Question 1, meaning that we can't use our usual techniques here to find Bayes estimators for μ or γ since we don't have a recognizable posterior distribution. Instead, we'll use a computational technique known as *Gibbs Sampling* to generate samples from this posterior distribution. Gibbs Sampling is particularly useful when we have more than one parameter, and the basic idea involves reducing our problem to a series of calculations involving one parameter at a time. In order to perform Gibbs Sampling, we need to find the conditional distributions

$$g(\mu \mid y, \gamma) \propto f(y \mid \mu, \gamma) f(\mu)$$

$$g(\gamma \mid y, \mu) \propto f(y \mid \mu, \gamma) f(\gamma)$$

We will use these conditional distributions to sample from the joint posterior $g(\mu, \gamma \mid y)$ according to the following algorithm:

(1) Start with initial values $\mu^{(0)}, \gamma^{(0)}$.

(2) Sample $\mu^{(t+1)} \sim g(\mu \mid y, \gamma = \gamma^{(t)})$.

(3) Sample $\gamma^{(t+1)} \sim g(\gamma \mid y, \mu = \mu^{(t+1)})$.

(4) Repeat many times.

It turns out that the resulting $\mu^{(0)}, \mu^{(1)}, \dots, \mu^{(N)}$ and $\gamma^{(0)}, \gamma^{(1)}, \dots, \gamma^{(N)}$ are samples from the joint posterior distribution $g(\mu, \gamma \mid Y)$, and we can use these sampled values to estimate quantities such as the posterior mean of each parameter $\hat{E}(\mu \mid y) = \frac{1}{N} \sum_{i=1}^{N} \mu^{(i)}$, $\hat{E}(\gamma \mid y) = \frac{1}{N} \sum_{i=1}^{N} \gamma^{(i)}$. Note that in practice we typically remove the initial iterations, known as the "burn-in" period: e.g., $\hat{E}(\mu \mid y) = \frac{1}{N-B} \sum_{i=B}^{N} \mu^{(i)}$.

2. Show that the conditional distributions $g(\mu \mid y, \gamma), g(\gamma \mid y, \mu)$ are proportional to $f(y \mid \mu, \gamma) f(\mu), f(y \mid \mu, \gamma) f(\gamma)$, respectively, as stated above.

ANSWER:

$$\begin{aligned} g(\mu \mid y, \gamma) &= \frac{f(\mu, y, \gamma)}{f(y, \gamma)} \\ &\propto f(\mu, y, \gamma), \text{ since } f(y, \gamma) \text{ doesn't depend on } \mu \\ &= f(y \mid \mu, \gamma) f(\mu, \gamma) \\ &= f(y \mid \mu, \gamma) f(\mu) f(\gamma), \text{ since } \mu, \gamma \text{ independent} \\ &\propto f(y \mid \mu, \gamma) f(\mu), \text{ since} f(\gamma) \text{ doesn't depend on } \mu \end{aligned}$$

A similar argument can be used to show $g(\gamma \mid y, \mu) \propto f(y|\mu, \gamma) f(\gamma)$.

3. Use this result to show that $\mu \mid y, \gamma \sim N\left(\frac{y\gamma + \frac{m}{v}}{\gamma + \frac{1}{v}}, \left[\gamma + \frac{1}{v}\right]^{-1}\right)$ and $\gamma \mid y, \mu \sim \text{Gamma}\left(\frac{1}{2} + a, \frac{1}{2}(y - \mu)^2 + b\right)$.

ANSWER:

$$
\begin{aligned}
g(\mu \mid y, \gamma) &\propto f(y \mid \mu, \gamma) f(\mu) \\
&= \left[(2\pi)^{-\frac{1}{2}} \gamma^{\frac{1}{2}} e^{-\frac{1}{2}\gamma(y-\mu)^2}\right] \left[(2\pi v)^{-\frac{1}{2}} e^{-\frac{1}{2v}(\mu-m)^2}\right] \\
&\propto e^{-\frac{1}{2}\gamma(y-\mu)^2 - \frac{1}{2v}(\mu-m)^2} \\
&= e^{-\frac{1}{2}\gamma(y^2-2\mu y+\mu^2) - \frac{1}{2v}(\mu^2-2\mu m+m^2)} \\
&\propto e^{-\frac{1}{2}\gamma(-2\mu y+\mu^2) - \frac{1}{2v}(\mu^2-2\mu m)} \\
&= e^{-\frac{1}{2}\left[\mu^2(\gamma+\frac{1}{v}) - 2\mu(y\gamma+\frac{m}{v})\right]} \\
&= e^{-\frac{1}{2}(\gamma+\frac{1}{v})\left[\mu^2 - 2\mu\left(\frac{y\gamma+\frac{m}{v}}{\gamma+\frac{1}{v}}\right)\right]} \\
&\propto e^{-\frac{1}{2}(\gamma+\frac{1}{v})\left[\mu^2 - 2\mu\left(\frac{y\gamma+\frac{m}{v}}{\gamma+\frac{1}{v}}\right) + \left(\frac{y\gamma+\frac{m}{v}}{\gamma+\frac{1}{v}}\right)^2\right]} \\
&=\propto e^{-\frac{1}{2\left(\gamma+\frac{1}{v}\right)^{-1}}\left[\mu - \left(\frac{y\gamma+\frac{m}{v}}{\gamma+\frac{1}{v}}\right)\right]^2}
\end{aligned}
$$

