Gibbs Sampler

May 1, 2018

So far, we have demonstrated MCMC for a single parameter. What if we seek the posterior distribution of multiple parameters, and that posterior distribution does not have a standard form? One option is to perform Metropolis-Hastings (M-H) by sampling candidates for all parameters at once, and accepting or rejecting all of those candidates together. While this is possible, it can get complicated. Another (simpler) option is to sample the parameters one at a time.

As a simple example, suppose we have a joint posterior distribution for two parameters θ and ϕ, written $p(\theta, \phi \mid y) \propto g(\theta, \phi)$. If we knew the value of ϕ, then we would just draw a candidate for θ and use $g(\theta, \phi)$ to compute our Metropolis-Hastings ratio, and possibly accept the candidate. Before moving on to the next iteration, if we don't know ϕ, then we can perform a similar update for it. Draw a candidate for ϕ using some proposal distribution and again use $g(\theta, \phi)$ to compute our Metropolis-Hastings ratio. Here we pretend we know the value of θ by substituting its current iteration from the Markov chain. Once we've drawn for both θ and ϕ, that completes one iteration and we begin the next iteration by drawing a new θ. In other words, we're just going back and forth, updating the parameters one at a time, plugging the current value of the other parameter into $g(\theta, \phi)$.

This idea of one-at-a-time updates is used in what we call Gibbs sampling, which also produces a stationary Markov chain (whose stationary distribution is the posterior). If you recall, this is the namesake of JAGS, "just another Gibbs sampler."

1 Full conditional distributions

Before describing the full Gibbs sampling algorithm, there's one more thing we can do. Using the chain rule of probability, we have:

$$p(\theta, \phi \mid y) = p(\theta \mid \phi, y) \cdot p(\phi \mid y)$$

Notice that the only difference between $p(\theta, \phi \mid y)$ and $p(\theta \mid \phi, y)$ is multiplication by a factor that **doesn't involve** θ. Since the $g(\theta, \phi)$ function above, when viewed as a function of θ is proportional to both these expressions, we might as well have replaced it with $p(\theta \mid \phi, y)$ in our update for θ.

This distribution $p(\theta \mid \phi, y)$ is called the **full conditional distribution** for θ. Why use it instead of $g(\theta, \phi)$? In some cases, the full conditional distribution is a standard distribution we know how to sample. If that happens, we no longer need to draw a candidate and decide whether to accept it. In fact, if we treat the full conditional distribution as a candidate proposal distribution, the resulting Metropolis-Hastings acceptance probability becomes exactly 1.

Gibbs samplers require a little more work up front because you need to find the full conditional distribution for each parameter. The good news is that all full conditional distributions have the same starting point: the full joint posterior distribution. Using the example above, we have

$$p(\theta \mid \phi, y) \propto p(\theta, \phi \mid y) \propto g(\theta, \phi)$$

where we simply now treat ϕ as a known number. Likewise, the other full conditional is

$$p(\phi \mid \theta, y) \propto p(\theta, \phi \mid y) \propto g(\theta, \phi)$$

where here, we consider θ to be a known number.

We always start with the full posterior distribution. Thus, **the process of finding full conditional distributions is the same as finding the posterior distribution of each parameter, pretending that all other parameters are known.**

2 Gibbs Sampling Algorithm

The Gibbs sampler sequentially samples from the collection of full (or complete) conditional distributions $p(\theta_i|\theta_{j\neq i}, y), i = 1, ..., k$, and it does, under fairly broad conditions, produce a Markov chain with the joint posterior density $p(\theta|y)$ as its stationary distribution. The algorithm was named by Geman and Geman (1984);Gelfand and Smith (1990) showed how the method could be applied to a wide variety of Bayesian inference problems.

The idea of Gibbs sampling is that we can update multiple parameters by sampling just one parameter at a time, cycling through all parameters and repeating. To perform the update for one particular parameter, we substitute in the current values of all other parameters.

