PHP 2530: BAYESIAN STATISTICAL METHODS
HOMEWORK IV SOLUTIONS

NICK LEWIS

Problem 1

The Importance Sampling Algorithm

a. Write a function which implements the importance sampling algorithm with the following five arguments:

(1) logTargetDensityFunc: the natural logarithm of the target density, as a function of one argument.
(2) logProposalDensityFunc: the natural logarithm of the proposal density, as a function of one argument.
(3) proposalNewFunc: a function of zero arguments that is equivalent to a draw from the proposal density.
(4) nSamples: the number of samples one wishes to draw.
(5) rejectionControlConstant: An argument whose default is NULL. This model parameter can only take positive values and is used to to make the rejection sampling more restrictive.

If rejectionControlConstant = NULL, this function should return the natural logarithm of the importance sampling weights as well as the Effective Sample Size (ESS). Otherwise when $rejectionControlConstant \neq NULL$, this function should return the number of accepted samples, the natural logarithm of the importance weights, the acceptance rate of the samples along with the ESS.

Solution:

The code can be found in the appendix. The algorithm to perform the rejection constant control is described below:

(1) if $rejectionControlConstant \neq NULL$, then proceed with the following set up. First, let $c = rejectionControlConstant$.
(2) Draw θ from your proposal distribution g(.)
(3) Calculate the weight $w(\theta) = \frac{f(\theta)}{g(\theta)}$. Remember f is your target distribution.
(4) Calculate $r = \min\{1, \frac{w(\theta)}{c}\}$. Draw $u \sim Unif(0, 1)$. If $r > u$ accept the draw. Otherwise reject and try again.
(5) If θ is accepted, modify the weight associated with θ by calculating $\hat{w}(\theta) = \frac{w(\theta)}{r}$.
(6) Using the modified weights calculate the Effective Sample Size (ESS).

NOTE: The normalized weights are returned instead of the weights.

b). Now consider the following mixture of two Normals set up in one dimension:

$$f(x) = \frac{1}{3}Normal(x \mid \mu = -2, \sigma^2 = 1^2) + \frac{2}{3}Normal(x \mid \mu = 2, \sigma^2 = 1^2)$$
$$g(x) = Normal(x \mid \mu = 0, \sigma^2 = 3^2)$$

Date: September 10, 2022.

c). Plot the two density functions f(x) and g(x) on the same plot. Take a look at it and observe that g(x) is a reasonable importance density function.
Solution:

The plot of the densities together is shown below:

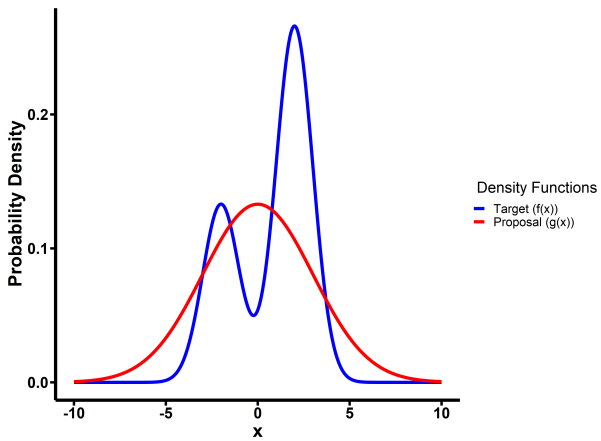

Figure 1. Probability Density Functions of the target distribution (Gaussian Mixture), and the proposal distribution (Gaussian)

The proposal distribution reasonably captures areas of high density for the target distribution. We can tell this since the range of the mixture is almost completely taken over by the normal proposal (look at areas of "zero" probability for the mixture and how they are contained within the unimodal normal).

d). Let $X \sim f(.)$. Theoretically compute

(1) $\mu_1 = E(X)$
(2) $\mu_2 = E(X^2)$
(3) $\theta = E(e^X)$

Solution:

All of these can be computed relatively straightforward. For simplicity call $Y_1 \sim N(-2, 1^2)$ and $Y_2 \sim N(2, 1^2)$. You will see why this is important in just a second.

$$
\begin{aligned}
\mu_1 = E(X) &= \int_{-\infty}^{\infty} x(\frac{1}{3}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+2)^2}{2}} + \frac{2}{3}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}})dx \\
&= \int_{-\infty}^{\infty} x\frac{1}{3}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+2)^2}{2}}dx + \int_{-\infty}^{\infty} x\frac{2}{3}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}}dx \\
&= \frac{1}{3}E(Y_1) + \frac{2}{3}E(Y_2) \\
&= \frac{-2}{3} + \frac{4}{3} \\
&= \frac{2}{3}
\end{aligned}
$$

Essentially our expected value for our gaussian mixture can be split into two integrals. Each integral has a gaussian distribution attached to it times some constant for the weights. We can do a similar trick for μ_2.

$$
\begin{aligned}
\mu_2 = E(X^2) &= \int_{-\infty}^{\infty} x^2 (\frac{1}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+2)^2}{2}} + \frac{2}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}}) dx \\
&= \int_{-\infty}^{\infty} x^2 \frac{1}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+2)^2}{2}} dx + \int_{-\infty}^{\infty} x^2 \frac{2}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} dx \\
&= \frac{1}{3}E(Y_1^2) + \frac{2}{3}E(Y_2^2) \\
&= \frac{1}{3}(Var(Y_1) + E(Y_1)^2) + \frac{2}{3}(Var(Y_2) + E(Y_2)^2) \\
&= \frac{1}{3}(1 + (-2)^2) + \frac{2}{3}(1 + 2^2) \\
&= \frac{5}{3} + \frac{10}{3} \\
&= 5
\end{aligned}
$$

Recall the Moment Generating function (MGF) for a continuous random variable $X \sim f(.)$ is defined as:

$$
MGF_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx
$$

For a gaussian random variable $V \sim N(\mu, \sigma^2)$, $MGF_V(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$. θ is essentially the Moment generating function at $t = 1$

$$
\begin{aligned}
\theta = E(e^X) &= \int_{-\infty}^{\infty} e^x (\frac{1}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+2)^2}{2}} + \frac{2}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}}) dx \\
&= \int_{-\infty}^{\infty} e^x \frac{1}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+2)^2}{2}} dx + \int_{-\infty}^{\infty} e^x \frac{2}{3}\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} dx \\
&= \frac{1}{3}E(e^{Y_1}) + \frac{2}{3}E(e^{Y_2}) \\
&= MGF_{Y_1}(1) + MGF_{Y_2}(1) \\
&= \frac{1}{3}e^{-\frac{3}{2}} + \frac{2}{3}e^{\frac{5}{2}} \\
&\approx 8.196
\end{aligned}
$$

e). Use the function you wrote ImpSampler( ), get the samples and the weights, compute the weighted average of the relevant h(.) functions and thus estimate μ_1, μ_2 and θ. Use m = 5000 samples without any rejection control and report the estimated ESS from your runs.
Solution:

The relevant functions to calculate expected values for are the ones that are above so $h_1(x) = x$, $h_2(x) = x^2$ and $h_3(x) = e^x$. Using our Importance Sampler without any rejection control, we report the following estimates for the expected values of those functions.

	μ_1	μ_2	θ	ESS
Estimated Values	0.658	5.028	8.311	3224
True Values	0.667	5.0	8.196	

As one can observe the Importance Sampling in this scenario produces highly accurate estimates.

f).Now vary the rejection control constant for $c \in \{1, 2, \ldots, 10\}$. For each such c, as in the previous part, use the function you wrote ImpSampler( ) and estimate μ_1, μ_2 and θ from m = 5000 samples with rejection control. For each c, report the estimated ESS (denoted by e_c), and the acceptance rate (denoted by a_c) of the samples from your runs. Let the error in the estimates be $|\hat{\mu}_1 - \mu_1|$, $|\hat{\mu}_2 - \mu_2|$ and $|\hat{\theta} - \theta|$. Finally, plot the three errors, ESS and acceptance rate against the rejection control constant (i.e. plot (c,$|\hat{\mu}_1 - \mu_1|$), (c,$|\hat{\mu}_2 - \mu_2|$), (c,$|\hat{\theta} - \theta|$), (c,e_c) and (c,a_c)). Now share your observations on the error in the estimators, acceptance rate, ESS, recommendations on the choice of c.

Solution:

	$\lvert\hat{\mu}_1 - \mu_1\rvert$	$\lvert\hat{\mu}_2 - \mu_2\rvert$	$\lvert\hat{\theta} - \theta\rvert$	e_c	a_c
$c = 1$	0.002	0.017	0.027	3019	0.6912
$c = 2$	0.022	0.130	0.298	2333	0.4708
$c = 3$	0.008	0.053	0.033	1669	0.3336
$c = 4$	0.022	0.076	0.370	1218	0.2436
$c = 5$	0.045	0.069	0.075	1022	0.2044
$c = 6$	0.006	0.096	0.288	845	0.1690
$c = 7$	0.035	0.172	0.687	726	0.1452
$c = 8$	0.052	0.007	0.137	584	0.1168
$c = 9$	0.014	0.097	0.028	538	0.1076
$c = 10$	0.026	0.257	0.529	481	0.0962

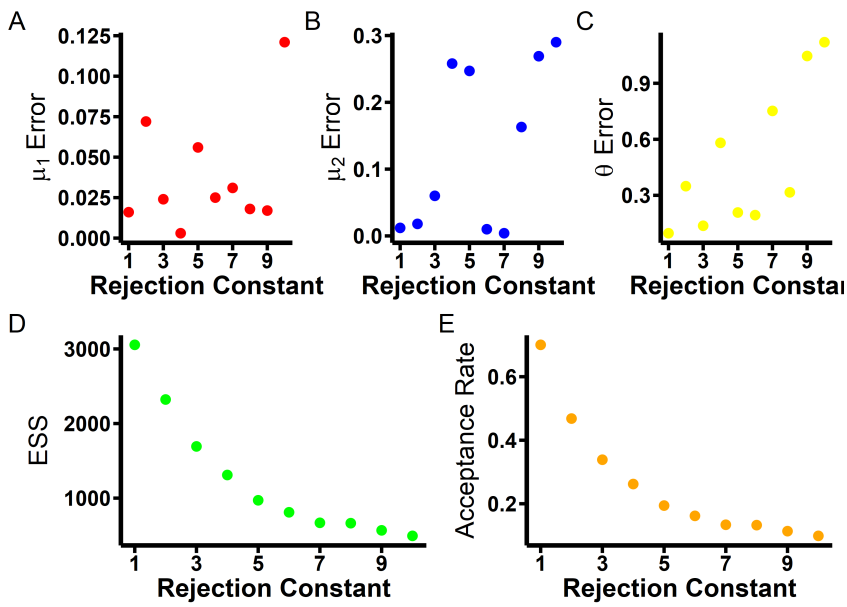

Figure 2. (A) $|\hat{\mu}_1 - \mu_1|$, (B) $|\hat{\mu}_2 - \mu_2|$, (C) $|\hat{\theta} - \theta|$, (D) Effective Sample Sizes, and (E) Acceptance Rate

Looking at these five plots, my recommendation would be to choose a rejection constant of either $c = 3$ or $c = 5$

Problem 2: BDA 3rd Ed. Exercise 10.5

Rejection sampling and importance sampling: Consider the model, $y_j \sim Binomial(n_j, \theta_j)$, where $\theta_j = logit^{-1}(\alpha + \beta x_j)$, for $j = 1, \ldots, J$, and with independent prior distributions, $\alpha \sim t_4(0, 22)$ and $\beta \sim t_4(0, 1)$. Suppose $J = 10$, the x_j values are randomly drawn from a

$Uniform(0,1)$ distribution, and $n_j Poisson^+(5)$, where $Poisson^+$ is the Poisson distribution restricted to positive values.

(a) Sample a dataset at random from the model.

Solution:

The procedure to sample a random dataset is relatively straightforward. Firstly, we need to set a seed for reproducibility. Then proceed as follows:

(1) draw J samples of $x \sim Uniform(0,1)$
(2) Draw J samples of $n \sim Poisson^+(5)$. This can be done by sampling from a $Poisson(\lambda = 5)$ and shifting the sample to the right by 1.
(3) Draw one sample of $\alpha \sim t_4(0, 2^2)$ and one sample of $\beta \sim t_4(0, 1^2)$. Let t_n denote the standard Student's t distribution with n degrees of freedom. Then for a Generalized t distribution $t_n(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$, $t_n(\mu, \sigma^2) = \mu + \sigma t_n$. This means we can draw values of α and β by just drawing samples $t_4^{(1)}$, $t_4^{(2)}$ from t_4 and calculating $\alpha = 2t_4^{(1)}$ and $\beta = t_4^{(2)}$
(4) With (α, β), calculate $\theta_j = \frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}$ for $j = 1, 2, \ldots, J$.
(5) For $j = 1, 2, \ldots, J$, draw $y_j \sim Binomial(n_j, \theta_j)$

For anyone who is interested in repeating my analysis to confirm it's validity, this is the sample dataset that I happened to produce:

Table 1. Probability that School j's coaching program is better than School k's.

