MATH 3190 Homework 6

Focus: Notes 8

Due March 30, 2024

Your homework should be completed in R Markdown or Quarto and Knitted to an html or pdf document. You will "turn in" this homework by uploading to your GitHub Math_3190_Assignment repository in the Homework directory.

Some of the parts in problems 1 and 2 require writing down some math-heavy expressions. You may either type it up using LaTeX style formatting in R Markdown, or you can write it by hand (neatly) and include pictures or scans of your work in your R Markdown document.

Problem 1 (10 points)

Three airlines serve a small town in Ohio. Airline A has 52% of all scheduled flights, airline B has 35% and airline C has the remaining 13%. Their on-time rates are 85%, 67%, and 41%, respectively. A flight just left on-time. What is the probability that it was a flight of airline A?

Problem 2 (13 points)

Suppose we have a data set with each observation x_i independent and identically exponentially distributed for $i = 1, 2, \ldots, n$. That is, $x_i \sim \text{Exp}(\lambda)$ where λ is the rate parameter. We would like to find a posterior (or at least a function proportional to it) for λ.

Part a (5 points)

Write down the likelihood function (or a function proportional to it) in this situation. We would call this $p(x|\lambda)$.

Part b (5 points)

Now let λ have a normal prior with mean 0.1 and variance 1: $\lambda \sim N(1/10, 1)$. Use this and the likelihood from part a to write down a function that is proportional to the posterior of λ given $\boldsymbol{x}$. We call this $p(\lambda|\boldsymbol{x})$.

Part c (3 points)

Which would be more appropriate here to obtain samples of λ, the Gibbs or Metropolis algorithm? Explain why. You may want to look on page 8 of Notes 8 in the conjugate prior table.

Problem 3 (26 points)

Suppose we have the vector `x = c(1.83, 1.72, 2.13, 2.49, 0.90, 2.01, 1.51, 3.12, 1.29, 1.54, 2.94, 3.02, 0.93, 2.78)` that we believe comes from a gamma distribution with shape of 10 and some rate β: $x_i \sim \text{Gam}(10, \beta)$. We will use sampling to obtain some information about β. Let's put a gamma prior on β with a shape of α_0 and a rate of 1: $\beta \sim \text{Gam}(\alpha_0, 1)$.

Part a (5 points)

Use the fact that this is a conjugate prior to write down what kind of distribution the posterior of β, which is $p(\beta|\boldsymbol{x})$, is.

Part b (5 points)

Let $\alpha_0 = 1$. In an **R** code chunk, sample 10,000 β values from the distribution you wrote down in part a using the `rgamma()` function and report the 95% credible interval for β using the 2.5th and 97.5th percentiles.

Part c (3 points)

Repeat part b with $\alpha_0 = 10$.

Part d (3 points)

Repeat part b with $\alpha_0 = 100$.

Part e (7 points)

Now suppose we have twice as much data (given in the **R** code chunk below). Repeat parts b, c, and d using this `x` vector instead and report the three 95% credible intervals. Note, this new vector x will change the shape and rate parameters used in the `rgamma()` functions.

```r
x <- c(1.83, 1.72, 2.13, 2.49, 0.90, 2.01, 1.51, 3.12, 1.29, 1.54,
       2.94, 3.02, 0.93, 2.78, 2.76, 1.70, 1.42, 2.16, 1.07, 2.21,
       2.38, 2.27, 1.72, 1.44, 1.54, 1.72, 1.87, 1.39)
```

Part f (3 points)

In this problem, the true β value is 5. Write a sentence or two about the effect adding more data has to these credible intervals by comparing the intervals from parts b-d to the intervals from part e.

Problem 4 (51 points)

Let's apply the Bayesian framework to a regression problem. In the GitHub data folder, there is a file called `treeseeds.txt` that contains information about species of tree, the count of seeds it produces, and the average weight of those seeds in mg.

Part a (3 points)

Read in the `treeseeds.txt` file and take the log of the counts and weights. Fit an OLS regression model using log(weight) to predict log(count).

