STAT 289 Homework 3

Wenjing Xu

For Questions 1-7, you will need the SAT coaching dataset given in Table 1 below. This dataset comes from a randomized experiment to test the effects of 8 different SAT coaching programs. For each program i, an observed effect y_i was measured as well as a standard error σ_i of the treatment effect. We want to analyze these treatments under a hierarchical model where

$$\begin{aligned} y_i &\sim \text{Normal}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normal}(\mu, \tau^2) \end{aligned}$$

We will assume that $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_n^2)$ are all known without error. We are interested in posterior inference for each θ_i, which is the underlying true treatment effect for program i, as well as the common treatment mean μ and variance τ^2.

1. Verify that using a flat prior of $p(\mu, \tau) \propto 1$ corresponds to a prior for (μ, τ^2) of $p(\mu, \tau^2) \propto \tau^{-1}$. With this prior, write out the full posterior distribution of the unknown parameters μ, τ^2 and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

   Proof: Since $p(\mu, \tau) \propto 1$, we can denote $t = \tau^2$, then $\tau = \sqrt{t} \Rightarrow \frac{d\tau}{dt} = \frac{1}{2\sqrt{t}} \;\Rightarrow\; p(\mu, t) \propto \frac{1}{\sqrt{t}}$, since $t = \tau^2$, then we can get $p(\mu, \tau^2) \propto \frac{1}{\tau}$
   Denote $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, then the full posterior distribution of $\mu, \tau^2, \boldsymbol{\theta}$ is:

$$\begin{aligned} p(\mu, \tau^2 | \mathbf{y}, \boldsymbol{\theta}^2) &\propto p(\mu, \tau^2) p(\boldsymbol{\theta} | \mu, \tau^2) p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\sigma}^2) \\ &\propto \frac{1}{\tau} \prod_{i=1}^{n} \left[N(\theta_i | \mu, \tau^2) N(y_i | \theta_i, \sigma_i^2) \right] \end{aligned}$$

2. The first approach is based on the realization that we can actually integrate out each θ_i from the above distribution, giving us the marginal distribution of our treatment effects y_i:

$$y_i \sim \text{Normal}(\mu, \sigma_i^2 + \tau^2)$$

   Write out the marginal posterior distribution of $p(\mu, \tau^2 | \mathbf{y}, \boldsymbol{\sigma}^2)$, and evaluate this posterior distribution over a grid of values of μ and and τ^2. Use a grid of (0,10) for τ^2.

Solution: The marginal distribution of y_i is $N(\mu, \sigma_i^2 + \tau^2)$, then the marginal posterior:

$$\begin{aligned} p(\mu, \tau^2|\mathbf{y}, \boldsymbol{\sigma}^2) &\propto p(\mu, \tau^2)p(\mathbf{y}|\mu, \tau^2, \boldsymbol{\sigma}^2) \\ &\propto \frac{1}{\tau}\prod_{i=1}^{n} N(y_i|\mu, \sigma_i^2 + \tau^2) \\ \Rightarrow p(\mu|\tau^2, \mathbf{y}, \boldsymbol{\sigma}^2) &\propto e^{-\sum \frac{(y_i-\mu)^2}{2(\sigma_i^2+\tau^2)}} \\ \Rightarrow p(\mu|\tau^2, \mathbf{y}, \boldsymbol{\sigma}^2) &= N(\mu|\beta, V^2), \end{aligned}$$

where $V^2 = (\sum \frac{1}{\sigma_i^2+\tau^2}^{-1})$, $\beta = (\sum \frac{y_i}{\sigma_i^2+\tau^2})V^2$.

$$\Rightarrow \; p(\tau^2|\mathbf{y}, \boldsymbol{\sigma}^2) = \frac{p(\mu, \tau^2|\mathbf{y}, \boldsymbol{\sigma}^2)}{p(\mu|\tau^2, \mathbf{y}, \boldsymbol{\sigma}^2)} \propto \frac{\frac{1}{\tau}\prod_{i=1}^{n} N(y_i|\mu, \sigma_i^2 + \tau^2)}{N(\mu|\beta, V^2)}$$

3. Use the grid sampling method to get 1000 samples from $p(\mu, \tau^2|\mathbf{y}, \boldsymbol{\sigma}^2)$. Calculate the mean, median and 95% posterior interval for μ and τ^2.