Data	A	B	C	D	E	F	G	H	I	J
y	0	2	2	1	2	0	0	1	0	0
n	10	9	7	7	12	7	7	6	6	5
x	0.457	0.957	0.453	0.678	0.573	0.103	0.900	0.246	0.042	0.328

The values for the parameters used to produce y are $\alpha = -1.429$ and $\beta = -1.295$

NOTE: This data only pertains to the R code. The data is different in Python, but a seed is still saved for reproducibility of the results.

(b) Use rejection sampling to get 1000 independent posterior draws from (α, β).

Solution:

To do rejection sampling, since your prior is proper, you could use that as your proposal distribution or the normal approximation.

If you take samples from your prior distribution your target distribution, aka the unnormalized posterior density, is defined as

$$q(\alpha, \beta \mid y, n, x) = p(y \mid \alpha, \beta, n, x)p(\alpha, \beta)$$

The enveloping distribution is then defined as

$$g(\alpha, \beta) = p(y \mid \hat{\alpha}, \hat{\beta}, n, x)p(\alpha, \beta)$$

where $(\hat{\alpha}, \hat{\beta})$ is the MLE of the likelihood function. $p(y \mid \alpha, \beta, n, x) \leq p(y \mid \hat{\alpha}, \hat{\beta}, n, x)$ so this distribution should work for the rejection sampling process. The MLE for this data is $(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE}) = (-2.958, 1.482)$

The normal approximation for the posterior distribution says that under certain conditions,

$$p(\theta \mid y) \approx N(\theta \mid \mu = \hat{\theta}, \Sigma = [I(\hat{\theta})]^{-1})$$

Where $\hat{\theta} = \text{argmax}_\theta\, p(\theta \mid y) = \text{argmax}_\theta\, log(p(\theta \mid y))$ is the mode of the posterior distribution. $[I(\hat{\theta})]^{-1}$ is the Inverse of the Fisher Information Matrix.
In other words, the ith row and j^{th} column of $I(\theta)$ is defined as $I(\theta)_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} log(p(\theta \mid y))$
If $\hat{\theta}$, the mode, is in the interior of parameter space, then the matrix $I(\hat{\theta})$ is positive definite. Recall that a matrix is positive definite if all of its eigenvalues are strictly positive. We care for two reasons: First a multivariate normal is only defined for positive definite matrices (i.e. look at the form. The determinant of positive definite s strictly positive). Secondly, if the inverse of the information matrix is positive definite at the mode, this implies the mode is actually the mode (i.e. the maximizer).
Both are valid proposal distributions, but since part (c) asks us to find the Normal Approximation I'm inclined to use the t priors as our proposal distribution.

To perform the Rejection Sampling to get 1000 samples we proceed as follows:

(1) draw J samples of $x \sim Uniform(0, 1)$
(2) Draw J samples of $n \sim Poisson^+(5)$. This can be done by sampling from a $Poisson(\lambda = 5)$ and shifting the sample to the right by 1.
(3) Draw one sample of $\alpha \sim t_4(0, 2^2)$ and one sample of $\beta \sim t_4(0, 1^2)$. Let t_n denote the standard Student's t distribution with n degrees of freedom. Then for a Generalized t distribution $t_n(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$, $t_n(\mu, \sigma^2) = \mu + \sigma t_n$. This means we can draw values of α and β by just drawing samples $t_4^{(1)}$, $t_4^{(2)}$ from t_4 and calculating $\alpha = 2t_4^{(1)}$ and $\beta = t_4^{(2)}$
(4) With (α, β), calculate $\theta_j = \frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}$ for $j = 1, 2, \ldots, J$.
(5) For $j = 1, 2, \ldots, J$, draw $y_j \sim Binomial(n_j, \theta_j)$

We plot the posterior distribution for our problem, along with statistics for our parameters below:

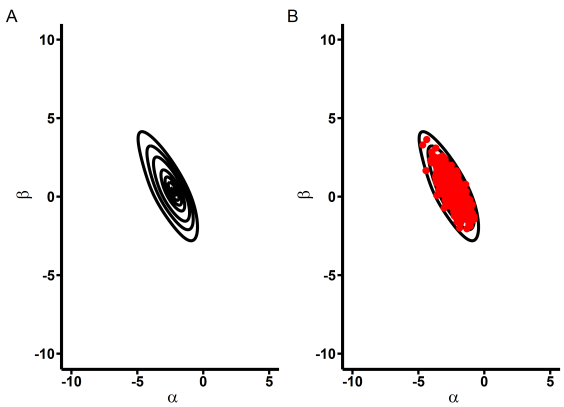

Figure 3. (A) Posterior Contours for $p(\alpha, \beta \mid x, y, n)$, (B) Posterior Draws

	Mean	Variance	2.5%	25%	50%	75%	97.5%
α	-2.363	0.586	-3.669	-2.720	-2.357	-1.965	-1.288
β	0.413	0.841	-1.194	-0.125	0.373	0.918	2.200

(c) Approximate the posterior density for (α, β) by a normal centered at the posterior mode with covariance matrix fit to the curvature at the mode

Solution:

The MAP (Maximum A Posterior, aka the mode of the posterior) and the Fisher Information Matrix for this problem are

$$\hat{\theta}_{MAP} = \begin{pmatrix} -2.231 \\ 0.317 \end{pmatrix}, [I(\hat{\theta}_{MAP})]^{-1} = \begin{pmatrix} 0.277 & -0.284 \\ -0.284 & 0.552 \end{pmatrix}.$$

We also take it upon ourselves to plot the posterior distribution against the normal approximation to the posterior distribution:

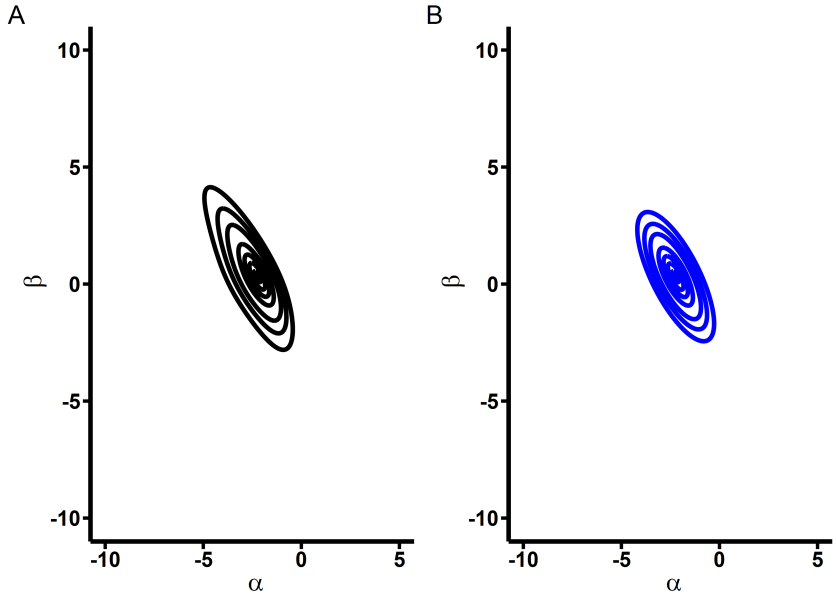

Figure 4. (A) Posterior Distribution (B) Normal Approximation to Posterior Distribution

As one can see the normal approximation

(d) Take 1000 draws from the two-dimensional t_4 distribution with that center and scale matrix and use importance sampling to estimate $E(\alpha \mid y)$ and $E(\beta \mid y)$.

Solution:

This problem is relatively simple. Using the importance sampler built in Problem 1, we can find the marginal posterior means to be

$$E(\alpha \mid y) = -2.351$$
$$E(\beta \mid y) = 0.425$$

In terms of performance the importance sampling seems to do well. Recall the mean of a multivariate normal distribution is just it's location parameter, and the values for the posterior means of α and β are extremely close to that. Likewise just calculating the means from the posterior draws of α and β appears to produce similar results.

(e) Compute an estimate of effective sample size for importance sampling using $S_{eff} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2}$, where $\tilde{w}(\theta^s) = \frac{w(\theta^s)}{\sum_{s'=1}^{S}(w(\theta^{s'})}$.

Solution:

Using the above formula we find that the Effective Sample Size for this problem is 899 samples.

Problem 3: BDA 3rd Ed. Exercise 10.8

Importance resampling with and without replacement:

(a) Consider the bioassay example introduced in Section 3.7. Use importance resampling to approximate draws from the posterior distribution of the parameters (α, β), using the normal approximation of Section 4.1 as the starting distribution. Sample $S = 10,000$ from the approximate distribution, and resample without replacement $k = 1000$ samples. Compare your simulations of (α, β) to Figure 3.3b and discuss any discrepancies.

Solution:

To see how the results differ let's look at the contour plots.

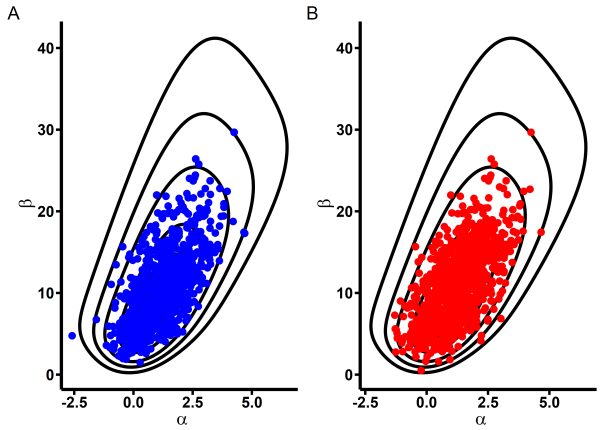

Figure 5. (A) Bioassay Posterior Contour with replacement draws, (B) Bioassay Posterior Contour without replacement draws

These plots look almost identical for the most part. Obviously each sample in the without replacement plot is unique, but for the most part sampling with replacement appears to cover most of the distribution as well.

(b) Comment on the distribution of the simulated importance ratios.

Solution:

We want to examine the distribution of the importance ratios. We can use the normalized ratios as well since thye are simply scaled by a constant. Below we plot the histogram and boxplot of the normalized importance ratios $\tilde{w}(\theta)$.

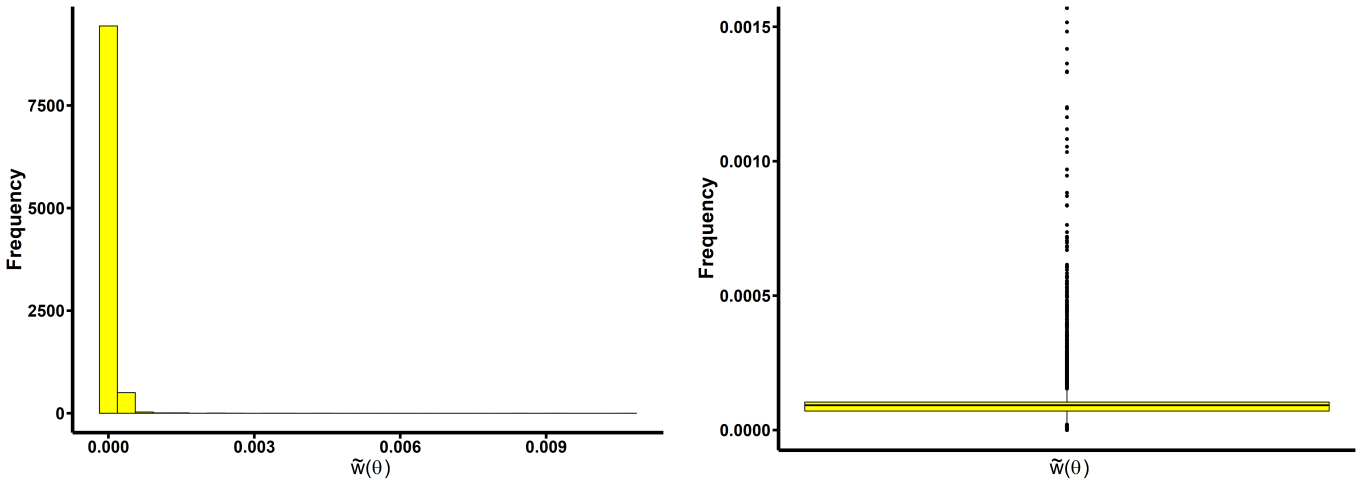

Figure 6. The left side is a histogram of the normalized importance ratios. The right side is a boxplot of the normalized importance ratios.

The histogram indicates right-skewness in the distribution of the normalized weights, and that they are highly concentrated towards zero meaning we have many small weights. We also see that our maximum normalized weight is around 0.02 meaning all the normalized weights are relatively small. This implies sampling with replacement should yield similar results to sampling without replacement since each sample has such a small probability of being chosen (recall this isn't a problem in sampling without. Once we choose a weight with a high probability, we can no longer choose it). Since the large weights are small (in this case small means far away from 1) we will more likely than not cover the entire distribution of our posterior. As for the boxplot, we see similar trends.

(c) Repeat part (a) using importance sampling with replacement. Discuss how the results differ.

Solution:

This problem is already solved for in part (a). I will add that we probably also don't see in discrepancies since we're picking 1,000 samples out of the 10,000 collected so it's possible we have enough variety here to not see any huge differences.