$$
\implies \mu \mid y, \gamma \sim N\left(\frac{y\gamma+\frac{m}{v}}{\gamma+\frac{1}{v}}, \left[\gamma+\frac{1}{v}\right]^{-1}\right)
$$

$$
\begin{aligned}
g(\gamma \mid y, \mu) &\propto f(y \mid \mu, \gamma) f(\gamma) \\
&= \left[(2\pi)^{-\frac{1}{2}} \gamma^{\frac{1}{2}} e^{-\frac{1}{2}\gamma(y-\mu)^2}\right] \left[\frac{b^a}{\Gamma(a)} \gamma^{a-1} e^{-b\gamma}\right] \\
&\propto \gamma^{\frac{1}{2}} \gamma^{a-1} e^{-\frac{1}{2}\gamma(y-\mu)^2} e^{-b\gamma} \\
&= \gamma^{\frac{1}{2}+a-1} e^{-\frac{1}{2}\gamma(y-\mu)^2 - b\gamma} \\
&= \gamma^{\frac{1}{2}+a-1} e^{-\gamma\left(\frac{1}{2}(y-\mu)^2+b\right)}
\end{aligned}
$$

$$
\implies \gamma \mid y, \mu \sim \text{Gamma}\left(\frac{1}{2}+a, \frac{1}{2}(y-\mu)^2+b\right)
$$

4. Suppose that we choose the following hyperparameters for our prior distributions—$m = 0, v = 1, a = 1, b = 1$—and that we observe $y = 2$. Write code to implement this Gibbs Sampler.

ANSWER:

```r
# set up priors
m <- 0
v <- 1
a <- 1
b <- 1

# set up data
y <- 2

# choose starting values by randomly sampling from our priors
# (this is just one possible way to choose starting values)
# (it's also useful to try out a few different starting values)
set.seed(1)
mu <- rnorm(1, mean = m, sd = sqrt(v))
gam <- rgamma(1, shape = a, rate = b)
```

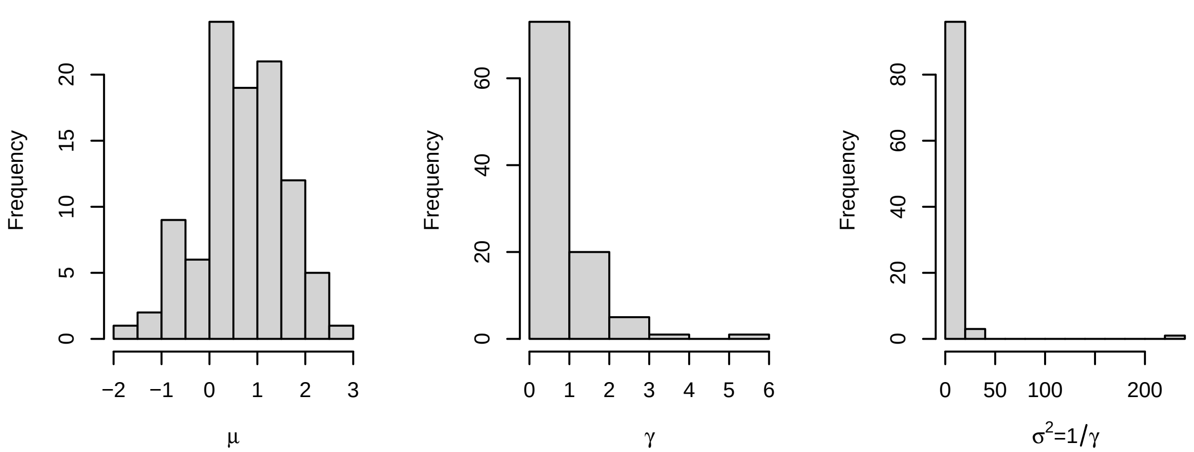
```r
# set up empty vectors to store samples
mus <- c()
gams <- c()

# store starting values in vectors of samples
mus[1] <- mu
gams[1] <- gam
```

```r
# choose number of iterations
# (we'll start with 100, but in practice you'd choose something much bigger)
N <- 100

# run through Gibbs Sampling for a total of N iterations
for(i in 2:N){
  # update mu
  m1 <- y*gam + m/v
  m2 <- gam + 1/v
  mu <- rnorm(n = 1, mean = (m1)/(m2), sd = sqrt(1/m2))

  # update gamma
  g1 <- 0.5 + a
  g2 <- 0.5*(y-mu)^2 + b
  gam <- rgamma(n = 1, shape = g1, rate = g2)

  # store new samples
  mus[i] <- mu
  gams[i] <- gam
}
```