Solution: 1000 samples were got from $p(\mu, \tau^2|\mathbf{y}, \boldsymbol{\sigma}^2)$ by using the grid sampling method.

	mean	median	95% posterior interval
μ	0.01546715	0.01252507	[-0.44199333,0.46522200]
τ^2	9.609489	9.709739	[8.668301,9.989991]

4. Write out the distribution $p(\theta_i|\mu, \tau^2, \mathbf{y}, \boldsymbol{\sigma}^2)$. Use the 1000 samples of μ and τ^2 to draw 1000 samples of each θ_i. Calculate the mean of each θ_i and compare to the observed treatment effects y_i.

Solution: Since $p(\mu, \tau^2, \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}^2) \;\propto\; \frac{1}{\tau}\prod_{i=1}^{n}[N(\theta_i|\mu, \tau^2)N(y_i|\theta_i, \sigma_i^2)]$, then:

$$\begin{aligned} p(\theta_i|\mu, \tau^2, \mathbf{y}, \boldsymbol{\sigma}^2) &\propto N(\theta_i|\mu, \tau^2)N(y_i|\theta_i, \sigma_i^2) \\ &\propto e^{-\frac{1}{2}(\frac{1}{\tau^2}+\frac{1}{\sigma_i^2})\theta_i^2+(\frac{\mu}{\tau^2}+\frac{y_i}{\sigma_i^2})\theta_i} \\ \Rightarrow p(\theta_i|\mu, \tau^2, \mathbf{y}, \boldsymbol{\sigma}^2) &= N(\theta_i|\mu_i^*, \tau_i^{2*}) \end{aligned}$$

1000 samples of $\boldsymbol{\theta}$ were drawn by using the samples of μ and τ^2. The mean of each θ_i is:

θ_1	θ_2	θ_3	θ_4
1.11757159	0.81483007	-0.14736517	0.71563212
θ_5	θ_6	θ_7	θ_8
-0.02328974	0.15142220	1.71089974	0.20240492

5. For Questions 5-7, consider τ^2 to be fixed and equal to $\tau^2 = \text{median}(\tau^2)$ that you calculated in Question 3. We now use a Gibbs sampler to draw samples from the posterior distribution $p(\mu, \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}^2)$. Remember that τ^2 is now a known quantity in this question, so we do not need to calculate the posterior distribution for it. Give the conditional distributions of the remaining unknown parameters given the other parameters.

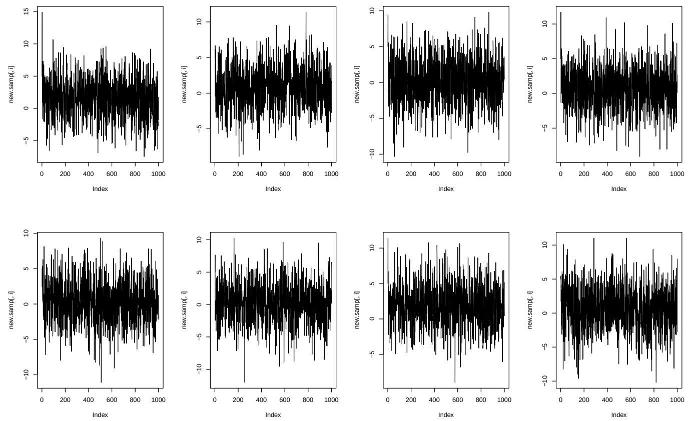
   Solution: When τ^2 is fixed, the joint posterior of μ, $\boldsymbol{\theta}$ is (we denote τ_0^2 =median(τ^2)):

$$p(\mu, \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}^2, \tau_0^2) \;\propto\; p(\mu)p(\boldsymbol{\theta}|\mu, \tau_0^2)p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\sigma}^2) \propto \prod_{i=1}^{n} \left[N(\theta_i|\mu, \tau_0^2)p(y_i|\theta_i, \sigma_i^2)\right]$$