Problem 4: BDA 3rd Ed. 11.2

Metropolis algorithm: Replicate the computations for the bioassay example of Section 3.7 using the Metropolis algorithm. Be sure to define your starting points and your jumping rule. Compute with log-densities (see page 261). Run the simulations long enough for approximate convergence.

Solutions:

For the Bioassay example detailed in chapter 3, recall we're given the following data:

Dose, x_j log g/ml	Number of animals, n_j	Number of deaths, y_j
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

To do this problem we use the following model:

$$p(\alpha, \beta|y, n, x) \propto p(\alpha, \beta) \prod_{j=1}^{N} [logit^{-1}(\alpha + \beta x_j)]^{y_j} [1 - logit^{-1}(\alpha + \beta x_j)]^{n_j - y_j}$$
$$p(\alpha, \beta) \propto 1$$

This give us our posterior, or **target distribution**.

We choose our **jumping distribution** simply to be a normal with mean equal to the previous draw and covariance equal to the Fisher Information of the above posterior at the MAP.

$$\theta^* \mid \theta^{(t-1)} \sim N(\theta^{(t-1)}, c^2[I(\hat{\theta}_{MAP})]^{-1})$$
$$[I(\hat{\theta}_{MAP})]^{-1} = \begin{pmatrix} 0.794 & -1.089 \\ -1.089 & 1.820 \end{pmatrix}.$$

Here, $c \approx \frac{2.4}{\sqrt{d}}$ is a scaling parameter meant to improve the efficiency of the jumping distribution. d refers to the dimension of the space (i.e. in our case $d = 2$ since we have two paraemeters (α, β)).

In order to make your **Metropolis-Hastings Algorithm** more efficient I recommend the following tips:

(1) Use the log-scale. That is to say when calculating $r = \min\{\frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^*|y)/J_t(\theta^{t-1}|\theta^*)}, 1\}$, calculate $log(r) = \min\{log(p(\theta^* \mid y)/J_t(\theta^* \mid \theta^{t-1})) - log(p(\theta^* \mid y)/J_t(\theta^{t-1} \mid \theta^*)), 0\}$ instead. Then accept your sample if $log(r) > log(u)$, where $u \sim Uniform(0, 1)$ and reject otherwise.

(2) Calculate your $U \sim Uniform(0, 1)$ draws external to the loop. That is to say if you're doing S iterations, draw S samples u^s from U before doing the loop. calling the uniform is slightly faster than calculating it inside the loop.

(3) Calculate the bottom part of the ratio, $log(p(\theta^* \mid y)/J_t(\theta^{t-1} \mid \theta^*))$ before performing the loop and set the bottom part equal to the top part, $log(p(\theta^* \mid y)/J_t(\theta^* \mid \theta^{t-1}))$, if you reach an acceptance. If you accept your sample, then $\theta^t = \theta^*$ which means in your next iteration the at timestep t+1, the bottom is equal to the top from timestep t . Likewise if you reject, $\theta^t = \theta^{t-1}$ so the bottom part of the ratio remains the same until you move on. By doing this you save precious time by avoiding recalculation of the bottom ratio at each iteration. This will probably make more sense in the code.

(4) only include a condition for acceptance. If you reject, your code should automatically set $\theta^t = \theta^{t-1}$.

We run our Metropolis Hastings Chain for $S = 50,000$ iterations. We choose $S = 50,000$ iterations for a couple of reasons. Firstly, we want to give the chains time to reach the stationary distribution. Secondly, we don't want to fool ourselves. As you'll see in Problem 8 there are times where it looks like the chain has converged, but instead is just hanging out in a certain section of the parameter space. We will consider the first 60% of our draws to be warm-up samples (i.e. samples taken before we reach the stationary distribution). Other acceptable warm-up periods could be 50%, as recommended by Gelman.

Using the above hints in R we are able to run the MH Algorithm for this particular problem in about 18 seconds, and achieve an acceptance rate of 38%.

Unlike with importance sampling and rejection sampling, our samples here are correlated (autocorrelated) with each other which makes it more difficult to get posterior quantities of interest

such as the mean, variance and quantiles. There are some checks we can do to make sure convergence is occurring. We show trace plots, mean plots, autocorrelation plots and other things below:

Trace Plots

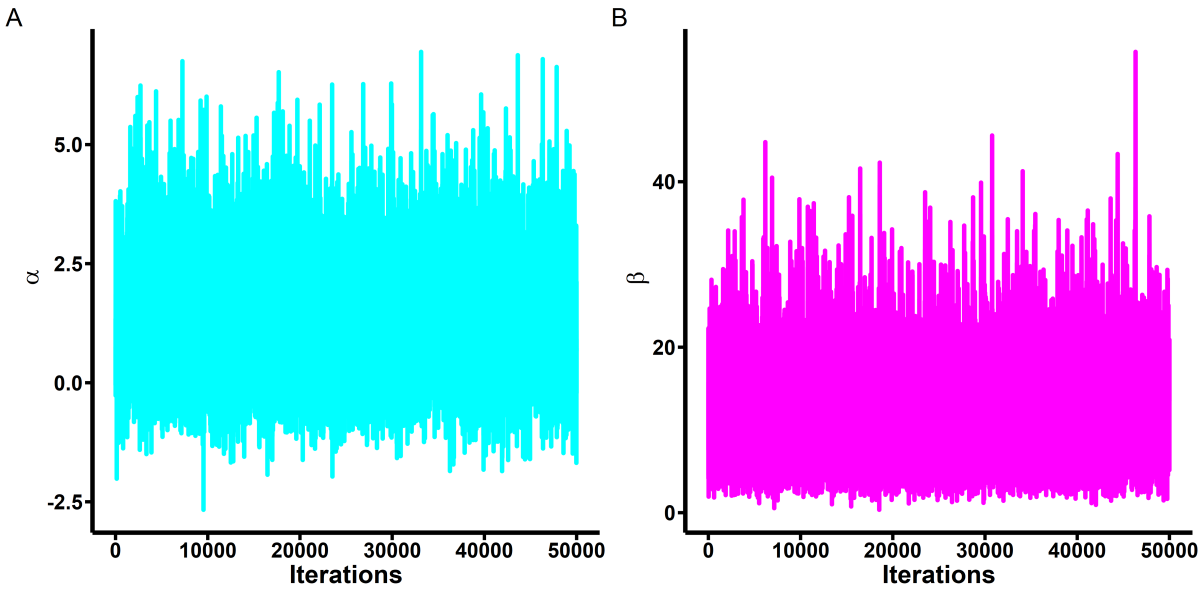

Figure 7. Trace plots for α chain (A) and β chain (B)

These trace plots actually look pretty good. When examining trace plots we want this "hairy" catepillar like shape as it indicates that our chain isn't bouncing around the distribution space too often to regions of low probability. The chain stays in areas of high probability, and occasionally bounces to low areas.

Cumulative Mean Plots

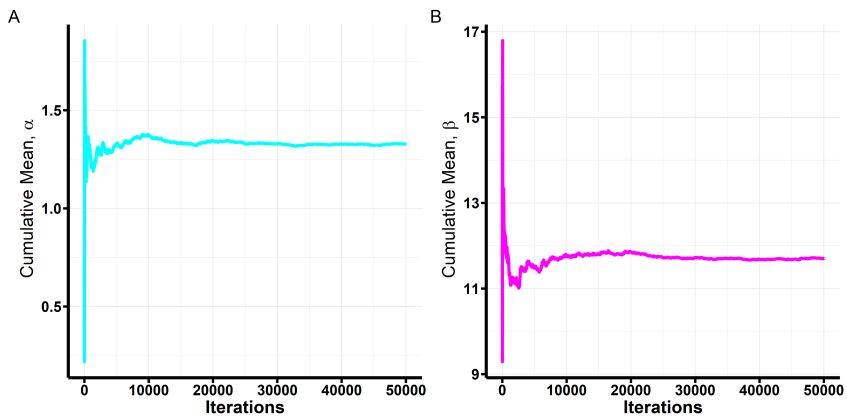

Figure 8. Cumulative Mean Plots for the α chain (A) and β chain (B)

According to the Birkhoff Ergodic Theorem, after a certain amount of iterations the cumulative average of the samples should converge to the first moment taken wrt the target distribution (i.e. $\lim_{N\to\infty} \frac{1}{N}\sum_{j=0}^{N} f(X_j) \to E_{\theta|X}(f)$). This essentially means that when we look at cumulative mean plots, we're looking for the graph to stabilize after a large amount of iterations.

Autocorrelation Plots

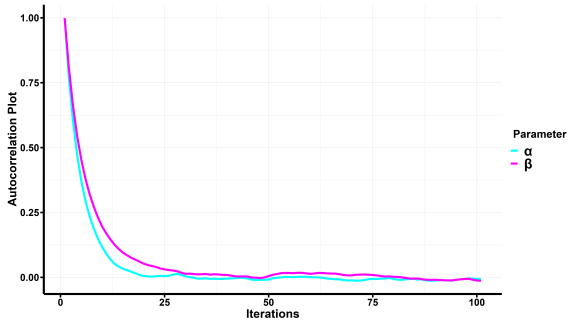

Figure 9. Autocorrelation Plots for the α chain and β chain

Another way to check for convergence is to look at the autocorrelations between the samples returned by our MCMC. The lag-k autocorrelation is the correlation between every sample and the sample k steps before. In other words the correlation between θ_n and θ_{n+k} at any time point n. Since this is a Markov chain, our draws aren't independent since the current draw is dependent on the previous draw.

As we add more and more samples we want this autocorrelation to decrease for the following reason: each sample provides us with information regarding the posterior distribution. If samples are highly correlated this will imply that previous draws are still highly influential in determining the next draw meaning that we will stay in one area of the posterior distribution and not explore the others. For this reason we desire the autocorrelation to decrease as the lag increases. The lower the autocorrelation, the less dependent samples are on previous ones meaning we'll get to explore more of the posterior space and add more information about our distribution.

Accepted Samples

We also plot our accepted samples to show not only did our algorithm work, but that our results are consistent with the results from section 3.7 in the Bayesian Data Analysis, 3rd Edition textbook. We also plot the statistics for α and β below:

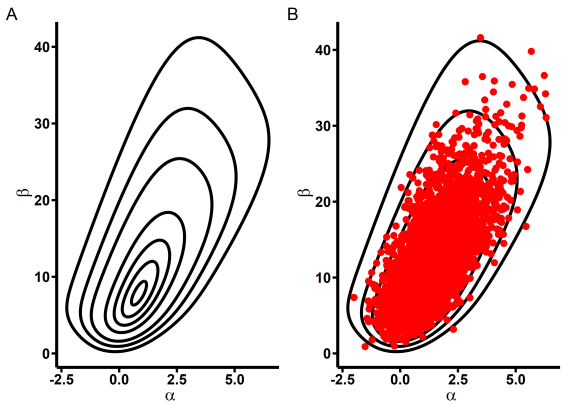

Figure 10. (A) Posterior Contours for $p(\alpha, \beta \mid x, y, n)$, (B) Posterior Draws

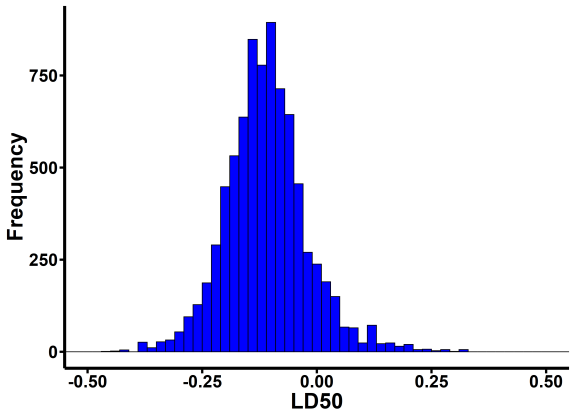

Figure 11. Histogram of the draws from the posterior distribution of the LD50 (on the scale of log dose in g/ml) in the bioassay example, conditional on the parameter β being positive.

	Mean	Standard Deviation	2.5%	25%	50%	75%	97.5%
α	1.320	1.102	-0.512	0.535	1.208	1.960	3.809
β	11.439	5.475	3.556	7.412	10.484	14.547	24.168

In total our results for this problem are consistent with section 3.7, and it appears that we have reached convergence.

Problem 5: BDA 3rd Ed. 11.3

Gibbs sampling: Table 11.4 contains quality control measurements from 6 machines in a factory. Quality control measurements are expensive and time-consuming, so only 5 measurements were done for each machine. In addition to the existing machines, we are interested in the quality of another machine (the seventh machine). Implement a separate, a pooled and hierarchical Gaussian model with common variance described in Section 11.6. Run the simulations long enough for approximate convergence. Using each of three models—separate, pooled, and hierarchical—report: (i) the posterior distribution of the mean of the quality measurements of the sixth machine, (ii) the predictive distribution for another quality measurement of the sixth machine, and (iii) the posterior distribution of the mean of the quality measurements of the seventh machine.

Table 2. Quality control measurements from 6 machines in a factory.

