Estimating Parameters of a Mixture Model

Zak Domingo

May 2019

1 Introduction

In this analysis, we seek to develop a model that accurately and reliably describes the given dataset. Specifically, our goals are to find parameters of the distributions from the data, find proportions of sub-populations that make up the dataset, and use diagnostic tools to examine the performance of the model. We will be using Bayesian techniques to accomplish the above.

2 Initial Analysis

Within this population, there is reason to believe that there exists two sub-populations. We begin by creating a histogram with a kernel density estimate overlay of the data to see if this is the case.

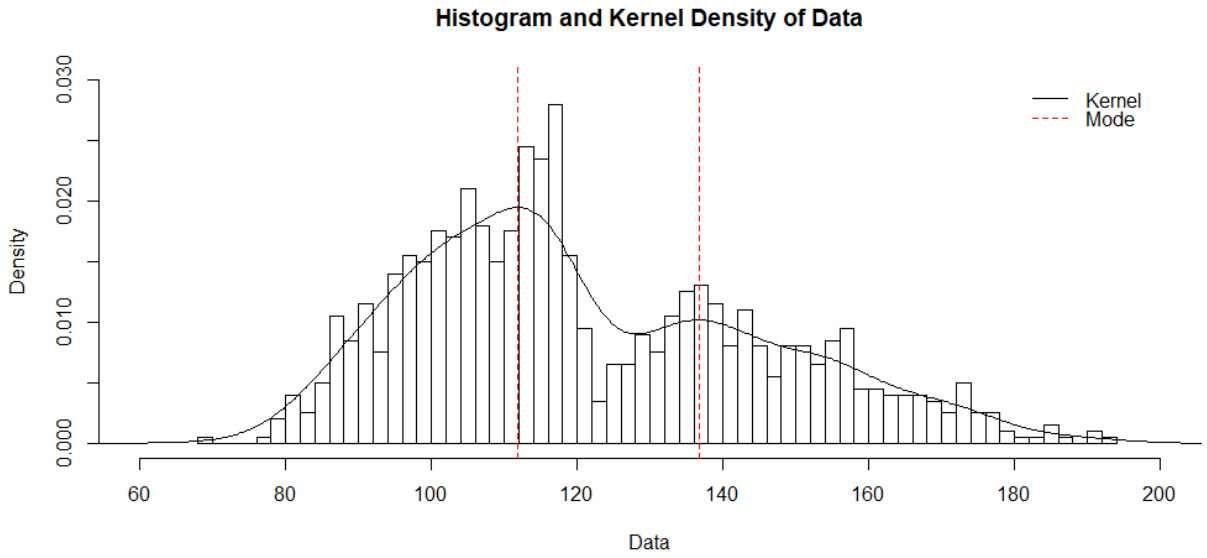

Figure 1: Histogram and kernel density showing a bimodal distribution

Figure 1 shows a right-skewed distribution that appears to be bimodal, giving more credence to our suspicions that there are two sub-populations. Furthermore, Anderson-Darling and Shapiro-Wilk normality tests both yield p-values less than 0.0001 suggesting that the data does not follow a normal distribution. It is plausible that the data comes from a mixture of two normal distributions with different parameters. The kernel density seems to show one sub-population with a lower mean and lower variance than the other. To get a better grasp of the sub-population parameters, Figure 2 shows a scatterplot of the observations. The scatterplot indicates where we partitioned the sample (at 123) and where the sample means of the sub-populations sit ($\hat{\mu}_1 = 104.96, \hat{\mu}_2 = 146.88$). The sample variances for each sub-population are $s_1^2 = 114.166$