Here is the algorithm. Suppose we have a joint posterior distribution for two parameters θ and ϕ, written $p(\theta, \phi \mid y)$. If we can find the distribution of each parameter at a time, i.e., $p(\theta \mid \phi, y)$ and $p(\phi \mid \theta, y)$, then we can take turns sampling these distributions like so:

1) Initalize ϕ_0 and θ_0

2) for $i = 1, ..., m$

- Using ϕ_{i-1}, draw θ_i from $p(\theta \mid \phi = \phi_{i-1}, y)$.
- Using θ_i, draw ϕ_i from $p(\phi \mid \theta = \theta_i, y)$.

Together, steps 1 and 2 complete one cycle of the Gibbs sampler and produce the draw for (θ_i, ϕ_i) in one iteration of a MCMC sampler. If there are more than two parameters, we can handle that also. One Gibbs cycle would include an update for each of the parameters.

3 Normal likelihood, unknown mean and variance

Let's return to the model we have gone through a few docs ago at "Non-conjugate models":

$$y_i|\mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2) i = i, ..., n$$
$$\mu|\sigma^2 \sim N(\mu_0, \sigma_0^2)$$
$$\sigma^2 \sim IG(\nu, \beta_0)$$

We chose a normal prior for μ because, in the case where σ^2 is known, the normal is the conjugate prior for μ. Likewise, in the case where μ is known, the inverse-gamma is the conjugate prior for σ^2. This will give us convenient full conditional distributions in a Gibbs sampler.

Let's first work out the form of the full posterior distribution. When we begin analyzing data, the JAGS software will complete this step for us. However, it is extremely valuable to see and understand how this works.

$$p(\mu, \sigma 2 \mid y1, y2, \ldots, yn) \propto p(y1, y2, \ldots, yn \mid \mu, \sigma^2) p(\mu) p(\sigma^2)$$
$$= \prod_{i=1}^{n} N(y_i \mid \mu, \sigma^2) \times N(\mu \mid \mu_0, \sigma_0^2) \times IG(\sigma^2 \mid \nu_0, \beta_0)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right] \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[\frac{\mu - \mu_0)^2}{2\sigma_0^2}\right] \times \frac{\beta_0^{\nu_0}}{\Gamma(\nu_0)} (\sigma^2)^{-(\nu 0+1)} \exp\left[-\frac{\beta_0}{\sigma^2}\right] I_{\sigma^2 > 0}(\sigma^2)$$

We drop some pieces that do not include μ and σ^2 and we get:

$$\propto (\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2}\right] \exp\left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right] (\sigma^2)^{-(\nu_0+1)} \exp\left[-\frac{\beta_0}{\sigma^2}\right] I_{\sigma^2 > 0}(\sigma^2)$$

From here, it is easy to continue on to find the two full conditional distributions we need. First let's look at μ, assuming σ^2 **is known** (in which case it becomes a constant and is absorbed into the normalizing constant):

$$p(\mu \mid \sigma^2, y_1, \ldots, y_n) \propto p(\mu, \sigma^2 \mid y_1, \ldots, y_n)$$
$$\propto \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2}\right] \exp\left[-dfrac(\mu - \mu_0)^2 2\sigma_0^2\right]$$
$$\propto \exp\left[-\frac{1}{2}\left(\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2} + \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)\right]$$

After having conducted some further calculations:

$$\propto N\left(\mu \mid \frac{n\bar{y}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}, \frac{1}{n/\sigma^2 + 1/\sigma_0^2}\right)$$

,

So, given the data and σ^2, μ follows this normal distribution.