Machine	Measurements
1	83, 92 ,92, 46, 67
2	117, 109, 114, 104, 87
3	101, 93, 92, 86, 67
4	105, 119, 116, 102, 116
5	79, 97, 103, 79, 92
6	57, 92, 104, 77, 100

Solution:

We wish to report three distributions of interest:

(1) $\theta_6 \mid y$, the the posterior distribution of the mean of the quality measurements of the sixth machine

(2) $\tilde{y}_6 \mid y$, the predictive distribution for another quality measurement of the sixth machine
(3) θ_7, the posterior distribution of the mean of the quality measurements of the seventh machine

We use three approaches (Pooled, Separate and Hierarchical) to report the three distributions above and they are described as follows:

Pooled

With the pooled model we assume each machine has the same distribution. This can be written mathematically as

$$y_{ij} \mid \theta, \sigma^2 \sim N(\theta, \sigma^2)$$
$$p(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Under this model the calculation of the posteriors is extremely straightforward. They are of the same form as has been calculated in chapter three of the Bayesian Data Analysis 3rd Edition textbook. Here they are below:

$$\theta \mid \sigma^2, y \sim N(\bar{y}, \frac{\sigma^2}{n})$$
$$\text{where } \bar{y} = \frac{1}{nJ} \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij} \text{ and } n = \sum_{j=1}^{J} n_j$$
$$\sigma^2 \mid y \sim Inv - \chi^2(s^2, nJ - 1)$$
$$\text{where } s^2 = \frac{1}{nJ - 1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

While not required, we display the summary statistics for our two parameters below:

	Mean	Variance	2.5%	25%	50%	75%	97.5%
θ	92.93	3.47	86.24	90.62	92.79	95.09	100.07
σ	18.51	2.57	14.42	16.74	18.17	20.06	24.11

Now to plot our three distributions of interest we do the following things. To plot $\theta_6 \mid y$ we use the posterior form of $\theta \mid y$ since this is a pooled model. To plot $\tilde{y}_6 \mid y$ we use the form of the likelihood, and use our draws from $\theta \mid y$ and the pooled variance posterior $\sigma^2 \mid y$. Finally to get draws from the seventh machine simply notice that since we assume each machine has the same mean and variance, the distribution of the seventh machine is the same as the sixth so we simply sample from that posterior again. The results are below:

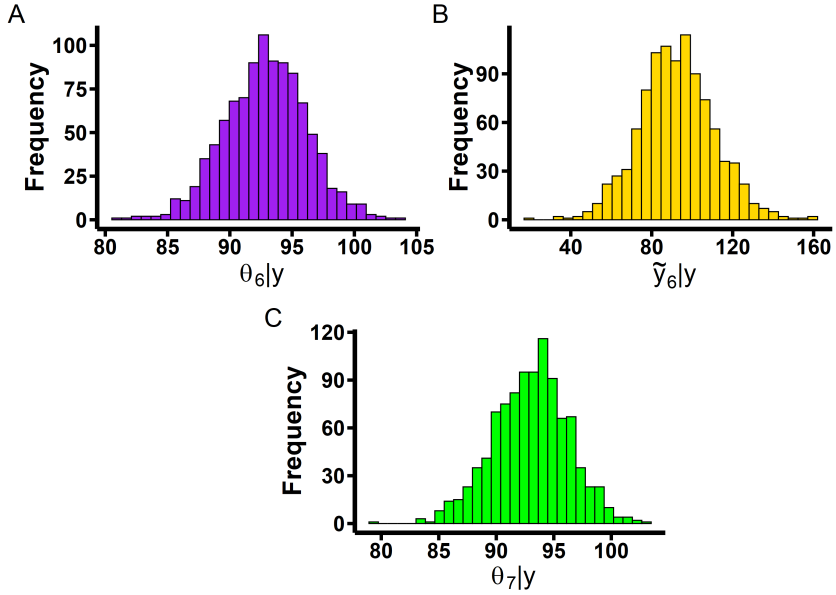

Figure 12. Histogram of distributions of interest, pooled model

Separate

The model for the separate case is presented as follows:

$$y_{ij} \mid \theta_j, \sigma^2 \sim N(\theta_j, \sigma^2)$$
$$p(\theta_1, \ldots, \theta_J, \sigma^2) \propto \frac{1}{\sigma^2}$$

Notice here each machine is assumed to have a different mean, but shared variance. This provides us with 7 posteriors that we will derive below:

$$\begin{aligned}
p(\theta_1, \ldots, \theta_J, \sigma^2 \mid y) &\propto p(\theta_1, \ldots, \theta_J, \sigma^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} \mid \theta_j, \sigma^2) \\
&\propto \frac{1}{\sigma^2} \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{(y_{ij}-\theta_j)^2}{2\sigma^2}} \\
&\propto \frac{1}{\sigma^2} \prod_{j=1}^{J} \left(\frac{1}{\sigma^2}\right)^{\frac{n_j}{2}} e^{-\sum_{i=1}^{n_j} \frac{(y_{ij}-\theta_j)^2}{2\sigma^2}} \\
&\propto \frac{1}{\sigma^2} \prod_{j=1}^{J} \left(\frac{1}{\sigma^2}\right)^{\frac{n_j}{2}} e^{-\frac{n_j(\theta_j-\bar{y}_{.j})^2+\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{.j})^2}{2\sigma^2}} \\
\text{since } (y_{ij}-\theta_j)^2 &= n_j(\theta_j-\bar{y}_{.j})^2 + \sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{.j})^2, \text{ where } \bar{y} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\sum_{j=1}^{J} n_j}{2}+1} e^{-\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{.j})^2}{2\sigma^2}} \prod_{j=1}^{J} e^{-\frac{n_j(\theta_j-\bar{y}_{.j})^2}{2\sigma^2}} \\
&= p(\sigma^2 \mid y) \prod_{j=1}^{J} p(\theta_j \mid \sigma^2, y)
\end{aligned}$$

This yields us the following posterior distributions:

$$\theta_j \mid \sigma^2, y \sim N(\bar{y}, \frac{\sigma^2}{n_j}) \quad \text{for } j = 1, \ldots, J$$

$$\text{where } \bar{y} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

$$\sigma^2 \mid y \sim Inv - \chi^2(S^2, N)$$

$$\text{where } N = \sum_{j=1}^{J} n_j \text{ and } S^2 = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2$$

Using the form of the posterior distributions we produce the following statistics:

	Mean	Variance	2.5%	25%	50%	75%	97.5%
θ_1	76.07	5.92	64.64	72.02	75.94	80.18	87.61
θ_2	106.44	5.67	94.76	102.68	106.44	110.28	117.46
θ_3	87.72	6.01	75.68	83.78	87.62	91.66	99.95
θ_4	111.65	5.87	100.02	107.57	111.83	115.68	122.66
θ_5	90.11	5.87	79.18	86.20	90.05	93.91	102.57
θ_6	86.14	6.19	74.58	81.99	86.23	90.12	98.65
σ	13.08	1.75	10.12	11.86	12.98	14.03	17.18

Now plotting the three distributions of interest is rather straightforward. The only trouble arises when trying to plot $\theta_7 \mid y$ since we have no information on measurements from the seventh machine so I'm actually not going to plot anything for that particular distribution.

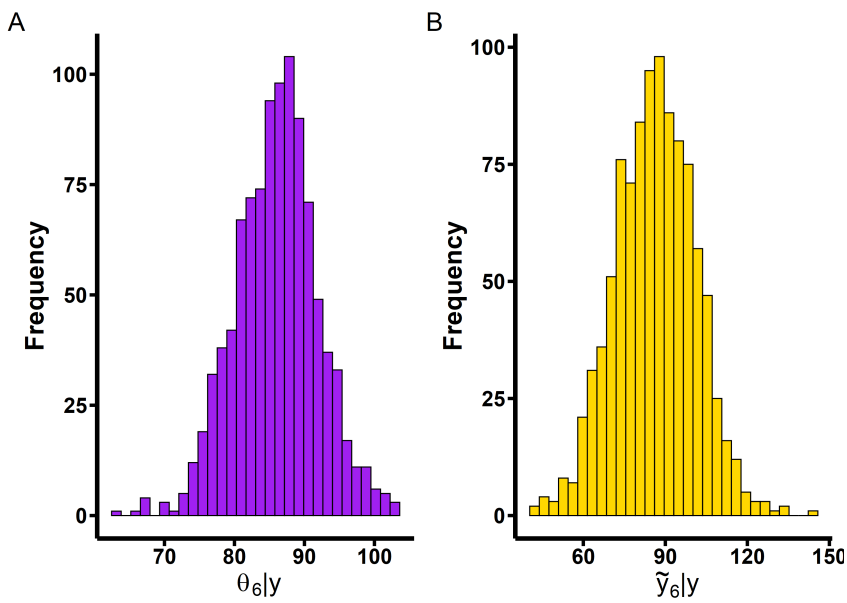

Figure 13. Histogram of distributions of interest, separate model

Hierarchical

$$
\begin{aligned}
&y_{ij} \mid \theta_j, \sigma^2 \sim N(\theta_j, \sigma^2), \text{ for } i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, J \\
&\theta_j \mid \mu, \tau \sim N(\mu, \tau^2), \text{ for } j = 1, 2, \ldots, J \\
&p(\mu, log(\sigma), log(\tau)) \propto \tau
\end{aligned}
$$

This provides us with a posterior distribution of the following form:

$$
\begin{aligned}
p(\theta, \mu, \sigma^2, \tau^2 \mid y) &\propto p(y \mid \theta, \mu, \sigma^2, \tau^2) p(\theta, \mu, \sigma^2, \tau^2) \\
&= p(y \mid \theta, \mu, \sigma^2, \tau^2) p(\theta \mid \mu, \sigma^2, \tau^2) p(\mu, \sigma^2, \tau^2) \\
&= p(\mu, \sigma^2, \tau^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} \mid \theta, \sigma^2) \prod_{j=1}^{J} p(\theta_j \mid \mu, \tau^2) \\
&= p(\mu, log\sigma, log\tau) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} \mid \theta, \sigma^2) \prod_{j=1}^{J} p(\theta_j \mid \mu, \tau^2)
\end{aligned}
$$

$$
\begin{aligned}
\theta_j \mid \sigma, \mu, \tau, y &\sim N(\hat{\theta}_j, V_j) \\
\hat{\theta}_j &= \frac{\frac{1}{\sigma^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \\
V_j &= \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}
\end{aligned}
$$

$$
\begin{aligned}
\tau^2 \mid \theta, \mu, \sigma, y &\sim Inv - \chi^2(J-1, \hat{\tau}^2) \\
\hat{\tau}^2 &= \frac{1}{J-1} \sum_{j=1}^{J} (\theta_j - \mu)^2
\end{aligned}
$$

$$
\mu \mid \theta, \sigma, \tau, y \sim N(\frac{1}{J} \sum_{j=1}^{J} \theta_j, \frac{\tau^2}{J})
$$

$$
\begin{aligned}
\sigma^2 \mid \theta, \mu, \tau, y &\sim Inv - \chi^2(\sum_{j=1}^{J} n_j, \hat{\sigma}^2) \\
\hat{\sigma}^2 &= \frac{1}{\sum_{j=1}^{J} n_j} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2
\end{aligned}
$$

As with any Gibbs Sampling chain we need to initialize starting values. There are a couple of ways to go about this. Firstly we can always rely on the Normal Approximation and sample a value from that distribution to start with. Secondly we can just give it our most reasonable guess. I will proceed with the second tactic since it will be easier to code.

Before we begin one should notice certain things about the conditional posteriors. For one, the conditional posterior of σ^2 depends only on θ. The distribution of μ depends only upon τ^2 and $\theta.\tau^2$ depends only on μ and θ. Given these dependencies if we're smart about which distribution

we sample from first, we'll only need to initialize a few parameter values. Everything depends on θ so that will need to be initialized. Likewise μ and τ are dependent on each other so we'll have to initialize at least one of them. Let's choose μ since we can just let $\mu = \frac{1}{J}\sum_{j=1}^{J}\theta_j^0$, where θ_j^0 is the initial value of θ_j. All in all we simplify the heavy lifting by initializing only J values (it is J since the J values of θ_j will determine what we initialize μ with).

