STA 602 HW 11

Ryan Tang

November 18th 2022

1 Exercise 8.1

(a) $Var[y_{i,j}|\mu, \tau^2]$ should be bigger than $Var[y_{i,j}|\theta_j, \sigma^2]$ that the variances of the population should always be larger or equal to the with-in group variances. The worse case is that the two variances are the same, $\theta_i = \mu \, \forall i$.

(b) $Cov[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2]$ should be closer to zero under the conditional iid assumption from the model. In other words, other observations in the same group don't tell us additional information if we already know the group's average performance.

$Cov[y_{i_1,j}, y_{i_2,j}|\mu, \tau^2]$ should be either negative or positive when the group-specific performance is unknown. The sign has to be determined by the actual observation differences and τ^2. If the two observation is far-part, then we think it gives us some information that the group means should be in the middle of the two — negative covariance. On the other hand, if the two observations are close, the group means will shrink towards the global mean μ — positive covariance.

(c) Following the conditional independence of the hierarchical model, we have the following. We can see the population variances are always higher than the within-group variances by τ^2.

$$
\begin{aligned}
Var[y_{i,j}|\theta_j, \sigma^2] &= \sigma^2 \\
Var[y_{i,j}|\mu, \tau^2] &= \mathbb{E}[Var[y_{i,j}|\theta_j, \sigma^2]|\mu, \tau^2] + Var[\mathbb{E}[y_{i,j}|\theta_j, \sigma^2]|\mu, \tau^2] \\
&= \mathbb{E}[\sigma^2] + Var[\sigma_j|\mu, \tau^2] \\
&= \sigma^2 + \tau^2 \\
Var[\bar{y}_{.,j}|\theta_j, \sigma^2] &= \frac{1}{N_j^2} Var[\sum_i y_{i,j}|\theta_j, \sigma^2] \\
&= \frac{\sigma^2}{N_j} \\
Var[\bar{y}_{.,j}|\mu, \tau^2] &= \mathbb{E}[\frac{1}{N_j^2} Var[y_{i,j}|\theta_j \sigma^2]|\mu, \tau^2] + Var[\frac{1}{N_j}\mathbb{E}[y_{i,j}|\theta_j]|\mu, \tau^2] \\
&= \mathbb{E}[\frac{\sigma^2}{N_j}|\mu, \tau^2] + Var[\theta_j|\mu, \tau^2] \\
&= \frac{m}{N}\sigma^2 + \tau^2 \\
Cov[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2] &= \mathbb{E}[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2] - \mathbb{E}[y_{i,j}|\theta_j, \sigma^2]\mathbb{E}[y_{i,j}|\theta_j, \sigma^2] \\
&= Var[y_{i,j}|\theta_j, \sigma^2] \\
&= \sigma^2 \\
Cov[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2] &= Var[y_{i,j}|\mu, \tau^2] \\
&= \sigma^2 + \tau^2
\end{aligned}
$$

(d) In words, the posterior of the population mean depends only on the posterior within group means. Of course, the with-in group posterior means have already incorporated all the information from the individual data points.

$$\begin{aligned} p(\mu|\theta,\sigma^2,\tau^2,Y) &\propto p(\mu)p(\theta|\mu,\tau^2)p(Y|\theta,\sigma^2) \\ &\propto p(\mu)p(\theta|\mu,\tau^2) \\ &\propto p(\theta,\mu|\tau^2) \\ &= p(\mu|\theta,\tau^2) \end{aligned}$$

2 Exercise 8.2

We are given the following hierarchical model.

$$\begin{aligned} \theta_A &= \mu + \delta \\ \theta_B &= \mu - \delta \\ \mu &\sim \mathcal{N}(\mu_o = 75, \lambda_o^2 = 100 = \tilde{\lambda}_o^{-1}) \\ \delta &\sim \mathcal{N}(\delta_o, \tau_o^2 = \tilde{\tau}_o^{-1}) \\ \gamma &= \frac{1}{\sigma^2} \sim Gamma(\nu_o = 1, \sigma_o^2 = 100) \\ y_i &\sim \mathcal{N}(\mu, \delta, \sigma^2) \end{aligned}$$