Now let's look at σ^2**, assuming μ is known:**

$$p(\mu \mid \sigma^2, y_1, \ldots, y_n) \propto p(\mu, \sigma^2 \mid y_1, \ldots, y_n)$$
$$\propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2}\right] (\sigma^2)^{-(\nu_0+1)} \exp\left[-\frac{\beta_0}{\sigma^2}\right] I_{\sigma^2 > 0}(\sigma^2)$$
$$\propto (\sigma^2)^{-(\frac{n}{2} + \nu_0 + 1)} \exp\left[-\frac{1}{\sigma^2}\left(\beta_0 + \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2}\right)\right] I_{\sigma^2 > 0}(\sigma^2)$$
$$\propto IG\left(\sigma^2 \mid \nu_0 + \frac{n}{2}, \beta_0 + \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2}\right)$$

,

These two distributions provide the basis of a Gibbs sampler to simulate from a Markov chain whose stationary distribution is the full posterior of both μ and σ^2. We simply alternate draws between these two parameters, using the most recent draw of one parameter to update the other.

4 Gibbs sampler in R

To implement the Gibbs sampler we just described, let's return to our running example where the data are the percent change in total personnel from last year to this year for $n = 10$ companies. We'll still use a normal likelihood, but now we'll relax the assumption that we know the variance of growth between companies, σ^2, and estimate that variance. Instead of the tt prior from earlier, we will use the conditionally conjugate priors, normal for $\mu\mu$ and inverse-gamma for σ^2.

The first step will be to write functions to simulate from the full conditional distributions we derived in the previous segment. The full conditional for μ, given σ^2 and data is:

$$N(\mu \mid \frac{n\bar{y}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}, \frac{1}{n/\sigma^2 + 1/\sigma_0^2})$$

```
In [5]: update_mu = function(n, ybar, sig2, mu_0, sig2_0) {
          sig2_1 = 1.0 / (n / sig2 + 1.0 / sig2_0)
          mu_1 = sig2_1 * (n * ybar / sig2 + mu_0 / sig2_0)
          rnorm(n=1, mean=mu_1, sd=sqrt(sig2_1))
        }
```

The full conditional for σ^2 given $\mu\mu$ and data is:

$$IG(\sigma^2 \mid \nu_0 + \frac{n}{2}, \beta_0 + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2})$$