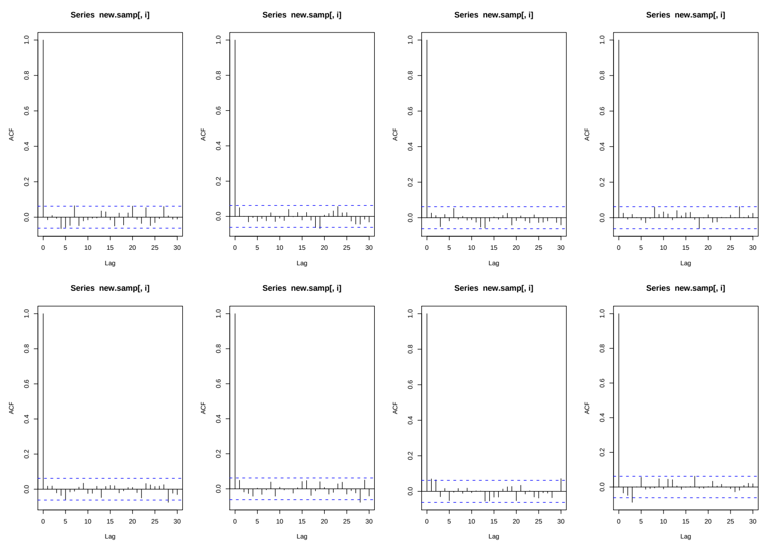
$$\begin{aligned}
P(\theta_i|\mu, \mathbf{y}, \boldsymbol{\sigma}^2, \tau_0^2) \;&\propto\; N(\theta_i|\mu_i, \tau_0^2)p(y_i|\theta_i, \sigma_i^2) \propto \; e^{-\frac{1}{2}(\frac{1}{\tau_0^2}+\frac{1}{\sigma_i^2})\theta_i^2+(\frac{\mu}{\tau_0^2}+\frac{y_i}{\sigma_i^2}\theta_i)} \\
&\Rightarrow\; p(\theta_i|\mu, \mathbf{y}, \boldsymbol{\sigma}^2, \tau_0^2) = N(\theta_i|\xi_i, \gamma_i^2) \\
&\gamma_i^2 = (\frac{1}{\tau_0^2} + \frac{1}{\sigma_i^2})^{-1},\; \xi_i = (\frac{\mu}{\tau_0^2} + \frac{y_i}{\sigma_i^2})\gamma_i^2 \\
p(\mu|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\sigma}^2, \tau_0^2) \;&\propto\; \prod_{i=1}^{n} N(\theta_i|\mu, \tau_0^2) \;\propto\; e^{-\frac{n}{2\tau_0^2}\mu^2+\frac{\sum\theta_i}{\tau_0^2}\mu} \\
&\Rightarrow\; p(\mu|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\sigma}^2, \tau_0^2) = N(\mu|\frac{\sum\theta_i}{n}, \frac{\tau_0^2}{n})
\end{aligned}$$

6. Use a Gibbs sampler based on the conditional distributions from the previous question to obtain 1000 samples from $p(\mu, \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}^2)$. Make sure that you only use samples after convergence and that your 1000 samples have low autocorrelation. Calculate the mean and posterior interval for each θ_i.

   Solution: We can plot the samples we simulated:



The ACF graphics are:



We can see from the plot of the samples, and the ACF plot that the samples have low autocorrelation and good convergence, so we use the samples we simulated to calculate the mean and posterior interval for each θ_i.
The mean and posterior interval for each θ_i is:

	mean	95% posterior interval
θ_1	1.4707507	[-4.28861989,7.4656290]
θ_2	0.9016350	[-5.01630629,6.6062699]
θ_3	0.0796483	[-6.00374072, 6.2903041]
θ_4	0.6655649	[-5.30738770,6.4243711]
θ_5	0.2933542	[-5.52278423,6.0655314]
θ_6	0.2238173	[-5.59824110,6.1768044]
θ_7	1.8908095	[-3.94686628,7.9728499]
θ_8	0.6777565	[-5.48349935,6.9848530]

7. Use your samples of $\boldsymbol{\theta}$ to calculate an estimate of the posterior probability that school A offers the best program.

   Solution: Simulated prediction values of y, the probability that school A offers the best program is 0.1709212.

   Questions 8-11 are based on the bicycle data in Table 2 below. We will focus only on the first two rows of the table (the residential streets with bike routes). We want to model the total amount of traffic y_i on each street (eg. $y_1 = 74$) as follows:

$$\begin{aligned} y_i &\sim \text{Poisson}(\theta_i) \\ \theta_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

8. Write out the posterior distribution of our unknown variables $p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y})$ using a flat prior distribution for α and β, $p(\alpha, \beta) \propto 1$.