So starting our chain off by initializing θ and μ, we sample from our conditionals in the following order:

$$\sigma^2 \mid \theta, \mu, \tau, y \sim Inv - \chi^2(\sum_{j=1}^{J} n_j, \hat{\sigma}^2)$$
$$\tau^2 \mid \theta, \mu, \sigma, y \sim Inv - \chi^2(J-1, \hat{\tau}^2)$$
$$\theta_j \mid \sigma, \mu, \tau, y \sim N(\hat{\theta}_j, V_j)$$
$$\mu \mid \theta, \sigma, \tau, y \sim N(\frac{1}{J}\sum_{j=1}^{J}\theta_j, \frac{\tau^2}{J})$$

We run our Gibbs Sampler for $N = 50,000$ iterations, and disregard the first 60% of samples as burn-in. This takes approximately 12 seconds to run and the statistics for our parameter distributions are shown below:

	Mean	Variance	2.5%	25%	50%	75%	97.5%
θ_1	79.79	6.81	66.56	75.23	79.73	84.18	93.26
θ_2	103.21	6.55	90.72	98.75	103.22	107.64	116.24
θ_3	88.95	6.06	76.82	85.04	88.97	93.00	100.76
θ_4	107.32	6.91	93.00	102.86	107.49	112.10	120.56
θ_5	90.64	6.06	78.40	86.68	90.75	94.65	102.48
θ_6	87.64	6.09	75.50	83.56	87.74	91.74	99.32
σ	14.97	2.29	11.23	13.34	14.73	16.33	20.16
μ	93.00	8.25	77.36	88.63	92.96	97.24	109.16
τ	16.28	10.26	4.07	10.33	14.24	19.51	41.04

Trace Plots

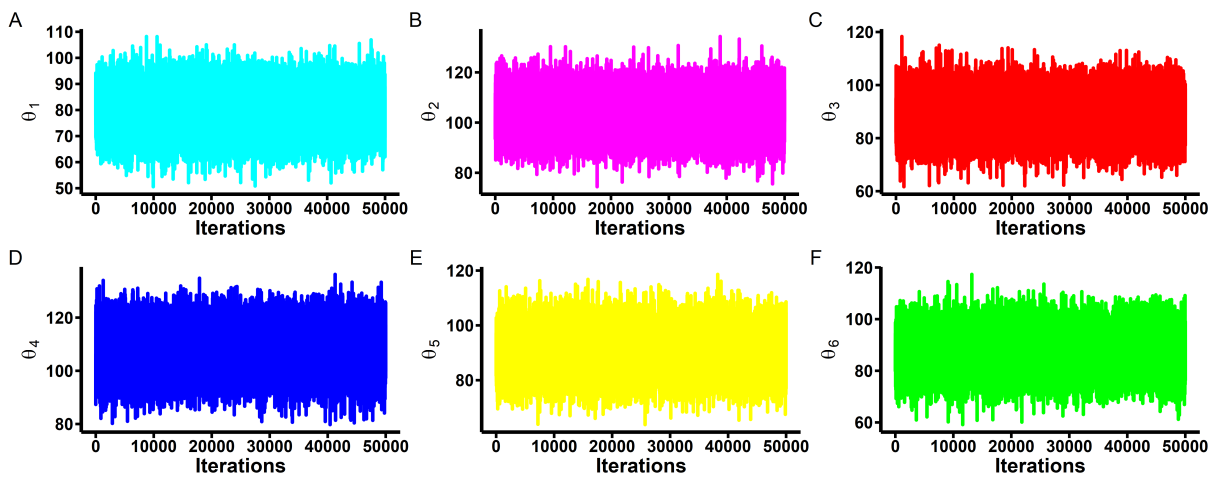

Figure 14. Trace Plots for θ parameters

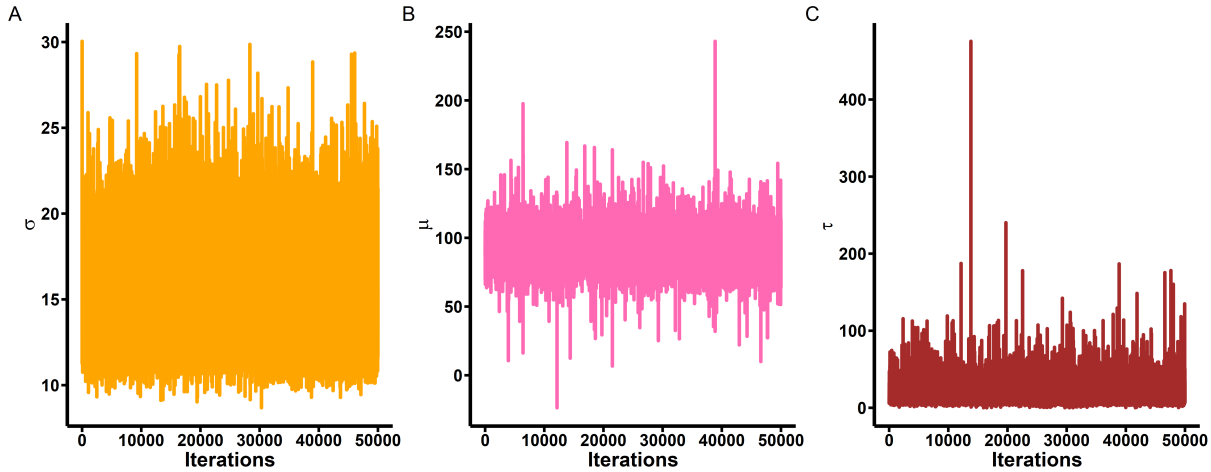

Figure 15. Trace Plots for our hyperparameters

In examining the trace plots nothing looks awry. Mixing appears to be good since we're seeing the hairy caterpillar form.

Autocorrelation Plots

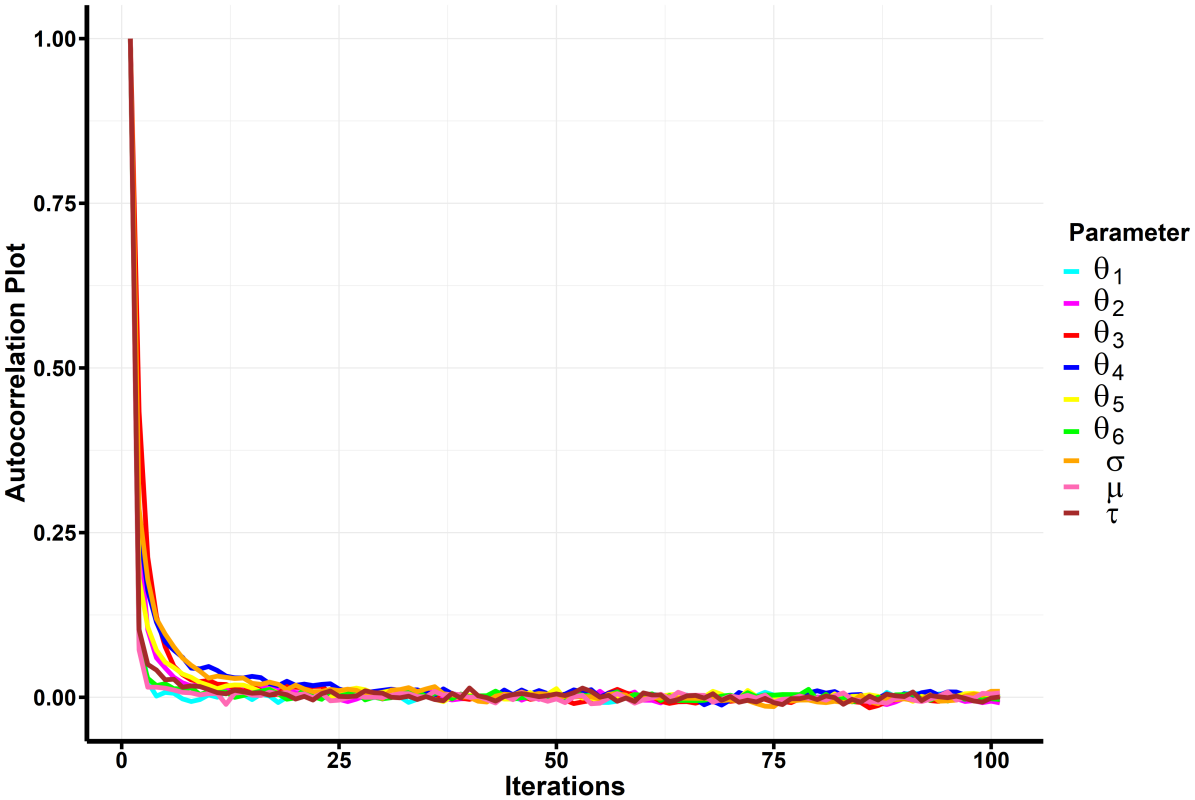

Figure 16. Autocorrelation Plots for our Parameters

Likewise with our autocorrelation plots we see the autocorrelation become virtually zero as the lag increases indicating that our chain is traversing the posterior distribution relatively well.

Once again, plotting the three distirubtions is a simple task. We already have $\theta_6 \mid y$ from our Gibbs sampling procedure, and calculating $\tilde{y}_6 \mid y$ is simply just plugging in $\theta_6 \mid y$ and $\sigma \mid y$ into the likelihood. To calculate the results of the new sample $\theta_7 \mid y$ we just need take our posterior samples $\mu \mid y$ and $\tau \mid y$ and plug them into the distribution of $\theta_j \mid \mu, \tau$.

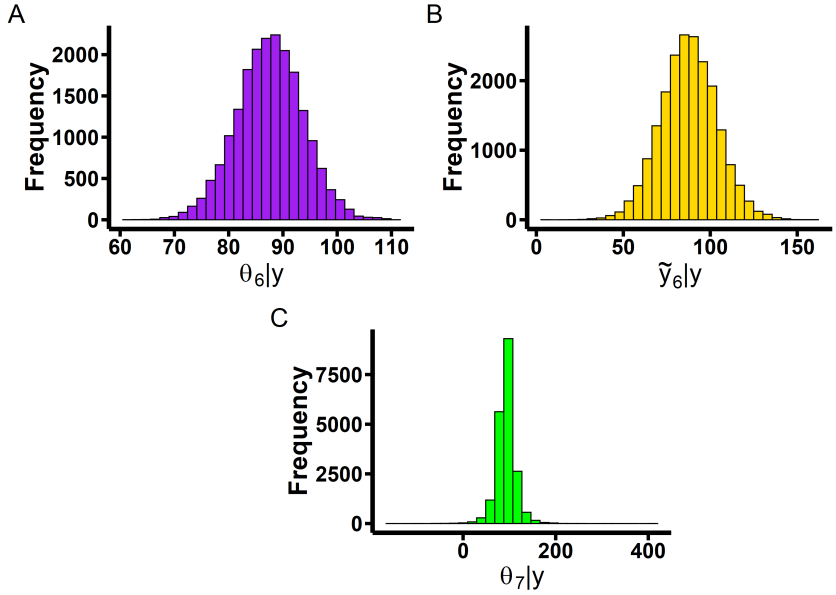

Figure 17. Histogram of distributions of interest, hierarchical model

Problem 6: BDA 3rd Ed. 11.4

Gibbs sampling: Extend the model in Exercise 11.3 by adding a hierarchical model for the variances of the machine quality measurements. Use an $Inv-\chi^2$ prior distribution for variances with unknown scale σ_0^2 and fixed degrees of freedom. (The data do not contain enough information for determining the degrees of freedom, so inference for that hyperparameter would depend very strongly on its prior distribution in any case). The conditional distribution of σ_0^2 is not of simple form, but you can sample from its distribution, for example, using grid sampling.

Solution:

For this problem we only consider the hierarchical model and we assume that each machine has its own variance as described in the hierarchical model below:

$$
\begin{aligned}
&y_{ij} \mid \theta_j, \sigma_j^2 \sim N(\theta_j, \sigma_j^2), \text{ for } i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, J \\
&\theta_j \mid \mu, \tau \sim N(\mu, \tau^2), \text{ for } j = 1, 2, \ldots, J \\
&\sigma_j^2 \mid \sigma_0, \nu \sim Inv-\chi^2(\nu, \sigma_0^2) \\
&p(\mu, log(\tau), log\sigma_0) \propto \tau
\end{aligned}
$$

This provides us with a posterior distribution of the following form:

$$
\begin{aligned}
p(\theta, \sigma^2, \mu, \tau^2, \sigma_0^2, \nu \mid y) &\propto p(y \mid \theta, \sigma^2, \mu, \tau^2, \sigma_0^2, \nu) p(\theta, \sigma^2, \mu, \tau^2, \sigma_0^2, \nu) \\
&= p(y \mid \theta, \sigma^2, \mu, \tau^2, \sigma_0^2, \nu) p(\theta \mid \mu, \tau^2) p(\sigma^2 \mid \sigma_0^2, \nu) \\
&= p(\mu, \tau^2, \sigma_0^2, \nu) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} \mid \theta, \sigma^2) \prod_{j=1}^{J} p(\theta_j \mid \mu, \tau^2) \prod_{j=1}^{J} p(\sigma_j^2 \mid \sigma_0^2, \nu)
\end{aligned}
$$

The process to get our conditional posteriors can be easily simplified by noticing dependencies. Firstly, τ, μ are not connected to σ_j^2, σ_0^2 and ν so their conditional posteriors are of the same form in Problem 5. Similarly with θ_j with the added caveat that the σ's are replaced with σ_j's as shown below:

$$
\begin{aligned}
\theta_j \mid \sigma, \mu, \tau, y &\sim N(\hat{\theta}_j, V_j) \\
\hat{\theta}_j &= \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \\
V_j &= \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}
\end{aligned}
$$