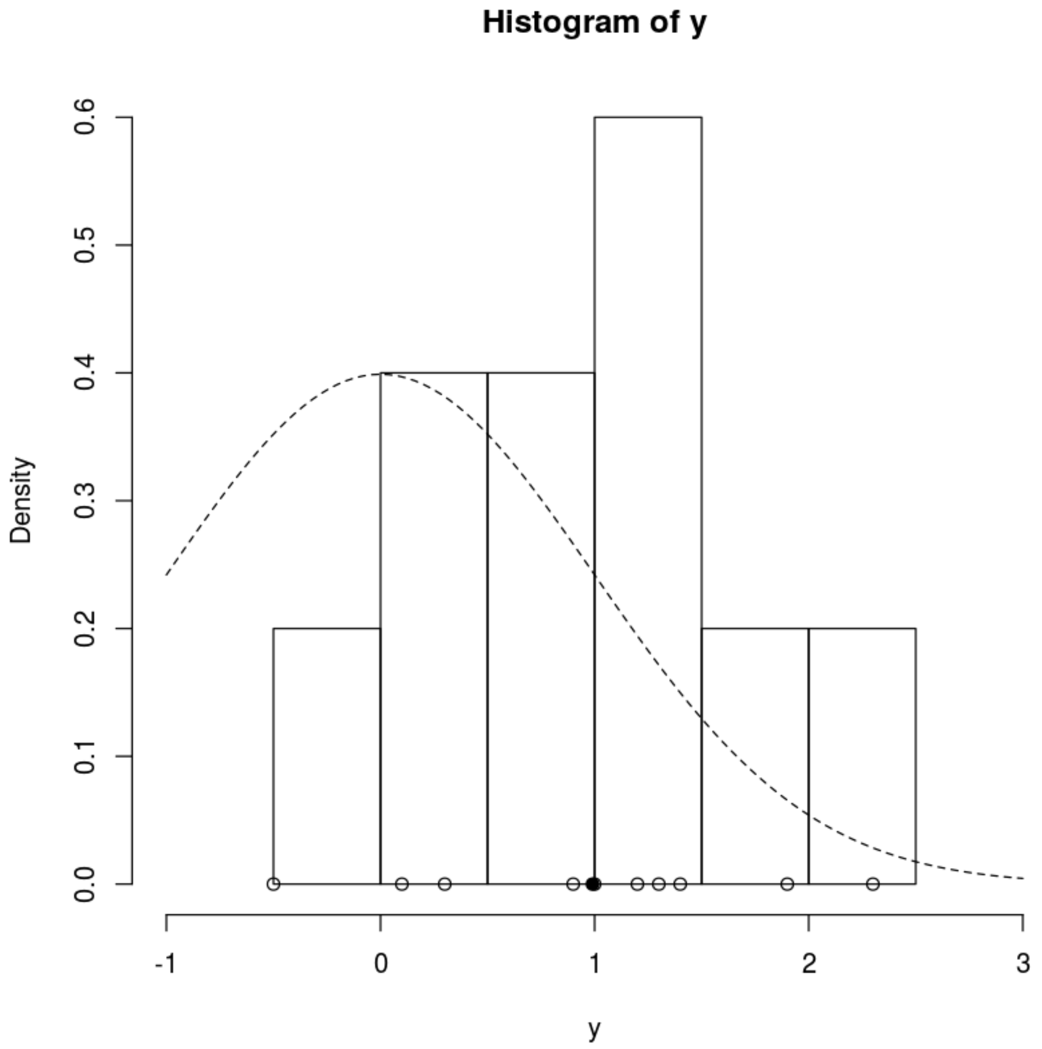
```
In [6]: update_sig2 = function(n, y, mu, nu_0, beta_0) {
          nu_1 = nu_0 + n / 2.0
          sumsq = sum( (y - mu)^2 ) # vectorized
          beta_1 = beta_0 + sumsq / 2.0
          out_gamma = rgamma(n=1, shape=nu_1, rate=beta_1) # rate for gamma is shape for inv-gam
          1.0 / out_gamma # reciprocal of a gamma random variable is distributed inv-gamma
        }
```

With functions for drawing from the full conditionals, we are ready to write a function to perform Gibbs sampling.

```
In [7]: gibbs = function(y, n_iter, init, prior) {
          ybar = mean(y)
          n = length(y)

          ## initialize
          mu_out = numeric(n_iter)
          sig2_out = numeric(n_iter)

          mu_now = init$mu

          ## Gibbs sampler
          for (i in 1:n_iter) {
            sig2_now = update_sig2(n=n, y=y, mu=mu_now, nu_0=prior$nu_0, beta_0=prior$beta_0)
            mu_now = update_mu(n=n, ybar=ybar, sig2=sig2_now, mu_0=prior$mu_0, sig2_0=prior$sig2
            sig2_out[i] = sig2_now
            mu_out[i] = mu_now
          }

          cbind(mu=mu_out, sig2=sig2_out)
        }
```

Now we are ready to set up the problem in R.

```
In [8]: y = c(1.2, 1.4, -0.5, 0.3, 0.9, 2.3, 1.0, 0.1, 1.3, 1.9)
        ybar = mean(y)
        n = length(y)

        ## prior
        prior = list()
        prior$mu_0 = 0.0
        prior$sig2_0 = 1.0
        prior$n_0 = 2.0 # prior effective sample size for sig2
        prior$s2_0 = 1.0 # prior point estimate for sig2
        prior$nu_0 = prior$n_0 / 2.0 # prior parameter for inverse-gamma
        prior$beta_0 = prior$n_0 * prior$s2_0 / 2.0 # prior parameter for inverse-gamma

        hist(y, freq=FALSE, xlim=c(-1.0, 3.0)) # histogram of the data
        curve(dnorm(x=x, mean=prior$mu_0, sd=sqrt(prior$sig2_0)), lty=2, add=TRUE) # prior for m
        points(y, rep(0,n), pch=1) # individual data points
        points(ybar, 0, pch=19) # sample mean
```