   Solution: $y_i \sim Poisson(\theta_i)$, $\theta_i \sim Gamma(\alpha, \beta)$

$$p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y}) \;\propto\; p(\alpha, \beta)p(\boldsymbol{\theta}|\alpha, \beta)p(\mathbf{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [Gamma(\theta_i|\alpha, \beta)Poisson(y_i|\theta_i)]$$

9. We want to use a Gibbs sampler to get samples from $p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y})$. What is the conditional distribution of θ_i given all the other parameters?

   Solution:

$$\begin{aligned} p(\theta_i|\boldsymbol{\theta}_{-i}, \alpha, \beta, \mathbf{y}) &= p(\theta_i|\alpha, \beta, \mathbf{y}) \\ &= Gamma(\theta_i|\alpha, \beta)Poisson(y_i|\theta_i) \\ &\propto \theta_i^{\alpha-1}e^{-\beta\theta_i}\theta_i^{y_i}e^{-\theta_i} \\ &\propto \theta_i^{\alpha+y_i-1}e^{-(\beta+1)\theta_i} \\ &= Gamma(\alpha + y_i, \beta + 1) \end{aligned}$$

10. The conditional distributions $p(\alpha, \beta|\mathbf{y})$ and $p(\beta|\alpha, \mathbf{y})$ are not easy to sample from, so you instead need to use a Metropolis (or Metropolis-Hastings) step to obtain samples from $p(\alpha, \beta|\boldsymbol{\theta}, \mathbf{y})$. Choose a proposal distribution for α and a proposal distribution for β.

    Solution: We can set $\alpha^* = \alpha^{(l)} + z$,and $\beta^* = \beta^{(l)} + z$, where $z \sim N(0, 1)$, as our proposal distribution for α and β, i.e., we need to simulate $\alpha^* \sim N(\alpha^{(l)}, 1)$ and $\beta^* \sim N(\beta^{(l)}, 1)$.

$$
\begin{aligned}
p(\alpha, \beta|\mathbf{y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} Gamma(\theta_i|\alpha, \beta) \\
&= \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{\alpha-1} e^{-\beta\theta_i} \\
&= \frac{\beta^{na}}{(\Gamma(\alpha))^n} (\prod_{i=1}^{n} \theta_i)^{\alpha-1} e^{-\beta\sum\theta_i}
\end{aligned}
$$

$$
\Rightarrow \ logp(\alpha, \beta|\mathbf{y}, \boldsymbol{\theta}) = n\alpha log\beta - nlog\Gamma(\alpha) + (\alpha - 1)\sum log\theta_i - \beta \sum \theta_i
$$

11. Combine the algorithm from Question 10 with the conditional distribution from Question 9 to form a Gibbs sampler, which iterates between sampling:

    1. sampling $\theta_i|\alpha, \beta, \mathbf{y}$ for each i.
    2. sampling $\alpha, \beta|\boldsymbol{\theta}, \mathbf{y}$.

    Use this algorithm to obtain 1000 samples from $p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y})$. Make sure that you only use samples after convergence and that your 1000 samples have low autocorrelation. Calculate means and 95% posterior intervals for α and β.

12. Give a 95% predictive interval for the total traffic on a new residential street with a bike lane.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Table 1: *Observed effects of special preparation on SAT-V scores in eight randomized experiments. Estimates are based on separate analyses for the eight experiments. From Rubin (1981).*

Type of street	Bike route?	Counts of bicycles/other vehicles
Residential	yes	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154
Fairly busy	yes	8/29, 35/415, 31/425, 19/42, 38/180, 47/675, 44/620, 44/437, 29/47, 18/462
Fairly busy	no	10/557, 43/1258, 5/499, 14/601, 58/1163, 15/700, 0/90, 47/1093, 51/1459, 32/1086
Busy	yes	60/1545, 51/1499, 58/1598, 59/503, 53/407, 68/1494, 68/1558, 60/1706, 71/476, 63/752
Busy	no	8/1248, 9/1246, 6/1596, 9/1765, 19/1290, 61/2498, 31/2346, 75/3101, 14/1918, 25/2318

Table 2: *Counts of bicycles and other vehicles in one hour in each of 10 city blocks in each of six categories. (The data for two of the residential blocks were lost.) For example, the first block had 16 bicycles and 58 other vehicles, the second had 9 bicycles and 90 other vehicles, and so on. Streets were classified as 'residential,' 'fairly busy,' or 'busy' before the data were gathered.*