MATH457 Midterm

Adam Rockett

May 2019

Introduction

With the data given and looking at Figure 1, we can clearly see that the Y values come from a mixture distribution where the mean of the smaller values appears to be around $\theta_1 = 110$, and the mean of the larger values is close to $\theta_2 = 150$. Therefore, when we come up with our prior we will use these values as the prior means. Also, looking at the kernel density line, we can clearly tell that the variance of the smaller distribution is greater than the larger distribution, so we will use values of $\sigma_1^2 = 100$, and $\sigma_2^2 = 64$. Judging by how similar the peaks of both the distributions are, I believe that an equal number observations come from each distribution, and my priors will reflect this.

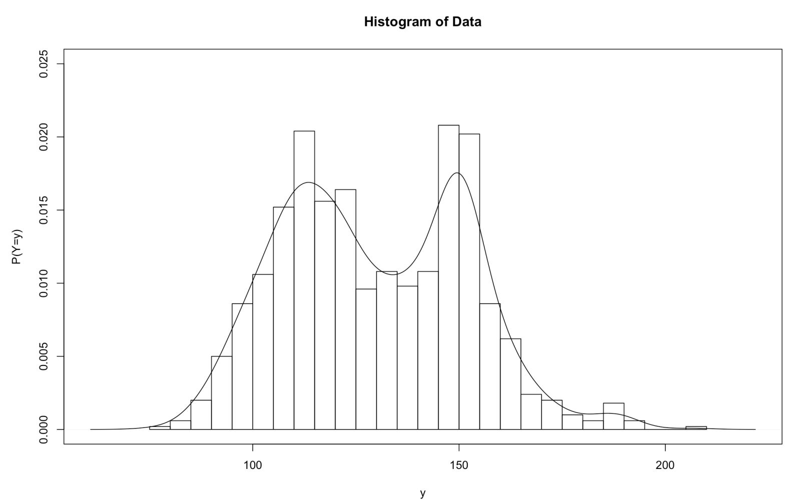

Figure 1: Histogram of Data Given

Priors

$$
\begin{aligned}
p &\sim beta(a,b) \\
X_i &\sim Bernoulli(p) \\
Y_i &\sim N(\theta_1, \sigma_1^2), if X_i = 1 \\
Y_i &\sim N(\theta_2, \sigma_2^2), if X_i = 0 \\
\theta_j &\sim N(\mu_0, \tau_0^2), j = 1, 2 \\
1/\sigma_j^2 &\sim Gamma(\nu_0/2, \nu_0\sigma_0^2/2)
\end{aligned}
$$

The joint distribution for all of the variables is

$$
\begin{aligned}
p(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p, \vec{X}|\vec{Y}) &\propto p(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p, \vec{X}, \vec{Y}) \\
&= p(\vec{Y}|\vec{X}, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \times p(\theta_j) \times p(1/\sigma_j^2) \times p(p) \times p(\vec{X}|p).
\end{aligned}
$$

Sampling Distribution

The sampling distribution for the Y values can be given by

$$
\begin{aligned}
p(\vec{Y}|\vec{X}, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p) &= \prod_{i=1}^{n} p(Y_i|X_i, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \\
&= \prod_{i=1}^{n} dnorm(y_i, \theta_1, \sigma_1^2)^{x_i} \times dnorm(y_i, \theta_2, \sigma_2^2)^{1-x_i},
\end{aligned}
$$

where $dnorm(y_i, \theta_1, \sigma_1^2)$ signifies the pdf of a $N(\theta_1, \sigma_1^2)$

Full Conditionals

The full conditional distribution of X_i can be given by

$$
\begin{aligned}
P(X_i = 1|Y_i, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p) &\propto \frac{P(X_i = 1|p) \times p(Y_i|x_i = 1, \theta_1, \sigma_1^2)}{p(Y_i|x_i, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2)} \\
&= \frac{p \times p(Y_i|x_i = 1, \theta_1, \sigma_1^2)}{p \times p(Y_i|x_i = 1, \theta_1, \sigma_1^2) + (1-p) \times p(Y_i|x_i = 0, \theta_2, \sigma_2^2)} \\
&= \frac{p \times dnorm(y_i, \theta_1, \sigma_1^2)}{p \times dnorm(y_i, \theta_1, \sigma_1^2) + (1-p) \times dnorm(y_i, \theta_2, \sigma_2^2)}.
\end{aligned}
$$