Part b (15 points)

We will walk through the mathematics of obtaining the posterior together here since this problem will focus on coding the Metropolis algorithm. Assuming the true errors are normal with mean 0 and variance σ^2, $\epsilon_i \sim N(0, \sigma^2)$, it can be shown that each y_i has the distribution

$$p(y_i|x_i, \beta_0, \beta_1\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

So, we can write the likelihood is

$$p(y_i|x_i, \beta_0, \beta_1\sigma^2) = \propto \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

where y_i is the log(count) for observation i and x_i is the log(weight) for observation i. Note that here we think of $\boldsymbol{y}$ as being random and $\boldsymbol{x}$ as being fixed. We could, in theory, think of the vector $\boldsymbol{x}$ as also being random and put a prior on it. But we won't do that here.

Now, let's just put uniform priors on β_0 and β_1 so the priors are proportional to 1. Also, let's assume $\sigma^2 = 1$. This seems reasonable since s_e^2, the MSE, is 0.877. Of course, we could put a prior on σ^2 as well and sample it too, but we will focus on only sampling β_0 and β_1.

Now, with those uniform priors, and plugging in 1 for σ^2, we have that the joint posterior of β_0 and β_1 is:

$$p(\beta_0, \beta_1 | \boldsymbol{x}, \boldsymbol{y}) =\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) = f(\beta_0, \beta_1 | \boldsymbol{x}, \boldsymbol{y}).$$

Then, we can take the log to get

$$\ln(f(\beta_0, \beta_1 | \boldsymbol{x}, \boldsymbol{y})) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Our goal now is to obtain samples of β_0 and β_1. Let's use the Metropolis algorithm to do this. Using the log of the function proportional to the joint posterior of β_0 and β_1, $\ln(f(\beta_0, \beta_1 | \boldsymbol{x}, \boldsymbol{y}))$, write a Metropolis algorithm in **R**. For β_0, you can use a normal proposal distribution centered at the previous value, $\beta_0^{(i)}$, with a standard deviation of 0.8 and for β_1, you can use a normal proposal distribution centered at the previous value, $\beta_1^{(i)}$, with a standard deviation of 0.1. The starting values don't matter too much, but we can use $\beta_0^{(0)} = 10$ and $\beta_1^{(0)} = -0.5$. It may be useful to look at the `Notes 8 Script.R` file that is on GitHub in the Notes 8 folder and is on Canvas.

Obtain at least 10,000 samples (set a seed, please) and plot the chains for β_0 and β_1. For this problem, include:

1. The plot for the β_0 chain.
2. The plot for the β_1 chain.
3. The 95% credible interval for β_0 based on the 2.5th and 97.5th percentiles.
4. The 95% credible interval for β_1 based on the 2.5th and 97.5th percentiles.

Part c (3 points)

Based on the plots of the chains from part b, does it look like the Metropolis sampling worked fairly well?

Part d (4 points)

Interpret both of the credible intervals from part b.

Part e (5 points)

Find and report the integrated autocorrelation time for the β_0 and β_1 chains. Each chain will have their own $\hat{\tau}_{int}$ value, so you should report two (although they will be similar).

Part f (3 points)

Based on the integrated autocorrelation time for the β_0 and β_1 chains, how many MCMC samples would you need to generate to get the equivalent of 10,000 independent samples?

Part g (3 points)

Let's compare these credible intervals to some other intervals. First, obtain the 95% t confidence intervals for β_0 and β_1 just using the `confint()` function and report them here.

Part h (10 points)

Now let's obtain confidence intervals using bootstrapping in a similar way we did with regularization in Notes 7 and HW 4 (this is known as bootstrapping the cases). Set a seed and then using at least 10,000 bootstrap samples, report the 95% percentile confidence intervals for β_0 and β_1 using the `quantile()` function on the values of β_0 and β_1 that you obtained in the bootstrap.

Part i (5 points)

Write a couple sentences comparing all of the intervals in parts b, g, and h.