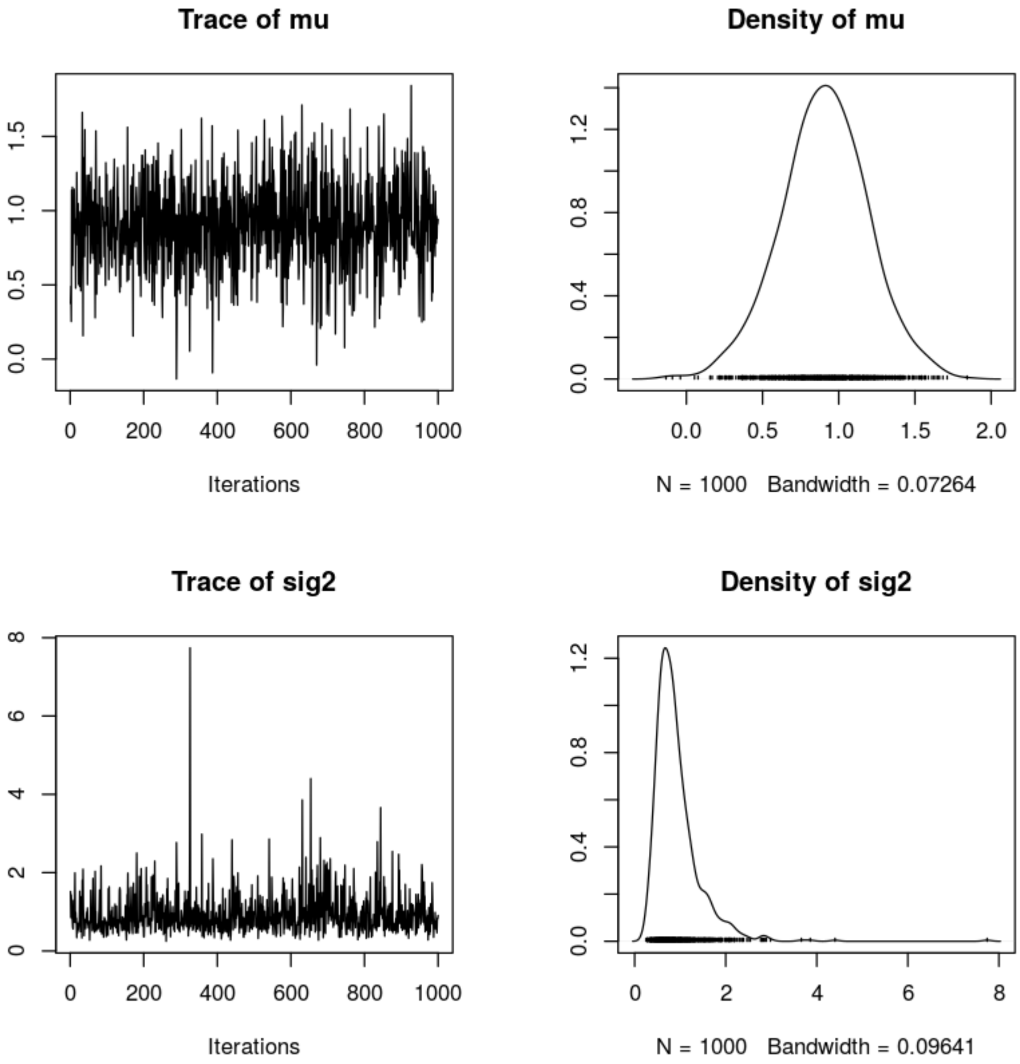


Finally, we can initialize and run the sampler!

```
In [9]: set.seed(53)

        init = list()
        init$mu = 0.0

        post = gibbs(y=y, n_iter=1e3, init=init, prior=prior)

        head(post)
```