Therefore, we can conclude that

$$X_i|y_i, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p \sim Bernoulli\left(\frac{p \times dnorm(y_i, \theta_1, \sigma_1^2)}{p \times dnorm(y_i, \theta_1, \sigma_1^2) + (1-p) \times dnorm(y_i, \theta_2, \sigma_2^2)}\right)$$

Now, finding the full conditional distribution of p, we get

$$\begin{aligned} p(p|\vec{Y}, \vec{X}, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) &\propto \text{Joint pdf of all from above} \\ &\propto p(p) \times p(\vec{X}|p) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \times \prod_{i=1}^{n} p^x(1-p)^{1-x} \\ &\propto p^{\sum x_i + a - 1}(1-p)^{n - \sum x_i + b - 1}. \end{aligned}$$

Looking at this result, we can conclude that p comes from a beta distribution, given by

$$p|\vec{Y}, \vec{X}, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2 \sim beta\left(\sum x_i + a, n - \sum x_i + b\right).$$

Now to find the full conditional for θ_1.

$$\begin{aligned} p(\theta_1|\vec{Y}, \vec{X}, p, \theta_2, \sigma_1^2, \sigma_2^2) &\propto \text{Joint pdf of all variables} \\ &\propto p(\vec{Y}|\vec{X}, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \times p(\theta_1) \\ &= \prod_{i=1}^{n} \left[dnorm(y_i, \theta_1, \sigma_1^2)^{x_i} \times dnorm(y_i, \theta_2, \sigma_2^2)^{1-x_i}\right] \times dnorm(\theta_1, \mu_0, \tau_0^2) \\ &\propto \prod_{i=1}^{n} \left[dnorm(y_i, \theta_1, \sigma_1^2)^{x_i}\right] \times dnorm(\theta_1, \mu_0, \tau_0^2) \\ &\propto exp\left\{-\frac{1}{2\tau_0^2}(\theta_1 - \mu_0)^2 - \frac{1}{2\sigma_1^2}\sum (y_i - \theta_1)^2\right\}. \end{aligned}$$

Then, ignoring the $-1/2$ in front, inside the $exp\{\}$ we have

$$\begin{aligned} \frac{1}{\tau_0^2}(\theta_1^2 - 2\theta_1\mu_0 + \mu_0^2) + \frac{1}{\sigma_1^2}(\sum y_i^2 - 2\theta_1 \sum y_i + n\theta_1^2) = a\theta_1^2 - 2b\theta_1 + c, \text{ where} \\ a = \frac{1}{\tau_0^2} + \frac{n}{\sigma_1^2}, b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma_1^2}, \text{ and } c = c(\mu_0, \tau_0^2, \sigma_1^2, \vec{y}). \end{aligned}$$

Now putting it all together, we get

$$
\begin{aligned}
p(\theta_1|\vec{Y},\vec{X},p,\theta_2,\sigma_1^2,\sigma_2^2) &\propto exp\left\{-\frac{1}{2}(a\theta_1^2-2b\theta_1)\right\} \\
&= exp\left\{-\frac{1}{2}a(\theta_1^2-2b\theta_1/a+b^2/a^2)+\frac{1}{2}b^2/a\right\} \\
&\propto exp\left\{-\frac{1}{2}a(\theta_1-b/a)^2\right\} \\
&= exp\left\{-\frac{1}{2}\left(\frac{\theta_1-b/a}{1/\sqrt{a}}\right)^2\right\}.
\end{aligned}
$$

This function follows the shape of a normal distribution where the mean is b/a, and the variance is $1/a$. Therefore, if $Y_1 = \{Y_i : X_i = 1\}$, n_1 is the size of Y_1, and $\bar{y_1}$ is the mean of values in Y_1, we can conclude

$$
\begin{aligned}
\theta_1|\vec{Y},\vec{X},p,\theta_2,\sigma_1^2,\sigma_2^2 &\sim N(\mu_{n,1},\tau_{n,1}^2), \text{ where} \\
\tau_{n,1}^2 &= \frac{1}{a} = \frac{1}{\frac{1}{\tau_0}+\frac{n_1}{\sigma_1^2}}, \text{ and} \\
\mu_{n,1} &= \frac{b}{a} = \left(\frac{\mu_0}{\tau_0^2}+\frac{n_1\bar{y_1}}{\sigma_1^2}\right)\times\tau_{n,1}^2.
\end{aligned}
$$