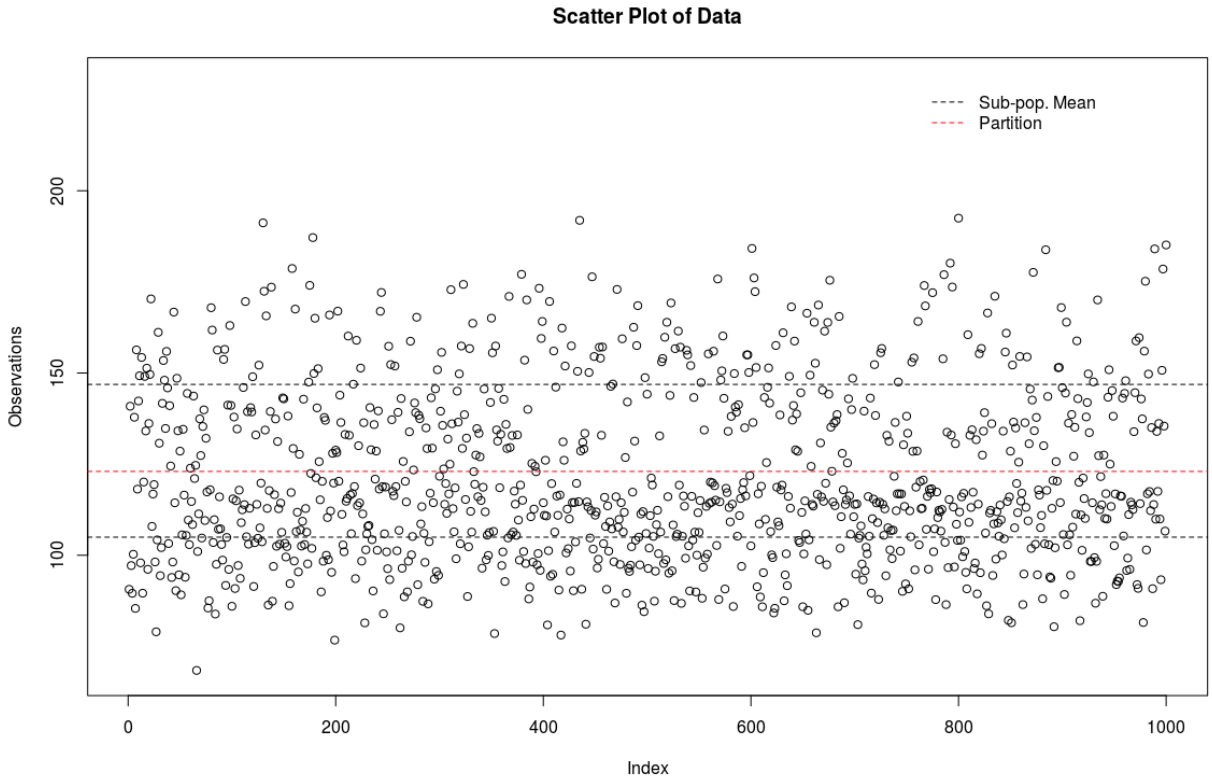

Figure 2: Scatter plot of data with partition and sub-population means

and $s_2^2 = 228.285$ with sub-population 1 making up 61.1% of the sample. We then proceed with developing a mixture model.

3 Model Selection

We will make use of the Gibbs sampling algorithm to determine the full posterior conditional distribution of each parameter of interest. Those parameters are θ_1, θ_2, σ_1^2, σ_2^2, p, where $Y_1 \sim N(\theta_1, \sigma_1^2)$, $Y_2 \sim N(\theta_2, \sigma_2^2)$ and $p \sim \text{Beta}(a, b)$.

3.1 Prior Specification

Based on the preliminary information, we believe this population is partitioned into two sub-populations. We are given that this proportion follows a beta distribution. We will select parameters $a = b = 1$ for the prior distribution, reflecting the fact that the membership variable X_i was unobserved before data collection and therefore do not know the relative proportions of the sub-populations.

To use the normal model given by Hoff (p.71) we will select prior parameters $\mu_0 = \bar{y} = 121.2695$, $\tau_0^2 = 1500$ to give us a relatively diffuse prior distribution. Although this selection for μ_0 does not seem congruent with a typical Bayesian framework, we lack preliminary information to help us determine an alternative value. According to the normal model, the prior distribution of σ_i^2 is inverse-gamma and we select the parameters $\nu_0 = 1$ and $\sigma_0^2 = Var(\boldsymbol{y}) = 576.3868$. Again, this selection of σ_0^2 is not congruent with a typical Bayesian framework, but the selection of $\nu_0 = 1$ allows us to impart as little as possible information into the posterior distribution of σ_i^2.

3.2 Sampling Distribution Specifications

We are given that each sub-population follows a normal distribution with different parameters. In order to incorporate the unobserved membership variable X_i, the sampling distribution of $p(\theta_i \mid \ldots)$ and $p(\sigma_i^2 \mid \ldots)$ takes the form

$$\prod_{i=1}^{n} \left(\text{dnorm}(y_i, \theta_1, \sigma_1^2)^{1-x_i} \cdot \text{dnorm}(y_i, \theta_2, \sigma_2^2)^{x_i}\right) \tag{1}$$

where $X_i = 0$ or 1 if the y_i observation is from sub-population 1 or 2 respectively. We will treat X_i as a Bernoulli(p_1) random variable where

$$p_1 = P(X_i = 0 \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, y_i, p) \tag{2}$$

$$= \frac{P(X_i = 0 \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p) \cdot p(y_i \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p, X_i = 0)}{P(y_i \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, p)} \tag{3}$$

$$= \frac{p \cdot \text{dnorm}(y_i, \theta_1, \sigma_1^2)}{p \cdot \text{dnorm}(y_i, \theta_1, \sigma_1^2) + (1-p) \cdot \text{dnorm}(y_i, \theta_2, \sigma_2^2)}. \tag{4}$$