5. Look at a histogram of your posterior samples for μ, γ and $\sigma^2 = \frac{1}{\gamma}$.

ANSWER:

```r
par(mfrow=c(1,3))
hist(mus, xlab = expression(mu), main = '')
hist(gams, xlab = expression(gamma), main = '')
hist(1/gams, xlab = expression(paste(sigma^2,'=',1/gamma)), main = '')
```



6. Estimate the posterior mean and median of μ.

ANSWER:

```
# posterior mean
mean(mus)
```

```
## [1] 0.6840056
```

```
# posterior median
median(mus)
```

```
## [1] 0.6525553
```

7. Find a 90% credible interval for μ, and estimate the probability that $\mu > 2$.

ANSWER:

```
# 90% credible interval
quantile(mus, probs = c(0.05, 0.95))
```

```
##         5%        95%
## -0.8862182  2.0094145
```

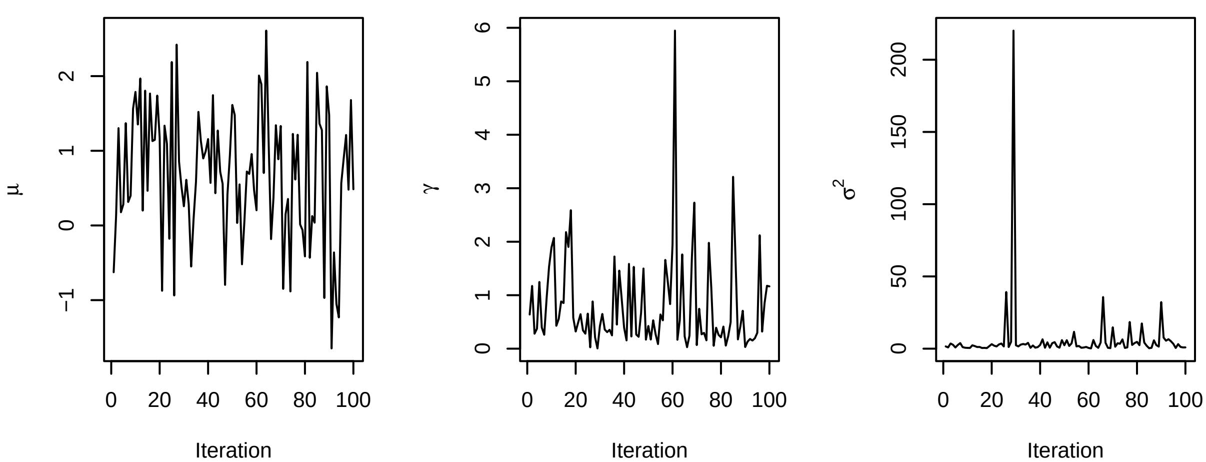
```
# P(mu > 2 | y)
mean(mus > 2)
```

```
## [1] 0.06
```

8. Create a *trace plot* showing the behavior of the samples over the N iterations.

ANSWER:

```
iterations <- 1:N

par(mfrow=c(1,3))
plot(mus ~ iterations, xlab = 'Iteration', ylab = expression(mu), type = 'l')
plot(gams ~ iterations, xlab = 'Iteration', ylab = expression(gamma), type = 'l')
plot(1/gams ~ iterations, xlab = 'Iteration', ylab = expression(sigma^2), type = 'l')
```



9. As mentioned above, in practice we usually pick a burn-in period of initial iterations to remove. This decision is often motivated by the fact that, depending on your choice of starting value, it may take awhile for your chain of samples to look like it is "mixing" well. Play around with your choice of starting value above to see if you can find situations in which a burn-in period might be helpful.

10. Want to see an interesting application of Gibbs Sampling that applies to my research in Statistical Genetics? Read the paper by Jonathan Pritchard et al., describing their program *STRUCTURE*, a Bayesian approach that is widely used for inferring ancestry from genetic data! See if you can derive the conditional distributions yourself. (Link Here)