Sadly, however, we need to put in some elbow grease to get conditional posteriors for the other parameters. Let's start with σ_j^2

$$
\begin{aligned}
p(\sigma_j^2 \mid \theta_j, \mu, \tau^2, \sigma_0^2, \nu \mid y) &\propto \prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij} \mid \theta_j, \sigma_j^2) \prod_{j=1}^{J} p(\sigma_j^2 \mid \sigma_0^2, \nu) \\
&\propto \prod_{i=1}^{n_j} p(y_{ij} \mid \theta_j, \sigma_j^2) p(\sigma_j^2 \mid \sigma_0^2, \nu) \\
&\propto \frac{1}{(\sigma_j^2)^{n_j/2}} e^{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2} \frac{1}{(\sigma_j^2)^{1+\nu/2}} e^{-\frac{\nu\sigma_0^2}{2\sigma_j^2}} \\
&\propto \frac{1}{(\sigma_j^2)^{1+(n_j+\nu)/2}} e^{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2} e^{-\frac{\nu\sigma_0^2}{2\sigma_j^2}} \\
&\propto \frac{1}{(\sigma_j^2)^{1+(n_j+\nu)/2}} e^{-\frac{1}{2\sigma_j^2}(\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2+\nu\sigma_0^2)} \\
&\propto \frac{1}{(\sigma_j^2)^{1+(n_j+\nu)/2}} e^{-\frac{n_j+\nu}{2\sigma_j^2}\frac{1}{n_j+\nu}(\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2+\nu\sigma_0^2)}
\end{aligned}
$$

So we see that

$$
\sigma_j^2 \mid \theta, \mu, \tau, \sigma_0^2, \nu, y \sim Inv-\chi^2(\nu + n_j, \frac{1}{\nu + n_j}\Big(\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2 + \nu\sigma_0^2\Big))
$$

Now we derive the conditional posterior for σ_0^2:

$$
\begin{aligned}
p(\sigma_0^2 \mid \theta, \sigma^2, \mu, \tau^2, \nu, y) &\propto p(\mu, \tau^2, \sigma_0^2, \nu) \prod_{j=1}^{J} p(\sigma_j^2 \mid \sigma_0^2, \nu) \\
&\propto \frac{1}{\sigma_0^2} \prod_{j=1}^{J} \frac{(\sigma_0^2 \frac{\nu}{2})^{\nu/2}}{\Gamma(\frac{\nu}{2})} e^{-\frac{\nu\sigma_0^2}{2\sigma_j^2}} \\
&\propto (\sigma_0^2)^{J\nu/2-1} e^{-\sum_{j=1}^{J}\frac{\nu\sigma_0^2}{2\sigma_j^2}}
\end{aligned}
$$

This can't be simplified into an easier form, but since this is a univariate conditional posterior we can just do a simple grid sampling procedure. Overall our schematic for Gibbs Sampling will be sampling from the above distributions.

As with any Gibbs Sampling chain we need to initialize starting values. There are a couple of ways to go about this. Firstly we can always rely on the Normal Approximation and sample a value from that distribution to start with. Secondly we can just give it our most reasonable guess. I will proceed with the second tactic since it will be easier to code.

Before we begin one should notice certain things about the conditional posteriors. For one, the conditional posterior of σ_0^2 depends only on the values of σ^2. The distribution of μ depends only upon τ^2 and θ. σ_j^2 depends only upon θ and σ_0^2. τ^2 depends only on μ and θ_j. Given these dependencies if we're smart about which distribution we sample from first, we'll only need to initialize a few parameter values. Almost everything depends on θ so that will need to be initialized. Likewise μ and τ are dependent on each other so we'll have to initialize at least one of them. Let's choose μ since we can just let $\mu = \frac{1}{J}\sum_{j=1}^{J}\theta_j^0$, where θ_j^0 is the initial value of θ_j. Likewise let's initialize σ_0^2 since σ_j^2 is dependent upon this value and it will be easier to pick one value than to pick J different values. All in all we simplify the heavy lifting by initializing only J+1 values (it is J + 1 since σ_0^2 is scalar and the J values of θ_j will determine what we initialize μ with).

So starting our chain off by initializing θ, μ and σ_0^2, our sampling scheme will be done via:

$$\sigma_j^2 \mid \theta, \mu, \tau, \sigma_0^2, y \sim Inv - \chi^2(\nu + n_j, \frac{1}{\nu + n_j}\Big(\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2 + \nu\sigma_0^2\Big))$$

$$p(\sigma_0^2 \mid \theta, \sigma^2, \mu, \tau^2, y) \propto (\sigma_0^2)^{J\nu/2-1} e^{-\sum_{j=1}^{J}\frac{\nu\sigma_0^2}{2\sigma_j^2}}$$

$$\tau^2 \mid \theta, \mu, \sigma, y \sim Inv - \chi^2(J-1, \hat{\tau}^2)$$

$$\theta_j \mid \sigma, \mu, \tau, y \sim N(\hat{\theta}_j, V_j)$$

$$\mu \mid \theta, \sigma, \tau, y \sim N(\frac{1}{J}\sum_{j=1}^{J}\theta_j, \frac{\tau^2}{J})$$

We run our Gibbs Sampler for $N = 50,000$ iterations, and disregard the first 60% of samples as burn-in. This takes approximately 25 seconds to run and the statistics for our parameter distributions are shown below:

	Mean	Variance	2.5%	25%	50%	75%	97.5%
θ_1	81.07	8.27	65.46	75.55	80.69	86.34	98.14
θ_2	104.07	5.80	92.06	100.50	104.27	107.83	114.86
θ_3	89.00	5.77	77.57	85.47	88.99	92.58	100.63
θ_4	109.18	5.41	96.53	106.26	109.66	112.63	118.60
θ_5	90.66	5.35	80.44	87.32	90.61	93.84	101.98
θ_6	88.11	7.21	73.90	83.51	88.16	92.57	102.61
σ_1	18.79	5.99	10.94	14.61	17.52	21.58	33.97
σ_2	13.07	4.30	7.39	10.08	12.20	15.01	23.60
σ_3	13.55	4.47	7.89	10.57	12.63	15.51	24.65
σ_4	10.78	4.19	5.74	8.00	9.86	12.41	21.56
σ_5	12.20	4.15	6.93	9.39	11.30	13.98	22.12
σ_6	18.01	5.53	10.52	14.20	16.93	20.61	31.91
μ	93.74	8.49	76.88	89.36	93.91	98.28	109.64
τ	16.72	9.75	5.73	10.76	14.50	19.90	40.70
σ_0	11.05	2.63	6.48	9.17	10.84	12.67	16.82

Looking at the statistics for the parameters in Problem 6 and Problem 5, there isn't a significant difference in the distributions of the θ's, τ and μ which subtlety implies this extra layer of complexity is unnecessary.

Trace Plots

As our statistics aren't too different from those in Problem 5 and we've already explained why we use different diagnostics such as Trace Plots, Autocorrelation Plots and Cumulative Mean Plots, I will just plot trace plots here and call it a day if they look nice enough.

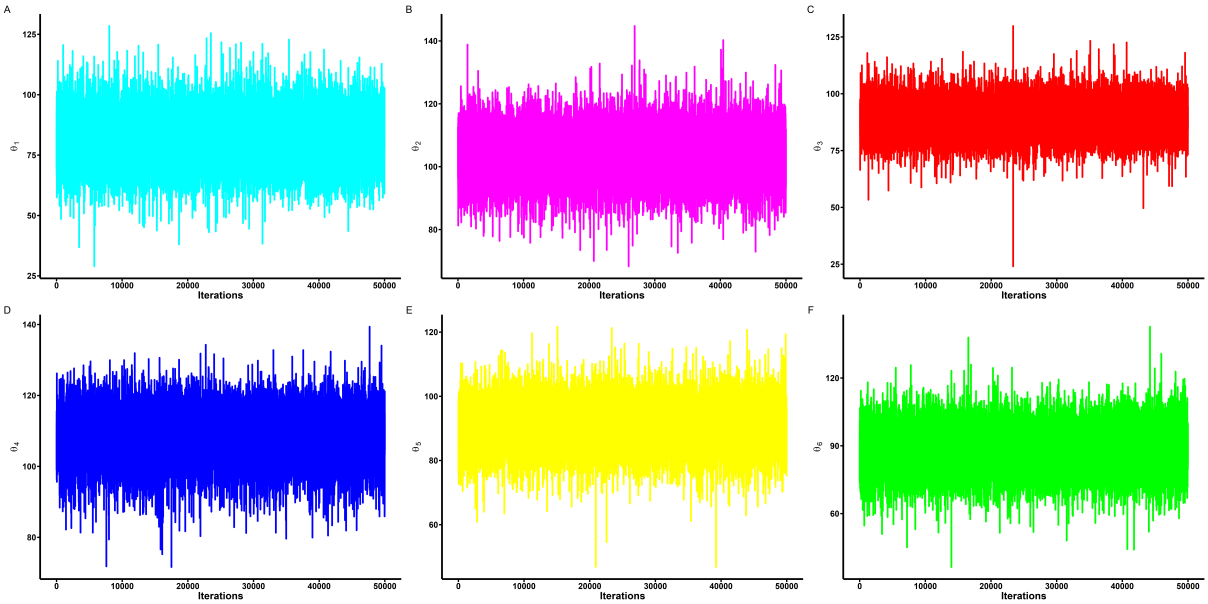

Figure 18. Trace Plots for our nine different parameters. Notice σ and τ are on the log scale.

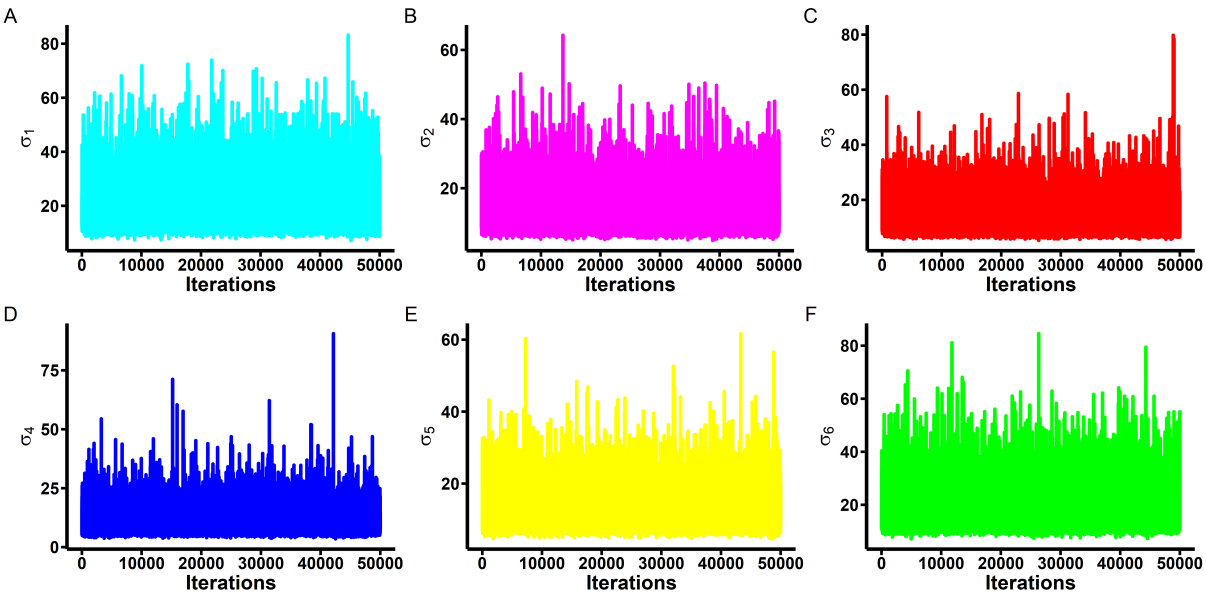

Figure 19. Trace Plots for σ parameters

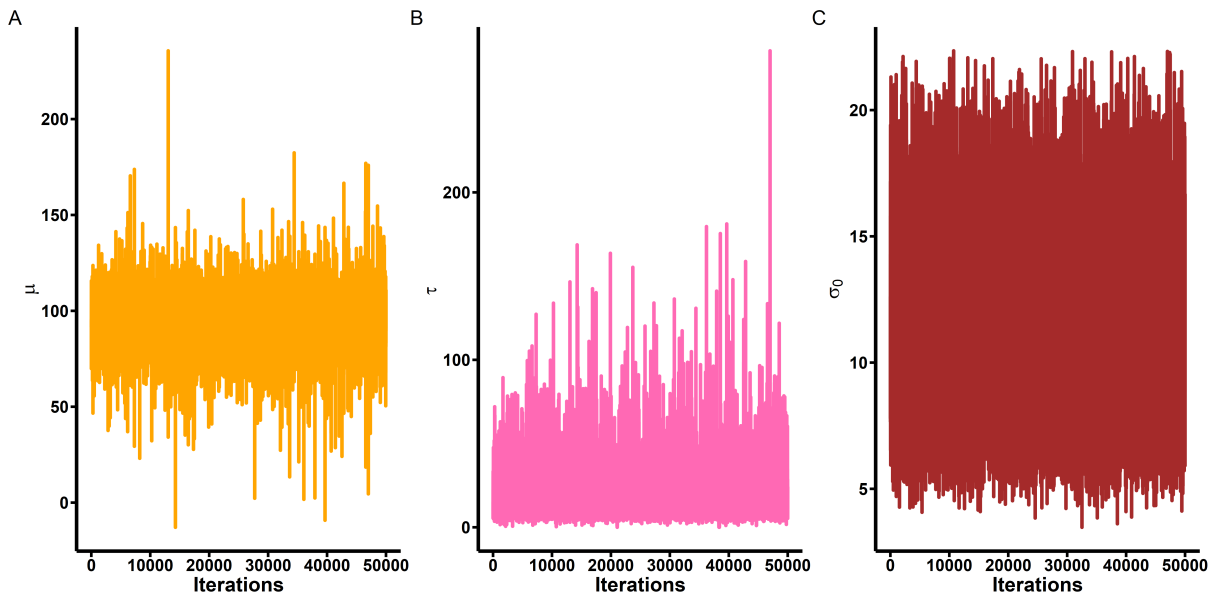

Figure 20. Trace Plots for our hyperparameters

The trace plots above look fine so I'll assume we've reached convergence. Once again, plotting the three distributions is a simple task. We already have $\theta_6 \mid y$ from our Gibbs sampling procedure, and calculating $\tilde{y}_6 \mid y$ is simply just plugging in $\theta_6 \mid y$ and $\sigma_6 \mid y$ into the likelihood. To calculate the results of the new sample $\theta_7 \mid y$ we just need take our posterior samples $\mu \mid y$ and $\tau \mid y$ and plug them into the distribution of $\theta_j \mid \mu, \tau$.