mu	sig2
0.3746992	1.5179144
0.4900277	0.8532821
0.2536817	1.4325174
1.1378504	1.2337821
1.0016641	0.8409815
1.1576873	0.7926196

```
In [10]: library("coda")
         plot(as.mcmc(post))
```



```
In [11]: summary(as.mcmc(post))
```

```
Iterations = 1:1000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

       Mean     SD Naive SE Time-series SE
mu   0.9051 0.2868  0.00907        0.00907
sig2 0.9282 0.5177  0.01637        0.01810

2. Quantiles for each variable:

       2.5%    25%    50%   75% 97.5%
mu   0.3024 0.7244 0.9089 1.090 1.481
sig2 0.3577 0.6084 0.8188 1.094 2.141
```

5 Review

Question 1

Which of the following descriptions matches the process of Gibbs sampling for multiple random variables?

a. Draw candidates for all J variables simultaneously using a multivariate proposal distribution. For each variable, calculate the acceptance ratio αj using the joint (unnormalized) density. Accept each candidate with probability min{1,αj} for j=1,. . . ,J. Repeat this cycle for many iterations.

b. Draw candidates for all variables simultaneously using a multivariate proposal distribution. Calculate the acceptance ratio α using the joint (unnormalized) density. Accept the candidates with probability min{1,α}. Repeat this step for many iterations.

c. Cycle through the variables, drawing from a proposal distribution for each variable and accepting the candidate with probability equal to the ratio of the candidate draw to the old value of the variable. Repeat this cycle for many iterations.

d. Cycle through the variables, drawing a sample from the full conditional distribution of each variable while substituting in the current values of all other variables. Repeat this cycle for many iterations.

Answer 1

d.

Gibbs sampling allows us to perform the updates one-at-a-time using full conditional distributions.

** Question 2**

Suppose we have a joint probability distribution for four variables, p(w,x,y,z). Which of the following expresses the full conditional distribution for variable x?

a. $p(x \mid y)$

b. $p(x \mid w, y, z)$

c. $p(x)$

d. $p(w, y, z \mid x)$

Answer 2

b.

It is the distribution of x, conditional on all other variables. It is proportional to p(w,x,y,z), where we consider w,y, and z as fixed constants.

Question 3

3. Suppose we have the following joint distribution for x,y, and z:

$$p(x,y,z) = 5e^{-5z} I_{z\geq 0} \frac{\Gamma(z+3)}{\Gamma(z)\Gamma(3)} y^{z-1}(1-y)^2 I_{0<y<1} \binom{10}{x} y^x (1-y)^{10-x} I_{x\in\{1,\ldots,10\}}.$$

The density for the full conditional distribution of z is proportional to which of the following?

Hint: The full conditional for z is proportional to the full joint distribution $p(x,y,z)$ where x and y are just constants.

a. $p(z \mid x, y) \propto 5e^{-5z} I_{z\geq 0}$

b. $p(z \mid x, y) \propto e^{-5z\frac{\Gamma(z+3)}{\Gamma(z)}} y^{z-1} I_{z\geq 0}$

c. $p(z \mid x, y) \propto \binom{10}{x} y^x (1-y)^{10-x} I_{x\in 1,\ldots,10}$

d. $p(z \mid x, y) \propto y^{z-1}(1-y)^2 y^x (1-y)^{10-x} I_{0<y<1}$

Answer 3

b.

Because we consider as variable only z and the rest are constants.

Question 4

The full conditional distribution in Question 3 is not a standard distribution that we can easily sample. Fortunately, it turns out that we can do a Metropolis-Hastings step inside our Gibbs sampler step for z.