Then $p_2 = 1 - p_1$. The sampling distribution of $p(p_i \mid \ldots)$ is Binomial(n,p_i) and we will pass the Binomial pmf either n_1 or n_2, the sample size of the sub-populations.

3.3 Full Conditional Distributions

The full conditional distribution of θ_1 is given by

$$p(\theta_1 \mid \sigma_1^2, \sigma_2^2, \theta_2, \boldsymbol{x}, \boldsymbol{y}, p) \propto p(\theta_1 \mid \sigma_1^2, \sigma_2^2, \theta_2, \boldsymbol{x}, p) \cdot p(\boldsymbol{y} \mid \sigma_1^2, \sigma_2^2, \theta_1, \theta_2, \boldsymbol{x}, p) \tag{5}$$

$$\propto p(\theta_1) \cdot \prod_{i=1}^{n} \left((\sigma_1)^{-1} \cdot \exp\left(\frac{-1}{2}\left(\frac{y_i - \theta_1}{\sigma_1}\right)^2\right)\right)^{x_i} \tag{6}$$

$$\cdot \left((\sigma_2)^{-1} \cdot \exp\left(\frac{-1}{2}\left(\frac{y_i - \theta_2}{\sigma_2}\right)^2\right)\right)^{1-x_i} \tag{7}$$

$$\propto \text{dnorm}(\theta_1, \mu_0, \tau_0^2) \cdot \prod_{i=1}^{n} \text{dnorm}(y_i, \theta_1, \sigma_1^2)^{x_i} \tag{8}$$

$$\propto \text{dnorm}(\theta_1, \mu_0, \tau_0^2) \cdot \prod_{y \in \boldsymbol{y}_1} \text{dnorm}(y, \theta_1, \sigma_1^2) \tag{9}$$

$$\propto \exp\Big(-\frac{1}{2\tau_0^2}(\theta_1 - \mu_0)^2\Big) \cdot \prod_{y \in \boldsymbol{y}_1} \exp\Big(-\frac{1}{2\sigma_1^2}(y - \theta_1)^2\Big) \tag{10}$$

$$\propto \exp\Big(\frac{1}{2\tau_0^2}(\theta_1 - \mu_0)^2\Big) \cdot \exp\Big(-\frac{1}{2\sigma_1^2}\sum_{y \in \boldsymbol{y}_1}(y - \theta_1)^2\Big) \tag{11}$$

$$\propto \text{See Hoff 70 for further details} \tag{12}$$

$$\propto N(\mu_{n(1)}, \tau_{n(1)}^2) \tag{13}$$

where

$$\tau^2_{n(1)} = \left((\tau^2_0)^{-1} + \frac{n_1}{\sigma^2_1}\right)^{-1} \tag{14}$$

$$\mu_{n(1)} = \frac{\mu_0(\tau^2_0)^{-1} + \frac{n_1}{\sigma^2_1}\bar{y}_1}{(\tau^2_0)^{-1} + \frac{n_1}{\sigma^2_1}}. \tag{15}$$

The full conditional of θ_2 is the same but with subscript 2 in place of 1.

The full conditional distribution of σ^2_1 is given by

$$p(\sigma^2_1 \mid \theta_1, \theta_2, \sigma^2_1, \boldsymbol{x}, \boldsymbol{y}, p) \propto p(\sigma^2_1 \mid \theta_1, \theta_2, \sigma^2_1, \boldsymbol{x}, p) \cdot p(\boldsymbol{y} \mid \theta_1, \theta_2, \sigma^2_1, \sigma^2_2, \boldsymbol{x}, p) \tag{16}$$

$$\propto p(\sigma^2_1) \cdot \prod_{i=1}^{n} \left((\sigma_1)^{-1} \cdot \exp\left(\frac{-1}{2}\left(\frac{y_i - \theta_1}{\sigma_1}\right)^2\right)\right)^{x_i} \tag{17}$$