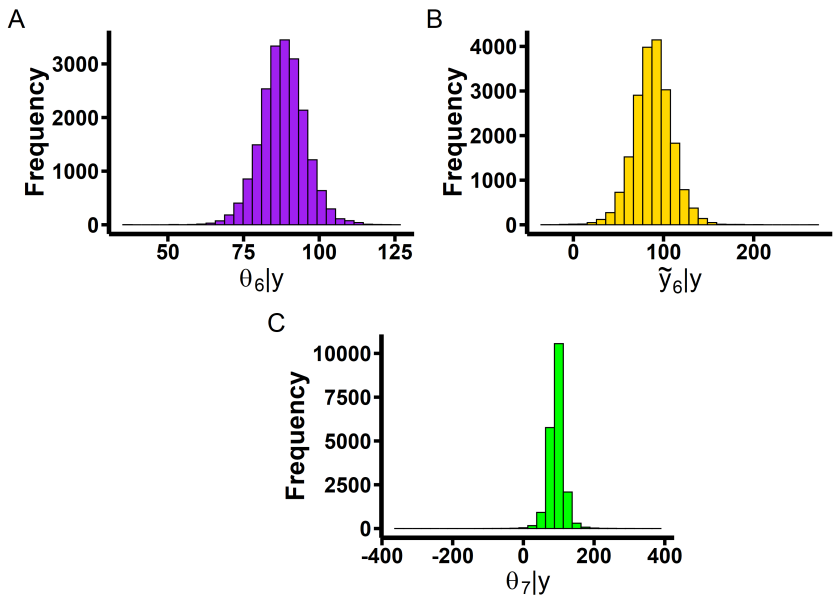

Figure 21. Histogram of distributions of interest, hierarchical model

Problem 7: BDA 3rd Ed. 13.5

Estimating the number of unseen species (see Fisher, Corbet, and Williams, 1943, Efron and Thisted, 1976, and Seber, 1992): suppose that during an animal trapping expedition the number of times an animal from species i is caught is $x_i \sim Poisson(\lambda_i)$. For parts (a)–(d) of this problem, assume a $Gamma(\alpha, \beta)$ prior distribution for the $\lambda_i's$, with a uniform hyperprior distribution on

(α, β). The only observed data are y_k, the number of species observed exactly k times during a trapping expedition, for $k = 1, 2, 3, \ldots$

(a) Write the distribution $p(x_i \mid \alpha, \beta)$.

Solution:

$$\begin{aligned} x_j|\lambda_j &\sim Poisson(\lambda_j) \\ \lambda_j|\alpha, \beta &\sim Gamma(\alpha, \beta) \\ p(\alpha, \beta) &\propto 1 \end{aligned}$$

This part is relatively simple. We integrate over the parameter λ_j to find the likelihood wrt the hyperparameters.

$$\begin{aligned} p(x_j \mid \alpha, \beta) \propto \int p(x_j \mid \lambda_j) p(\lambda_j \mid \alpha, \beta) d\lambda_j &\propto \int \frac{\lambda_j^{x_j}}{y_j!} e^{-\lambda_j} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_j^{\alpha-1} e^{-\beta\lambda_j} d\lambda_j \\ &= Neg - Binom(x_j \mid r = \alpha, p = \frac{\beta}{\beta+1}) \end{aligned}$$

Many details are skipped here, mainly because a similar derivation is done in the HW III solutions, Problem 5.14.

(b) Use the distribution of x_i to derive a multinomial distribution for y given that there are a total of N species.

Solution:

For this problem it is highly important for us to define the parameters.

$$\begin{aligned} x_i &= \text{ Number of animals of species i that were captured} \\ y_k &= \text{Number of species observed exactly k times} \end{aligned}$$

Now let's define some things. Firstly, how far does the index k of y_k extend? Well the k refers to the number of times a species was observed. Obviously you can't observe any given species an infinite amount of times so let's put an upper bound on this, call it K. Essentially K is the maximum number of times a species can be observed.

$$y_k = \text{Number of species observed exactly k times } \text{ for } k = 1, \ldots, K$$

Secondly, we're given information that there are exactly N species. Given the definition of y_k this means that each y_k captures a unique amount of species. To make this more concrete, think about this: say we observe 1 kangaroo, 1 dolpin and 1 spider monkey. We observe 2 jaguars, 2 wallaby's , 2 beetles and 2 eagles. We observe 5 rats. We observe 10 squirrels, 10 walruses and 10 rabbits. This means $y_1 = 3$, $y_2 = 4$, $y_5 = 1$ and $y_{10} = 2$. Since we observe 10 walruses, this species only contributes mass to y_{10}. Likewise since we only see 5 rats, they only contribute to y_5. I say all this to point out that each y_k therefore can be viewed as a set and each y_k is therefore disjoint from each other. This has to mean that the following identity must hold:

$$\sum_{k=0}^{K} y_k = N$$

y_0 stands for the number of species we don't see.
Thirdly, for our purposes, captured and observed are virtually identical. If a species was captured, it was obviously found and observed by the study scientist. Therefore since we only observe a value for y_k if $x_i = k$ for some species i. This implies $\theta_i = P(x_i = k \mid \alpha, \beta)$. Of course, though, one must realize the parameter space of a negative binomial is infinite. We have that

$$\sum_{j=0}^{\infty} P(x_i = j \mid \alpha, \beta) = 1$$

This may seem problematic as we essentially will deal with an infinite multinomial, which I don't even know if this is well-defined, but everything works itself out in the end. Since we only observe a maximum K of any species, $y_k = 0$ for $k \geq K + 1$ so the likelihood reduces to:

$$\begin{aligned} p(y \mid \theta) &= \frac{N!}{\prod_{k=0}^{\infty} y_k!} \prod_{k=0}^{\infty} P(x_i = j \mid \alpha, \beta)^{y_k} \\ &= \frac{N!}{\prod_{k=1}^{K} y_k!} \prod_{k=1}^{K} P(x_i = j \mid \alpha, \beta)^{y_k} \end{aligned}$$

Assuming $y_0 = 0$ as well. Note this is a skeleton for the problem. I'm actually not even sure if this is the correct approach. Regardless,below is a plot of the posterior's contours that arise from this model when using the hyperprior $p(\alpha, \beta) \propto 1$.

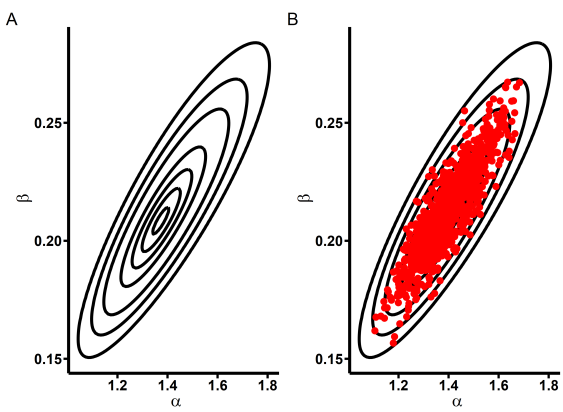

Figure 22. (A) Posterior Contours for $p(\alpha, \beta \mid y)$, (B) Posterior Draws

(c) Suppose that we are given $y = (118, 74, 44, 24, 29, 22, 20, 14, 20, 15, 12, 14, 6, 12, 6, 9, 9, 6, 10, 10, 11, 5, 3, 3)$, so that 118 species were observed only once, 74 species were observed twice, and so forth, with a total of 496 species observed and 3266 animals caught. Write down the likelihood for y using the multinomial distribution with 24 cells (ignoring unseen species). Use any method to find the mode of α, β and an approximate second derivative matrix.

Solution:

Using the above likelihood, as well as the flat prior on α and β, we find the MAP and the Fisher Information Matrix for this problem to be

$$\hat{\theta}_{MAP} = \begin{pmatrix} 1.372 \\ 0.208 \end{pmatrix}, [I(\hat{\theta}_{MAP})]^{-1} = \begin{pmatrix} 0.0095 & 0.0014 \\ 0.0014 & 0.0003 \end{pmatrix}.$$

(d) Derive an estimate and approximate 95% posterior interval for the number of additional species that would be observed if 10,000 more animals were caught.

Solution:

From part c we're given that

$$\sum_{k=0}^{24} k y_k = 3,266$$
$$\sum_{k=0}^{24} y_k = 496$$

In other words we catch 3,266 animals and observe 496 species. Overall the maximum number of times a species is observed is 24 times. Now, however, we observe an additional 10,000 animals. This means our upper bound of 24 is bound to change, call the new maximum number of time N. Our goal in this problem is to estimate $\sum_{k=0}^{N} y_k$. Firstly we have to realize two things (Credit to Ben for pointing this out). Firstly, the number of already known species can increase. That is to say if we saw 20 rabbits before, it's entirely possible we see 200 rabbits this time meaning rabbits would contribute mass to y_{200} instead of y_{20}.

Let's give rudimentary estimates first and foremost. If we observed 496 unique species out of the 3,266 animals that were caught, this implies we see ≈ 15 unique animals per 100 animals caught which roughly translates to seeing 1500 unique species out of 10,000 animals caught. At the same time out of these new 10,000 animals it is entirely possible that we observe the previous 496 seen before so we'd expect our lower bound to be around 1,000 and our upper bound to be around 1,500. This is an extremely crude estimate, but it provides necessary intuition into what is reasonable.

Perhaps this is better formulated through the language of the $x_i \mid \alpha, \beta$ distribution. Recall x_i represents the number of times species i is caught. If a species is caught, this means it must have been observed implying the relation $\sum_{i=1}^{\gamma} x_i \mid \alpha, \beta = \sum_{k=1}^{N} k y_k = 13,266$, where γ is the number of unique species observed after the additional 10,000 observations. We can then estimate the distribution of γ and simply subtract 496 from the resulting 95% Credible Interval since those previously observed 496 species are implicitly included in the estimate. There are three possible scenarios:

(1) We only observe new species in the additional 10,000, meaning we don't see any of the original 496 again.
(2) We observe all of the original 496 in the additional 10,000 captured animals.
(3) We only observe some of the original 496 species in the additional 10,000 captured animals.

Truth be told I don't know if what I'm doing is correct. Honestly everything in part (d) could just amount to gibberish. Anyways here is how I do the sampling:

(1) derive S samples of $\{(\alpha^s, \beta^s)\}_{s=1}^S$ using rejection sampling, importance resampling, gibbs sampling, Metropolis hastings, etc.
(2) for each tuple (α^s, β^s) draw from $x_i \mid \alpha^s, \beta^s$ until $\sum x_i \mid \alpha^s, \beta^s > 10000$. The number of iterations it took to pass this threshold should be the number of new species to expect.

Doing this method yields us a 95% Credible Interval of (1452, 1593) for new species observed in the additional 10,000 species caught which seems to be consistent with what we would heuristically expect.

(e) Evaluate the fit of the model to the data using appropriate posterior predictive checks.

Solution:

To evaluate the fit of the model let's think for a second what we're trying to capture. Firstly we want to make sure that we can consistently predict the total number of animals caught. Secondly we want to make sure that we see the same amount of species seen. Therefore let's define our two test statistics to be (i) $T_1(y) = \sum_{j=1}^{496} y_j$ and (ii) $T_2(y) = \max x_j$.