Without loss of generality, we can follow a similar process to come up with the distribution for θ_2, and will end up with

$$
\begin{aligned}
\theta_2|\vec{Y},\vec{X},p,\theta_1,\sigma_1^2,\sigma_2^2 &\sim N(\mu_{n,2},\tau_{n,2}^2), \text{ where} \\
\tau_{n,2}^2 &= \frac{1}{\frac{1}{\tau_0}+\frac{n_2}{\sigma_2^2}}, \text{ and} \\
\mu_{n,2} &= \left(\frac{\mu_0}{\tau_0^2}+\frac{n_2\bar{y}}{\sigma_2^2}\right)\times\tau_{n,2}^2,
\end{aligned}
$$

where $Y_2 = \{Y_i : X_i = 0\}$, n_2 is the size of Y_2, and $\bar{y_2}$ is the mean of values in Y_2.

Finally, we must find the full conditional distribution of σ_1^2 and σ_2^2. Since we know the prior distribution of $1/\sigma_1^2$, we will find the full conditional based on the precision. Therefore,

$$
\begin{aligned}
p(1/\sigma_1^2|\vec{Y},\vec{X},p,\theta_1,\theta_2,\sigma_2^2) &\propto \text{Joint pdf of all variables} \\
&\propto p(\vec{Y}|\vec{X},p,\theta_1,\theta_2,\sigma_2^2) \times p(1/\sigma_1^2) \\
&\propto \left((\tilde{\sigma}_1^2)^{n_1/2} exp\left\{ -\tilde{\sigma}_1^2 \sum_{i=1}^{n_1} (y_i - \theta_1)^2/2 \right\} \right) \\
&\quad \times \left((\tilde{\sigma}_1^2)^{\nu_0/2-1} exp\left\{ -\tilde{\sigma}_1^2 \nu_0 \sigma_0^2/2 \right\} \right) \\
&= (\tilde{\sigma}_1^2)^{(\nu_0+n_1)/2-1} \times exp\left\{ -\tilde{\sigma}_1^2 \times \left[\nu_0\sigma_0^2 + \sum (y_i - \theta_1)^2 \right] /2 \right\},
\end{aligned}
$$

where all the y_i values have corresponding x_i values that are 1. This function takes the shape of a gamma distribution, therefore,

$$
\begin{aligned}
1/\sigma_1^2|\vec{Y},\vec{X},p,\theta_1,\theta_2,\sigma_2^2 &\sim Gamma\left(\frac{\nu_{n,1}}{2}, \frac{\sigma_{n_1}^2 \nu_{n,1}}{2} \right), \text{ where} \\
\nu_{n,1} &= \nu_0 + n_1, \text{ and,} \\
\sigma_{n,1}^2 &= \frac{1}{\nu_{n,1}} \left[\nu_0 \sigma_0^2 + n_1 s_{n_1}^2(\theta_1) \right], \text{ where,} \\
s_{n,1}^2(\theta_1) &= \sum (y_i - \theta_1)^2/n_1.
\end{aligned}
$$

Again, without loss of generality, we can follow the same logic to come up with the full conditional distribution for $1/\sigma_2^2$, which will end up being,

$$
\begin{aligned}
1/\sigma_2^2|\vec{Y},\vec{X},p,\theta_1,\theta_2,\sigma_1^2 &\sim Gamma\left(\frac{\nu_{n,2}}{2}, \frac{\sigma_{n_2}^2 \nu_{n,2}}{2} \right), \text{ where} \\
\nu_{n,2} &= \nu_0 + n_2, \text{ and,} \\
\sigma_{n,2}^2 &= \frac{1}{\nu_{n,2}} \left[\nu_0 \sigma_0^2 + n_2 s_{n_2}^2(\theta_2) \right], \text{ where,} \\
s_{n,2}^2(\theta_2) &= \sum (y_i - \theta_2)^2/n_2.
\end{aligned}
$$

Gibb's Sampler

I implemented a Gibb's Sampler next with values of $a = b = 1$, because I didn't really know anything about the probability of people from the low mean distribution versus the high mean distribution, $\mu_0 = 140$, because it was the mean of the entire set of y values, $\sigma_0^2 = 625$, $\tau_0^2 = 625$, because they seemed like reasonable numbers for the variances, and $\nu_0 = 5$. I then used the full conditional distributions from above and ran the Gibb's Sampler using 100,000 iterations.