The full conditionals for all the parameters μ, γ, δ are given below.

$$\begin{aligned} N_A &= \sum_i^N \mathbb{I}(y_i \in A) \\ N_B &= \sum_i^N \mathbb{I}(y_i \in B) \\ N &= N_A + N_B \\ p(\delta|\mu,Y,\sigma^2) &= p(\delta)\prod_i^{N_A}\mathcal{N}(y_{A_i}|\mu+\delta,\sigma^2)\prod_i^{N_B}\mathcal{N}(y_{B_i}|\mu-\delta,\sigma^2) \\ &\sim \mathcal{N}(\delta_n = \tau_n^2[\tilde{\tau}_o\delta_o + \gamma(N_A\bar{\tilde{Y}}_A + N_B\bar{\tilde{Y}}_B)], \tau_n^2 = (\tilde{\tau}_o + \gamma N)^{-1}) \\ \bar{Y}_{A,\delta} &= \frac{1}{N_A}\sum_{i\in A}(y_i - \mu) \quad \bar{Y}_{B,\delta} = \frac{1}{N_B}\sum_{i\in B}(\mu - y_i) \end{aligned}$$

$$\begin{aligned} p(\mu|\delta,Y,\sigma^2) &\sim \mathcal{N}(\delta_n = \lambda_n^2[\tilde{\lambda}_o\delta_o + \gamma(N_A\bar{\tilde{Y}}_A + N_B\bar{\tilde{Y}}_B)], \lambda_n^2 = (\tilde{\lambda}_o + \gamma N)^{-1}) \\ \bar{Y}_{A,\mu} &= \frac{1}{N_A}\sum_{i\in A}(y_i - \delta) \quad \bar{Y}_{B,\mu} = \frac{1}{N_B}\sum_{i\in B}(y_i + \delta) \end{aligned}$$

$$p(\gamma|\delta,\mu,Y) \sim Gamma(\nu_u = \nu_0 + N, \sigma_n^2 = \frac{1}{\nu_n}(\nu_o\sigma_o^2 + \sum_{i\in A}(y_{A_i} - \mu - \delta)^2 + \sum_{i\in B}(y_{B_i} - \mu + \delta)^2))$$

(a) Sensitivity Analysis Below are the posterior results from a wide range of different combinations of prior parameters. All chains behave quite nicely after 1000 burn-in periods. And all statistics were calculated using 10,000 MCMC samples.

- We are certainly seeing some evidence that δ is negative, the mean of group A is greater than group B.

- We see that larger prior variances, τ_o, leads to a higher probability that θ_B is less than θ_A when δ_0 is positive. However, it also contributes to a wider confidence interval because we only have 32 observations. Posterior correlation also shrinks towards 0 when we have higher variance priors

δ_o	τ_0^2	$P[\delta < 0\|Y]$	95% CI LB	95% CI UB	Post Theta Corr	Prior Theta Corr
-4	10	0.88	-7.83	2.01	0.40	0.82
-4	50	0.70	-8.74	4.85	0.08	0.33
-4	100	0.66	-9.03	5.85	0.01	0.01
-4	500	0.62	-9.42	6.75	-0.06	-0.67
-2	10	0.75	-6.53	3.20	0.40	0.82
-2	50	0.66	-8.39	5.56	0.07	0.33
-2	100	0.63	-8.82	6.23	0.00	-0.01
-2	500	0.62	-9.15	6.86	-0.07	-0.67
0	10	0.57	-5.27	4.50	0.40	0.82
0	50	0.59	-7.78	6.05	0.07	0.33
0	100	0.60	-8.47	6.48	0.00	0.00
0	500	0.61	-8.80	7.00	-0.05	-0.67
2	10	0.38	-3.96	5.68	0.40	0.82
2	50	0.54	-7.32	6.68	0.05	0.34
2	100	0.58	-8.09	6.73	-0.01	0.00
2	500	0.60	-9.10	6.64	-0.05	-0.67
4	10	0.21	-2.85	6.92	0.41	0.82
4	50	0.49	-6.93	7.24	0.06	0.33
4	100	0.54	-7.81	7.20	-0.00	-0.01
4	500	0.59	-8.89	7.05	-0.07	-0.67