$$\cdot \left((\sigma_2)^{-1} \cdot \exp\left(\frac{-1}{2}\left(\frac{y_i - \theta_2}{\sigma_2}\right)^2\right)\right)^{1-x_i} \tag{18}$$

$$\propto \text{Inv-Gamma}(\sigma^2_1, \nu_0, \sigma^2_0\nu_0/2) \cdot \prod_{y \in \boldsymbol{y}_1} \text{dnorm}(y, \theta_1, \sigma^2_1) \tag{19}$$

$$\propto \exp\left((\sigma^2_1)^{-(\nu_0/2)-1} \exp\left(-\frac{1}{\sigma^2_1}\sigma^2_0\nu_0/2\right)\right) \tag{20}$$

$$\cdot (\sigma^2_1)^{-n/2} \exp\left(-\frac{1}{2\sigma^2_1}\sum_{y \in \boldsymbol{y}_1}(y - \theta_1)^2\right) \tag{21}$$

$$\propto \text{See Hoff 93 for further details} \tag{22}$$

$$\propto \text{Inv-Gamma}(\nu_{n(1)}/2, \sigma^2_{n(1)}(\theta_1)\nu_{n(1)}/2) \tag{23}$$

where

$$\nu_{n(1)} = \nu_0 + n_1 \tag{24}$$

$$\sigma^2_{n(1)}(\theta_1) = \nu_n^{-1}\left(\nu_0\sigma^2_0 + n s^2_{n(1)}(\theta_1)\right) \tag{25}$$

$$s^2_{n(1)}(\theta_1) = \sum_{y \in \boldsymbol{y}_1}(y - \theta_1)^2/n_1 \tag{26}$$

$$\tag{27}$$

The full conditional of σ^2_2 is the same but with subscript 2 in place of 1.
Full conditional distribution of p:

$$p(p \mid \theta_1, \theta_2, \sigma^2_1, \sigma^2_2, \boldsymbol{x}, \boldsymbol{y}) \propto p(p) \cdot p(\theta_1, \theta_2, \sigma^2_1, \sigma^2_2, \boldsymbol{x}, \boldsymbol{y}, \mid p) \tag{28}$$

$$\propto p(p) \cdot p(\boldsymbol{x} \mid p) \cdot p(\boldsymbol{y} \mid \theta_1, \theta_2, \sigma^2_1, \sigma^2_2, \boldsymbol{x}) \cdot p(\theta_1, \theta_2, \sigma^2_1, \sigma^2_2) \tag{29}$$

$$\propto p(p) \cdot p(\boldsymbol{x} \mid p) \tag{30}$$

$$\propto \text{dbeta}(p, a, b) \cdot \text{dbinom}(n_1, n, p) \tag{31}$$

$$\propto p^{a-1}(1-p)^{b-1} \cdot p^{n_1}(1-p)^{n_2} \tag{32}$$

$$= p^{a+n_1-1}(1-p)^{b+n_2-1} \tag{33}$$

$$= \text{beta}(a + n_1, b + n_2) \tag{34}$$

Now that prior, sampling, and full conditional posterior distributions are determined we can proceed with the Gibbs sampler. The Gibbs sampling algorithm proceeds as follows:

1. sample $p_n \sim \text{beta}(a + n_1,\, b + n_2)$
2. sample $\theta_1^{(s+1)} \sim p(\theta_1 \mid \boldsymbol{x},\, \boldsymbol{y},\, p,\, \theta_2,\, \sigma_1^{2(s)},\, \sigma_2^{2(s)})$
3. sample $\theta_2^{(s+1)} \sim p(\theta_2 \mid \boldsymbol{x},\, \boldsymbol{y},\, p,\, \theta_1,\, \sigma_1^{2(s)},\, \sigma_2^{2(s)})$
4. sample $\sigma_1^{2(s)} \sim p(\sigma_1^2 \mid \boldsymbol{x},\, \boldsymbol{y},\, p,\, \sigma_2^{2(s)},\, \theta_1^{(s+1)},\, \theta_2^{(s+1)})$
5. sample $\sigma_2^{2(s)} \sim p(\sigma_2^2 \mid \boldsymbol{x},\, \boldsymbol{y},\, p,\, \sigma_1^{2(s)},\, \theta_1^{(s+1)},\, \theta_2^{(s+1)})$
6. let $\phi^{(s+1)} = \{\theta_1^{(s+1)},\, \theta_2^{(s+1)},\, \tilde{\sigma}_1^{2(s+1)},\, \tilde{\sigma}_2^{2(s+1)},\, p^{(s+1)}\}$

We iterate this procedure 10,000 times.