Diagnostics

Looking to Figure 3 for the means and 95% Confidence Intervals for all of the parameters of interest, our initial guesses for the θ_j values look to be pretty close, but our original guesses for the σ_j^2 weren't that close at all. It also appears as if about 46.7% of people come from the distribution with smaller mean, and the rest are from the larger distribution.

Parameter	2.5%	50%	97.5%
θ_1	108.87	11.79	112.70
θ_2	143.10	146.43	149.12
σ_1^2	106.65	129.55	158.89
σ_2^2	187.83	236.89	306.44
p	0.393	0.467	0.529

Figure 2: Means and 95% Confidence Intervals

From both Figure 2 and Figure 3, one can see that θ_2 has a wider spread in θ_2 when compared to θ_1. This is another consequence of there being a significant amount of extreme values one the high end of the original data set, making θ_2 vary more.

According to the Gibb's Sampler, if we take a sample of $\tilde{Y}$ from the corresponding θ_j and σ_j, we get a similar looking kernel density which can be seen in Figure 4. The shape of the distribution with the smaller mean from the original sample looks pretty similar to the mixture model, however the distribution

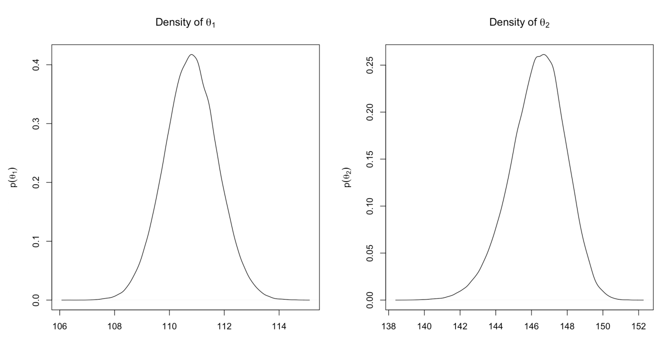

Figure 3: Posterior Density of θ_1 and θ_2

with the larger mean doesn't quite match up. That's because when we run the Gibb's Sampler, $\sigma_2 = 15.45$, which is quite high. I think this happens because there are a couple extreme values on the high end of the original sample which might be influencing the variance of the higher mean distribution.

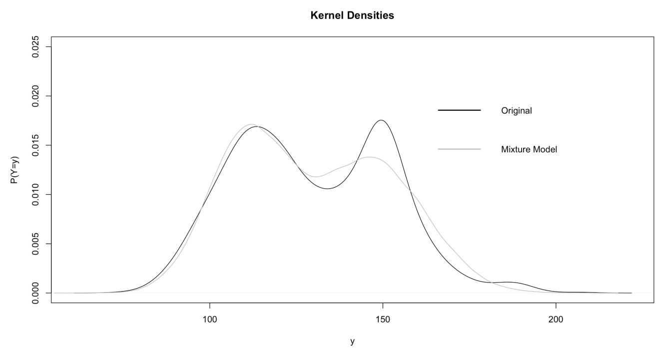

Figure 4: Kernel Densities

I also found that, given an individual whose y value is 120 has a 78% chance of coming from the distribution with with a smaller mean.

Looking at the auto correlation functions in Figure 5, we can see that the θ_j values seem to be quite correlated, but as time goes on, they eventually become uncorrelated. This is why I did such a large number of iterations with the Gibb's Smapler, so I could get large enough effective samples sizes for both of the θ_j values. Those effective sample sizes were 5111.7 for θ_1 and 3952.5 for θ_2.

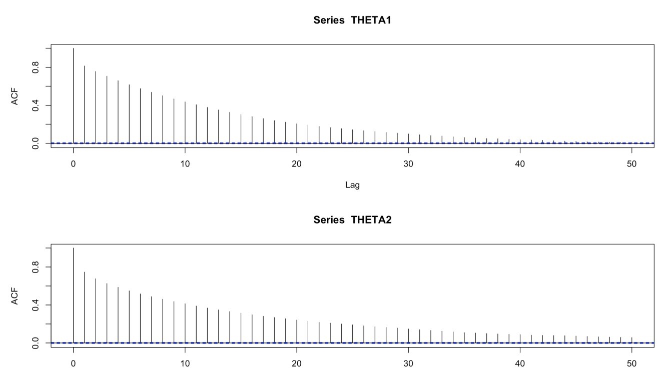

Figure 5: Kernel Densities