(b) Prior Opinions

- If one has no strong opinion or belief $\tau_o^2 = 500$, she can use a diffuse prior with $\delta = 0$. In this case, $P[\delta < 0|Y] = 0.61$ with a wide 95% confidence interval. In other words, she will need more evidence to confirm the differences.
- If one is sure that Group A under-performs Group B on average but uncertain about the truthfulness of this belief, she can opt for a $\delta_o = -4$, $\tau_o^2 = 500$ prior. It, too, leads to a $p[\delta < 0|Y] = 0.62$ and a wide 95% confidence interval.
- One can also be strongly confident that Group A under-performs Group B; she can choose a $\delta_o = -4$, $\tau_o^2 = 10$ prior.
- One can also use unit information prior, which corresponds to $\delta_o = -2$, $\tau_0^2 = 50$.

3 Exercise 8.3

We are given data from 8 schools and would like to use the hierarchical normal model to do inference among schools.

$$
\begin{aligned}
y_{ij}|\theta_j, \sigma^2 &\sim \mathcal{N}(y_{ij}|\theta_j, \sigma^2) && i = 1 \dots N,\ j = 1 \dots M \\
\theta_j|\mu, \tau^2 &\sim \mathcal{N}(\theta_j|\mu, \tau^2) \\
\mu &\sim \mathcal{N}(\mu|\mu_o, \gamma_o^2) \\
\frac{1}{\tau^2} &\sim Gamma(\eta_o, \tau_o^2) \\
\frac{1}{\sigma^2} &\sim Gamma(\nu_o, \sigma_o^2)
\end{aligned}
$$

And the posterior full conditionals are given below.

$$\begin{aligned}
\theta_j|Y,\mu,\tau^2,\sigma^2 &\propto p(\theta_j|\mu,\tau^2)p(y_{.j}|\theta_j,\sigma^2) \\
&\propto p(\theta_j|\mu,\tau^2)\prod_i^{n_j} p(y_{ij}|\theta_j,\sigma^2) \\
&\sim \mathcal{N}\left[\theta_j|\mu_n = \tau_n^2(\frac{1}{\sigma^2}n_j\bar{y_j} + \frac{1}{\tau^2}\mu), \tau_n^2 = (\frac{1}{\sigma^2}n_j + \frac{1}{\tau^2})^{-1}\right]
\end{aligned}$$

$$\begin{aligned}
\mu|Y,\tau^2,\theta &\propto \left[\prod_j^m p(\theta_j|\mu,\tau^2)\right] p(\mu|\mu_o,\gamma_o^2) \\
&\sim \mathcal{N}\left[\mu|\mu_n = \gamma_n^2(\frac{1}{\tau^2}m\bar{\theta} + \frac{1}{\gamma_o^2}\mu_o), \gamma_n^2 = (\frac{1}{\tau^2}m + \frac{1}{\gamma_o^2})^{-1}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{1}{\tau^2}|Y,\mu,\theta &\propto p(\tau^2|\eta_o,\tau_o^2)\prod_j^m p(\theta_j|\mu,\tau^2) \\
&\sim Gamma\left[\alpha = \frac{m+\eta_o}{2}, \beta = \frac{1}{2}(\eta_o\tau_o^2 + \sum_j(\theta_j-\mu)^2)\right]
\end{aligned}$$

$$\begin{aligned}
\frac{1}{\sigma^2}|Y,\theta &\propto p(\sigma^2|\nu_o,\sigma_o^2)\prod_j^m\prod_i^{n_j} p(y_{ij}|\theta_j,\sigma^2) \\
&\sim Gamma\left[\alpha = \frac{n+\nu_o}{2}, \beta = \frac{1}{2}(\nu_o\sigma_o^2 + \sum_j\sum_i(y_{ij}-\theta_j)^2)\right]
\end{aligned}$$

(a) Gibbs Sampling Using the above full conditionals, I drew 10,000 MCMC samples after a 1,000 burn-in period. Below are the trace plots, auto-correlation plots, and the corresponding density plots for each variable. We can see the chains perform just fine with good mixing and converge quite successfully.