4 Model Results

The expected values of the parameters and their highest posterior density intervals are listed below:

- $\theta_1 = 105.5491$
  95% HD Interval = (103.8240, 107.2913)
- $\sigma_1 = 11.9817$
  95% HD Interval = (10.8233, 13.0918)
- $\theta_2 = 143.2256$
  95% HD Interval = (137.7766, 148.1031)
- $\sigma_2 = 18.7384$
  95% HD Interval = (15.8090, 21.6566)
- $p = 0.5803$
  95% HD Interval = (0.4927, 0.6546)

It appears that about 58% of observations are from sub-population 1, and 42% are from sub-population 2.

5 Diagnostics

First we will see if the Gibbs sampler had enough iterations. Examining the below traceplots for θ_1 and σ_1^2 reveal homoscedasticity, indicating the Gibbs sampler had sufficient amount of iterations.

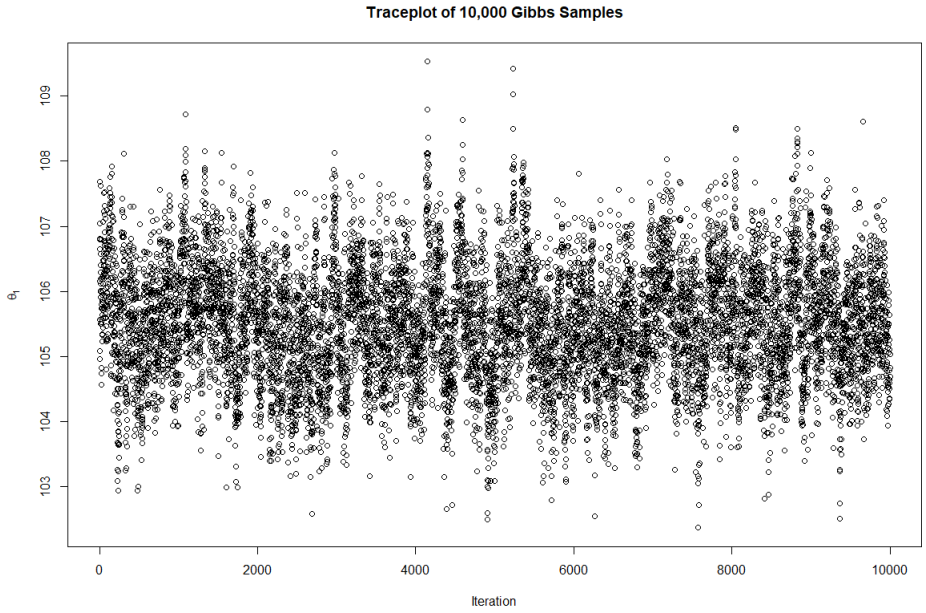

Figure 3: Traceplot of Gibbs sampler for θ_1

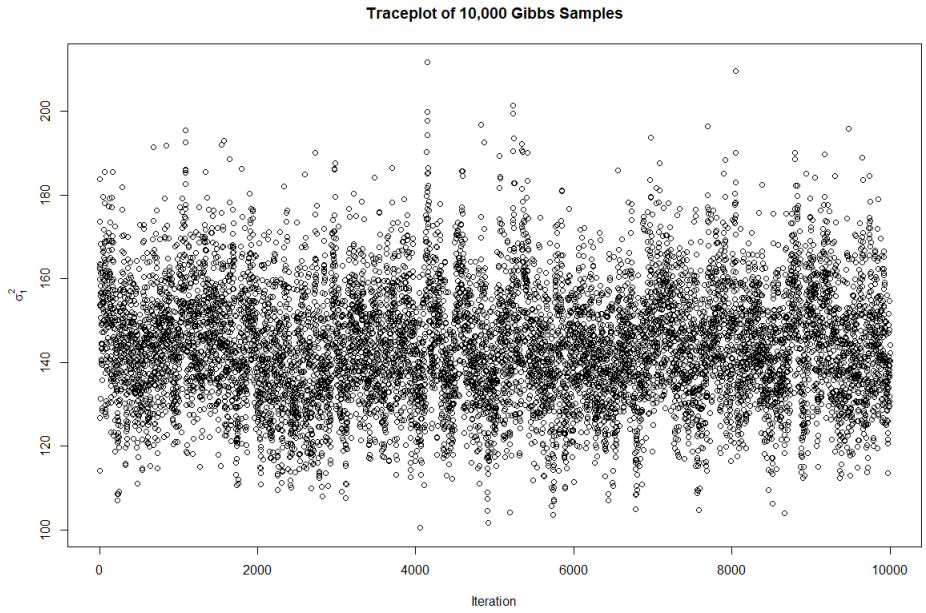

Figure 4: Traceplot of Gibbs sampler for σ_1^2

We then compare the observations to the predictions made at each iteration of the Gibbs sampler, shown in the plot below.