If we employ that strategy in a Gibbs sampler for y and z (always conditioning on x), then the algorithm would look like this:

For iteration i in 1 to m, repeat:

1. a) Draw z* from a proposal distribution q.

b) Calculate the acceptance ratio (alpha) using the full conditional distribution for z | x,y and the candidate distribution q, plugging in the previous iteration's value y_{i-1} for y.

c) Accept the candidate with probability min{1,alpha} and set the value for z_i accordingly.

2. ____________________________.

end.
What would go in step 2 to complete the Gibbs sampler?

a. Draw yi from the full conditional p(y|x,z), plugging in the candidate z∗ for z.

b. Draw yi from the marginal distribution p(y).

c. Draw yi from the full conditional p(y|x,z), plugging in the value zi just drawn in step 1 for z.

d. Draw yi from the full conditional p(y|x,z), plugging in the previous iteration's value zi−1 for z.

Answer 4

c.

This is just the usual Gibbs step for y.
Question 5
For Questions 6 to 8, consider the example from the lesson where the data are percent change in total personnel since last year for $n = 10$ companies.

In our model with normal likelihood and unknown mean μ and unknown variance σ^2, we chose a normal prior for the mean and an inverse-gamma prior for the variance. What was the major advantage of selecting these particular priors?

a. Because these priors are conjugate for their respective parameters, they guarantee the smallest possible Monte Carlo standard error for posterior mean estimates.

b. These priors allowed us to bypass MCMC, providing a joint conjugate posterior for μ and σ^2.

c. Each prior was conjugate in the case where the other parameter was known, causing the full conditional distributions to come from the same distribution families as the priors (and therefore easy to sample).

d. Because these priors are conjugate for their respective parameters, they guarantee the most accurate posterior distribution possible for the given likelihood.

Answer 5

c.

In hierarchical models, selecting conjugate priors at any level will result in a simple Gibbs update for the parameter involved.

Question 6 Suppose we repeat the analysis for n=6 companies in another industry and the data are:

```
In [12]: y = c(-0.2, -1.5, -5.3, 0.3, -0.8, -2.2)
```

Re-run the Gibbs sampler in R for these new data (5000 iterations using the same priors and initial values as in the Lesson) and report the posterior mean for μ. Round your answer to two decimal places.

```
In [18]: post = gibbs(y=y, n_iter=5e3, init=init, prior=prior)
         round(post[5000,1],2)
```

mu: -1.05

```
In [14]: head(post)
```

mu	sig2
0.09519318	3.588773
-0.91893502	4.974997
-1.31098864	2.047403
-0.27033688	7.362965
-1.84699337	18.198883
-1.25042204	2.194644

Question 7 An industry expert is surprised by your results from Question 7 and insists that growth in this sector should be positive on average. To accommodate this expert's prior beliefs, you adjust the prior for μ to be normal with a mean 1.0 and variance 1.0. This is a fairly informative and optimistic prior (the prior probability that μ>0 is about 0.84).

What happens to the posterior mean of μ? Re-run the analysis on the new data with this new prior. Again, use 5000 iterations and the same prior for σ2 and initial values as before).

a. The posterior mean for μ is less than -0.25, suggesting that despite the optimistic prior, the data strongly favor estimating growth to be negative in this industry.

b. The posterior mean for μ is between -0.25 and 0.25, suggesting that the data are not as optimistic about growth as the prior, but we are inconclusive about whether growth is positive or negative.

c. The posterior mean for μ is between 0.25 and 1.0, suggesting that the data are not informative enough to contradict this expert's opinion.

d. The posterior mean for μ is above 1.0, suggesting that the optimistic prior was actually not optimistic enough.

Answer 7

```
In [19]: init$mu = 1.0
         post = gibbs(y=y, n_iter=5e3, init=init, prior=prior)
         round(post[5000,1],2)
```

mu: -0.71
Answer 7

a.

```
In [ ]:
```