- ESS for σ^2 is 8,995
- ESS for τ^2 is 9,551
- ESS for μ is 5,948

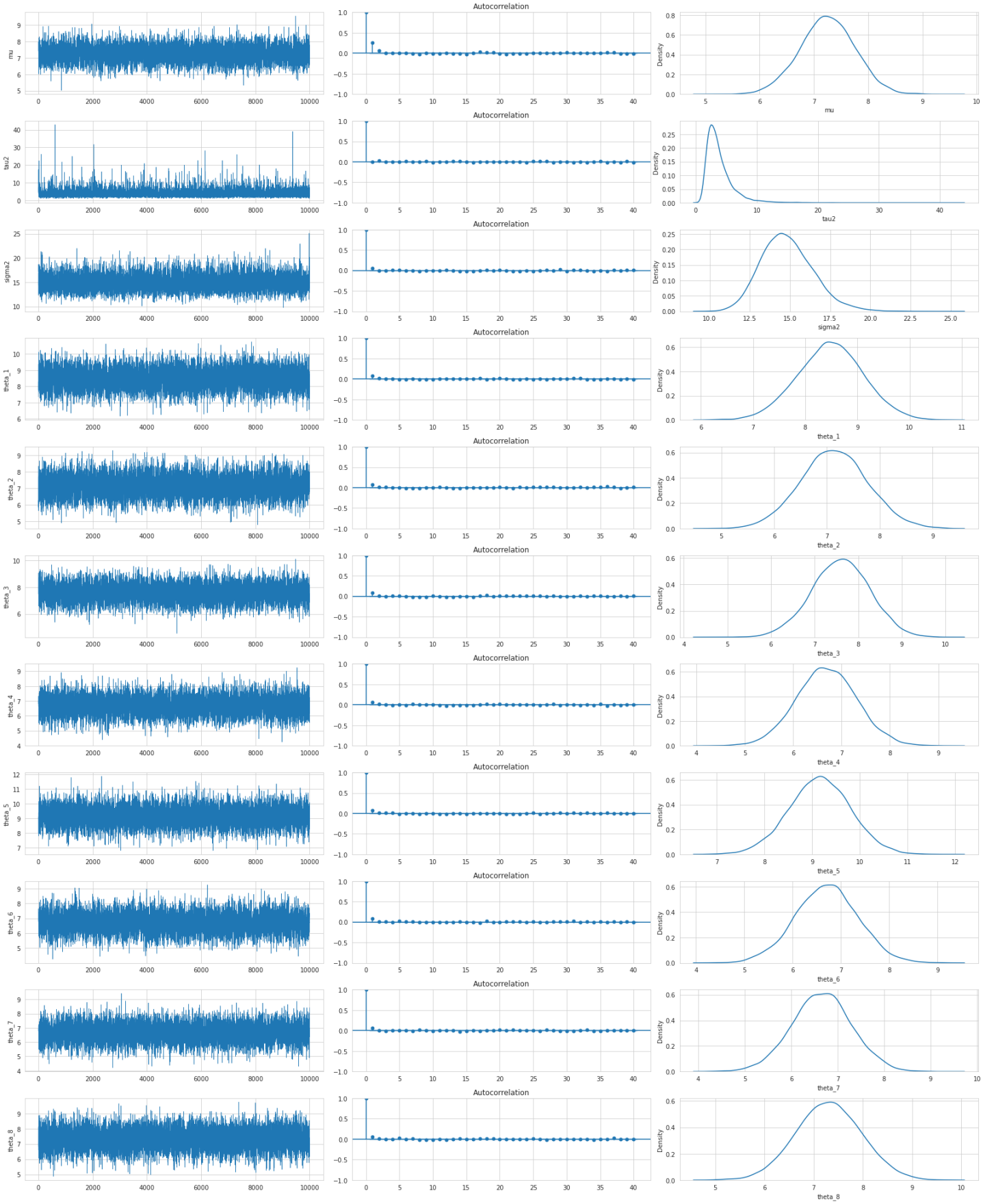

Figure 1: MCMC Diagnostics

(b) Posterior Statistics Here are some statistics and prior and posterior comparisons.