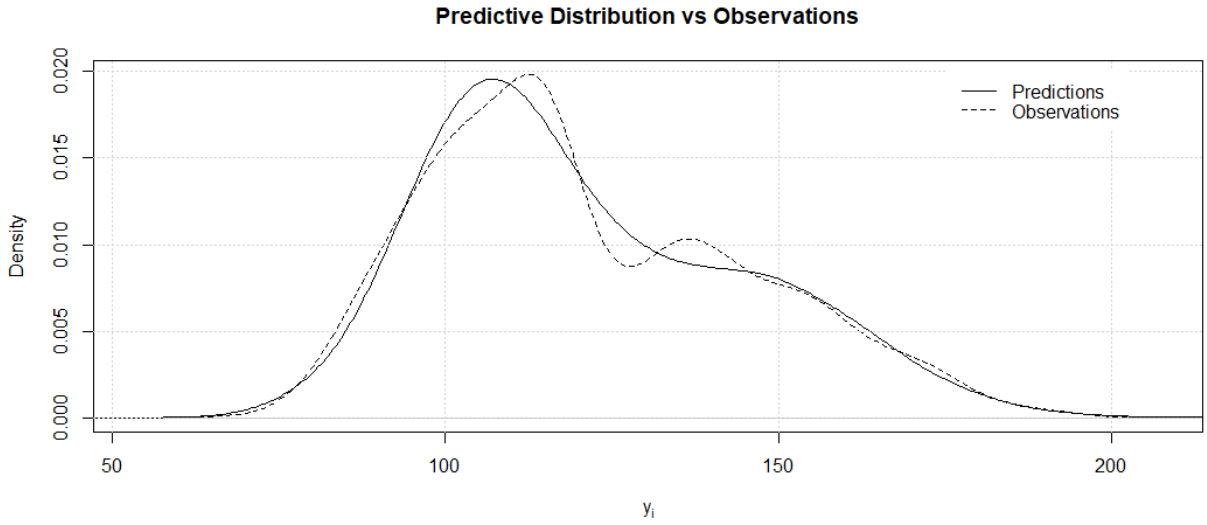

Figure 5: Kernel Distributions of Observations and Predictions

The plot reveals that the mixture model approximated the observed distribution fairly well. We then wish to compute $P(y_i \text{ from group } 1 \mid y_i = 120)$. This can be found by sending the single value $y_i = 120$ through Gibbs sampler with the parameters we previously found. This yields a probability of 0.3166.

5.1 Prior Sensitivity

We execute the Gibbs sampler under varying prior conditions to test the sensitivity of the mixture model to the prior parameter selections. First we will decrease the τ_0^2 value to 100, holding all other parameters the same. This yields

- $\theta_1 = 105.2655, \qquad \sigma_1 = 11.8425$
- $\theta_2 = 141.7426, \qquad \sigma_2 = 19.50112$
- $p = 0.5595.$

Next, we will let $\mu_0 = \bar{y}_1 = 104.9641$ and $\sigma_0^2 = s_1^2 = 114.166$, giving

- $\theta_1 = 105.3674, \qquad \sigma_1 = 11.8567$
- $\theta_2 = 142.8469, \qquad \sigma_2 = 18.9036$
- $p = 0.5734.$

Lastly, we let $\mu_0 = \bar{y}_2 = 146.8803$ and $\sigma_0^2 = s_1^2 2 = 228.2855$. This gives us

- $\theta_1 = 105.5444, \qquad \sigma_1 = 11.9554$
- $\theta_2 = 143.3608, \qquad \sigma_2 = 18.6719$
- $p = 0.5824.$

We can see the results are nearly identical under different prior parameter selections, indicting that this model is not very sensitive to different prior parameters.

6 Conclusion

Using a Gibbs sampler to find parameters of a mixture model yields reasonable results, given that the sampler has enough iterations. We observe that the model we developed is quite robust to varying prior parameter choices and appears to be adequate in describing this dataset.

7 References

1. Hoff, Peter D. *First Course in Bayesian Statistical Methods.* Springer, 2010.