- μ posterior mean is 7.25 and 95% CI $[6.254, 8.221]$
- τ^2 posterior mean is 3.85 and 95% CI $[1.422, 9.800]$
- σ^2 posterior mean is 14.84 and 95% CI $[11.988, 18.393]$

We can see that the posterior global group mean μ certainly becomes narrower than before. The posterior global group precision $\frac{1}{\tau^2}$ is also slightly higher than the prior. The most significant changes are from the individual precision posterior $\frac{1}{\sigma^2}$; it has become really narrow than the diffuse prior. In other words, the observation data gave us information saying the global between-group mean is around 7.5 and has lower variances than our priors between and with-in groups. Lastly, all θ_j have been shrinking much narrower than the priors.

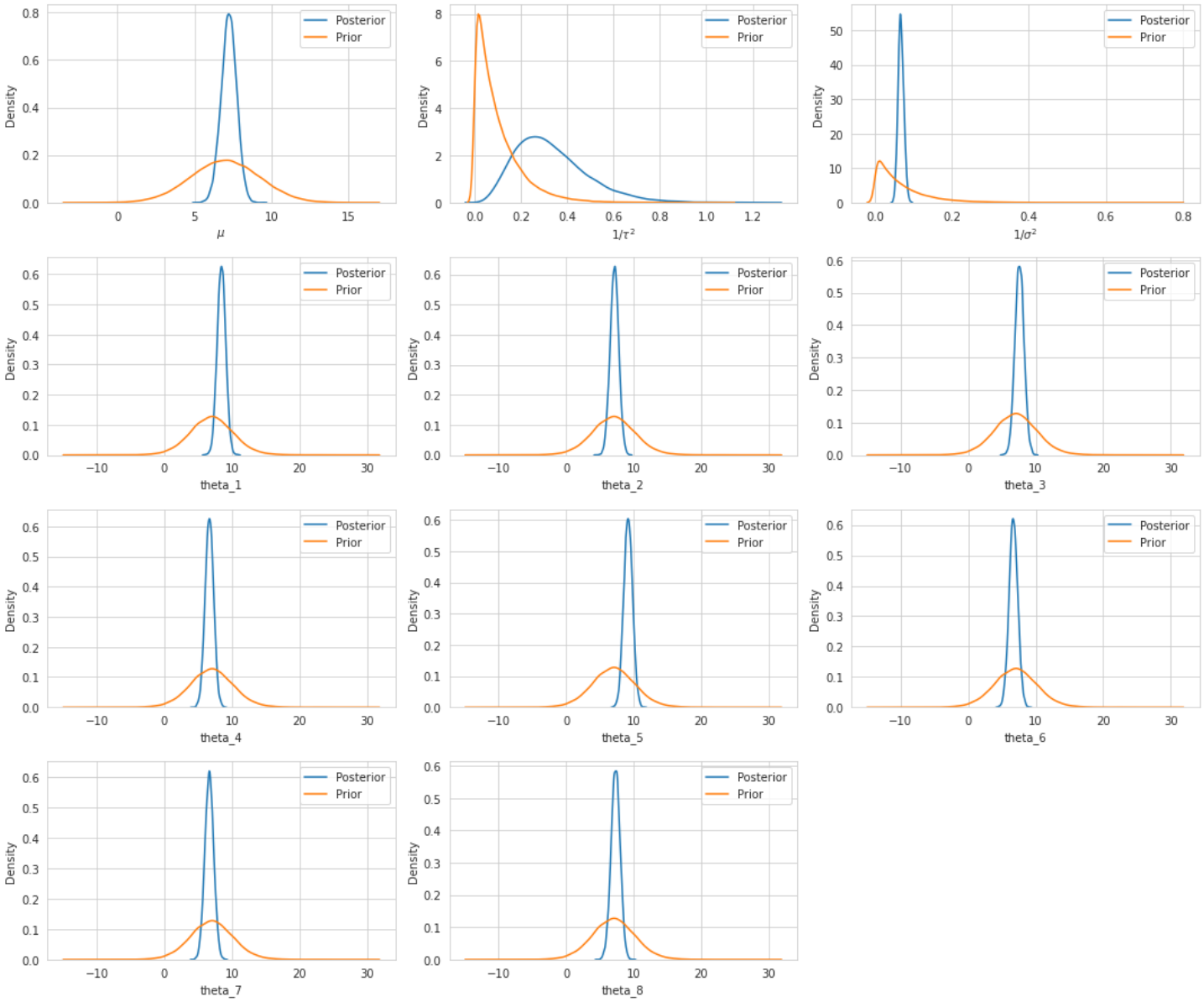

Figure 2: Posteriors of all parameters

(c) Initially, we set our priors assuming the between-school and individual student variances are similar. However, after sampling, we can see from the below graph that the between-school variations are certainly less than individual student variances.

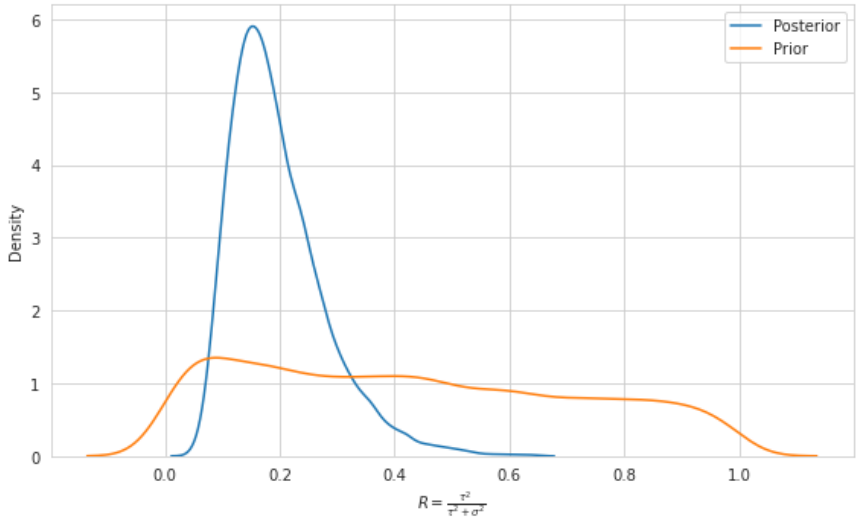

Figure 3: Between School Variations

(d)

- The $P[\theta_7 < \theta_6|Y] = 0.5172$
- The $P[\theta_7 < min\{\theta_1, ...\theta_6, \theta_8\}|Y] = 0.2898$

(e) Here I plot the various mean using a bar plot. The posterior shrinks the mean value towards $\mu|Y$.

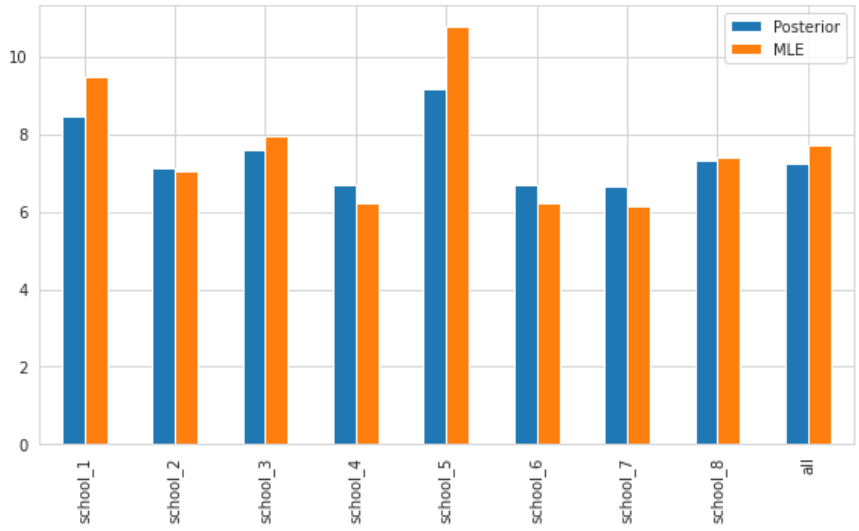

Figure 4: Posterior Means vs MLE Means