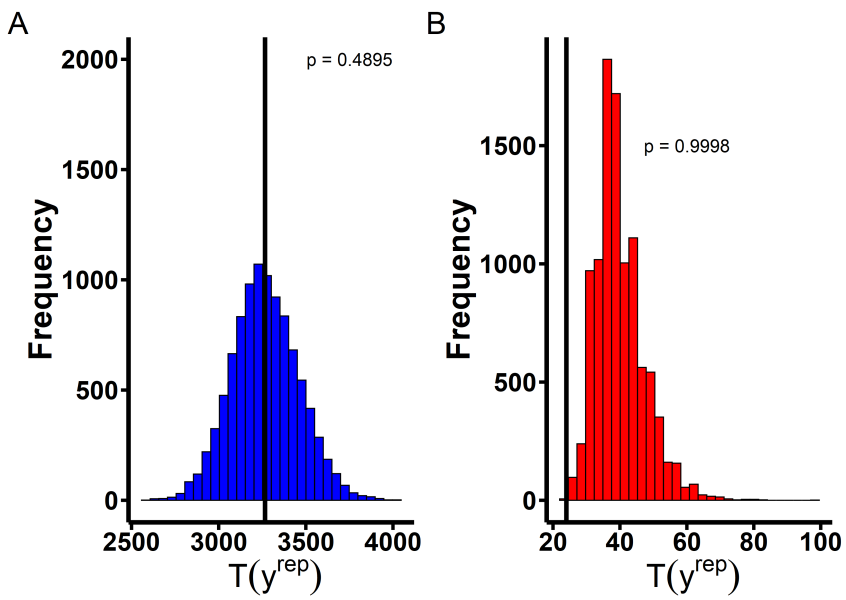

Figure 23. (A) Posterior Contours for $p(\alpha, \beta \mid y)$, (B) Posterior Draws

(f) Discuss the sensitivity of the inference in (d) to each of the model assumptions.

Solution:

Probably one of the more interesting model assumptions is that of modeling the number of captures as a Poisson distribution. In a sense this is probably not the wisest model assumption due to the fact that captures are most likely not independent. Capturing one animal makes it less likely to capture another. Less importantly is the vagueness in determining captures. The problem states nothing of whether animals are captured and released so it could be possible we see the same animal more than once. Also, given the large amount of animals that are present within the dataset, our likelihood is more likely to dominate inference with our posterior so the uniform hyperprior on (α, β) is most likely having negligible influence.

Problem 8

Generate/simulate n = 120 observations from a three component mixture Gaussian model with

$$
\begin{aligned}
(p_1, p_2, p_3) &= (0.1, 0.3, 0.6) \\
(\mu_1, \sigma_1^2) &= (0, 1) \\
(\mu_2, \sigma_2^2) &= (-2, 2) \\
(\mu_3, \sigma_3^2) &= (3, 16)
\end{aligned}
$$

Use these n data points to estimate (p_i, μ_i, σ_i^2) using:

(a) EM Algorithm: Implement EM to find their MLE

Solution:

We begin by defining the model at hand. Assume we have J data points $\{y_j\}_{j=1}^J$ and K mixture components.

$$
\begin{aligned}
y_j \mid z_k = 1, \theta &\sim N(\mu_k, \sigma_k^2) \text{ for } j = 1, 2, \ldots, J \\
z \mid \lambda &\sim Multinomial(1; \lambda_1, \ldots, \lambda_K)
\end{aligned}
$$

Here we can define $z = (z_1, z_2, \ldots, z_K)$ where

$$
z_k = \begin{cases} 1 & \text{if drawn from kth mixture component} \\ 0 & \text{otherwise} \end{cases}
$$

thus the joint distribution for this model will be of the form

$$
\begin{aligned}
p(y, z \mid \theta, \lambda) &= p(z \mid \lambda) p(y \mid z, \theta) \\
&= \prod_{j=1}^{J} \Big(\frac{\Gamma(\sum_{k=1}^{K} z_{jk})}{\prod_{k=1}^{K} z_{jk}!} \prod_{k=1}^{K} \lambda_k^{z_{jk}} \Big) \prod_{j=1}^{J} \Big(\prod_{k=1}^{K} p(y_j \mid \theta_k)^{z_{jk}} \Big) \\
&= \prod_{j=1}^{J} \prod_{k=1}^{K} (\lambda_k p(y_j \mid \theta_k))^{z_{jk}}
\end{aligned}
$$

Since $z \sim Multinomial(1; \lambda_1, \ldots, \lambda_K)$, this means $\Gamma(\sum_{k=1}^{K} z_{jk}) = Gamma(1) = 1$. Since we're concerned with finding the MLE, we don't concern ourselves with priors; our only concern is the likelihood. Now recall the two steps of the EM algorithm:

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) &= E(logp(z, y \mid \theta) \mid y, \theta^{(t)}), \text{ Expectation Step} \\
\theta^{(t+1)} &= \underset{\theta}{\operatorname{argmax}}\, Q(\theta \mid \theta^{(t)}), \text{ Maximization Step}
\end{aligned}
$$

From the above we have already obtained the form of $p(y, z \mid \theta)$. However, notice that this expecation has both $y, \theta^{(t)}$ in the conditional so the expecation is of the form:

$$
E(logp(z, y \mid \theta) \mid y, \theta^{(t)}) = \int logp(z, y \mid \theta) p(z \mid y, \theta^{(t)}) dz
$$

Since z is a discrete random variable, this means our expectations will take the form:

$$E(logp(z, y \mid \theta) \mid y, \theta^{(t)}) = \sum_{k=1}^{K} logp(y \mid \theta)p(z \mid y, \theta^{(t)})$$

Notice by the chapter 5 identity,

$$\begin{aligned} p(z_k = 1 \mid y, \theta^{(t)}, \lambda^{(t)}) &= \frac{p(z_k = 1, y \mid \theta^{(t)})}{p(y \mid \theta^{(t)})} \\ &= \frac{\lambda_k^{(t)} N(y \mid \mu_k^{(t)}, \sigma_k^{2(t)})}{\sum_{k=1}^{K} \lambda_k^{(t)} N(y \mid \mu_k^{(t)}, \sigma_k^{2(t)})} \end{aligned}$$

So we find that the expectation step has the form:

$$Q(\theta \mid \theta^{(t)}) = \sum_{k=1}^{K} \Big(\sum_{j=1}^{J} log\lambda_k + \frac{1}{2} log(2\pi\sigma_k) - \frac{1}{2\sigma_k^2}(y_j - \mu_k)^2 \Big) \frac{\lambda_k^{(t)} N(y \mid \mu_k^{(t)}, \sigma_k^{2(t)})}{\sum_{k=1}^{K} \lambda_k^{(t)} N(y \mid \mu_k^{(t)}, \sigma_k^{2(t)})}$$

Maximizing this function over the parameters yields the following MLE's that we'll need to update

$$\begin{aligned} \lambda_k^{(t+1)} &= \frac{\sum_{j=1}^{J} p(z_k = 1 \mid y, \theta^{(t)}, \lambda^{(t)})}{J} \\ \mu_k^{(t+1)} &= \frac{\sum_{j=1}^{J} p(z_k = 1 \mid y, \theta^{(t)}, \lambda^{(t)}) y_j}{\sum_{j=1}^{J} p(z_k = 1 \mid y, \theta^{(t)}, \lambda^{(t)})} \\ \sigma_k^{2(t+1)} &= \frac{\sum_{j=1}^{J} p(z_k = 1 \mid y, \theta^{(t)}, \lambda^{(t)})(y_j - \mu_k)^2}{\sum_{j=1}^{J} p(z_k = 1 \mid y, \theta^{(t)}, \lambda^{(t)})} \end{aligned}$$

Now since our log likelihood $l(y \mid \theta, \lambda)$ is what we're trying to maximize in the first place we should iterate until convergence occurs. We can declare convergence when $|l(y \mid \theta^{(t+1)}, \lambda^{(t+1)}) - l(y \mid \theta^{(t)}, \lambda^{(t)})| < \epsilon$, where ϵ is sufficiently small.

For this problem a tolerance of $\epsilon = 10^{-8}$ should be sufficient to determine convergence. We also need to set initial conditions. We'll set them pretty simply here. Let $\lambda^0 = (1/3, 1/3, 1/3)$ and $\sigma^0 = (1, 1, 1)$. We plot the results below:

	(μ_1, μ_2, μ_3)	$(\sigma_1, \sigma_2, \sigma_3)$	$(\lambda_1, \lambda_2, \lambda_3)$	Iterations
True Values	(0,-2,3)	$(1, \sqrt{2}, 4)$	(0.1,0.3,0.6)	
Estimated Values ($n = 120$)	(-5.8,-1.8,4.2)	(1.2,0.50,3.5)	(0.025,0.345,0.63)	488

The Iterations column refers to the number of iterations it took until approximate convergence was reached. The results aren't terrible, but the model doesn't do an adequate job of approximating the standard deviations or the means so we can expect there to be huge spikes in our mixture model. The increased sample size case does a much better job of controlling for this.

The above are just numbers, however, let's see how the values look in comparison to the true Gaussian Mixture.

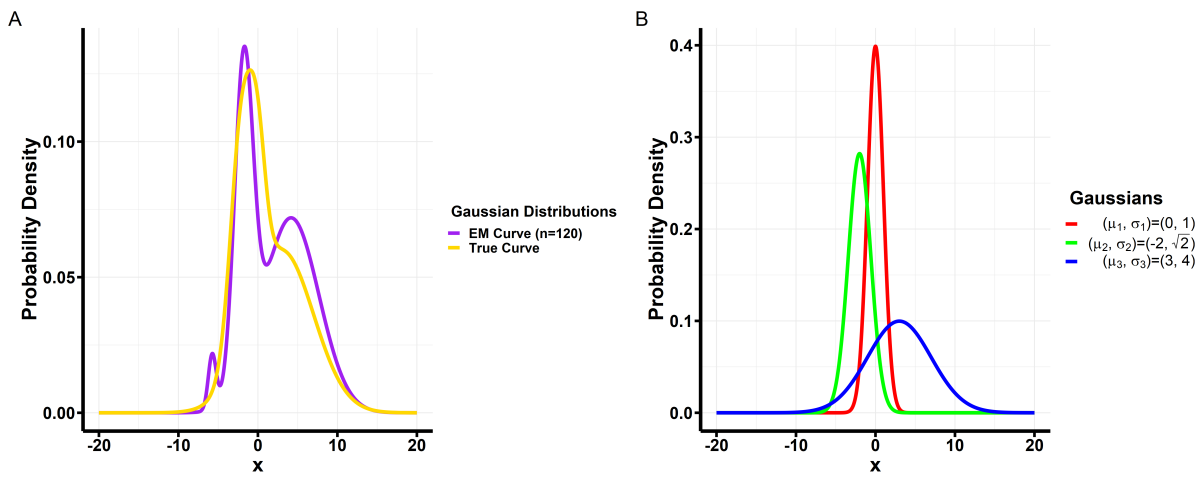

Figure 24. (A) Gaussian Mixture Curve from the EM algorithm vs the True Gaussian Mixture Curve. (B) Individual Gaussians from the true mixture curve

One can see that our runs from the EM algorithm produce an alright approximation to the true mixture curve. This can be attributed to sample size and the fact that our three individual gaussians are hard to tell apart, especially the first and second. There's also the fact that given different starting values and re-running the algorithm, we arrive at different, and often, poor estimates to the true parameters.

(b) MCMC (Gibbs): use proper diffused prior (p_i, μ_i, σ_i^2) implement Gibbs algorithm to find the posterior distribution of (p_i, μ_i, σ_i^2).

Solution:

Starting from the model in part (a) we have that

$$\begin{aligned} y_j \mid z_k, \theta &\sim N(\mu_k, \sigma_k^2) \text{ for } k = 1, 2, \ldots, K \\ z \mid \lambda &\sim Multinomial(1; \lambda_1, \ldots, \lambda_K) \end{aligned}$$

Unlike in part (a) we're not trying to find the MLE so we actually need to assign priors in this case. We want proper diffuse priors so we can make our lives easier by exploiting conjugacy. Since $z \mid \lambda \sim Multinomial(1; \lambda_1, \ldots, \lambda_K)$, let $\lambda \sim Dirichlet(\underbrace{1, 1, \ldots, 1}_{K \text{ times}})$. Likewise let's assume prior independence between our location and scale parameters. We can then assume $\mu_j \sim N(\epsilon, \alpha)$ and $\sigma_j^2 \sim Inv - Gamma(\eta, \nu)$, where ϵ, α, η and ν are fixed constants. Our model can then be described as

$$\begin{aligned} y_j \mid z_k, \theta &\sim N(\mu_k, \sigma_k^2) \text{ for } k = 1, 2, \ldots, K \\ z \mid \lambda &\sim Multinomial(1; \lambda_1, \ldots, \lambda_K) \\ \mu_j &\sim N(\epsilon, \alpha^2) \text{ for } k = 1, 2, \ldots, K \\ \sigma_j^2 &\sim Inv - Gamma(\eta, \nu) \text{ for } k = 1, 2, \ldots, K \end{aligned}$$

The posterior of interest is then

$$
\begin{aligned}
p(\mu, \sigma^2, \lambda \mid y, z) &\propto p(y, z \mid \mu, \sigma^2, \lambda) p(\mu, \sigma^2, \lambda) \\
&= p(y \mid z, \mu, \sigma^2, \lambda) p(z \mid \lambda) p(\mu) p(\sigma^2) p(\lambda) \\
&= p(z \mid \lambda) p(\mu) p(\sigma^2) p(\lambda) \prod_{j=1}^{N} \Big(\prod_{k=1}^{K} p(y_j \mid z_{jk}, \mu, \sigma^2, \lambda) \Big) \\
&\propto \Big(\prod_{k=1}^{K} e^{-\frac{(\mu_k - \epsilon)^2}{2\alpha^2}} \Big) \Big(\prod_{k=1}^{K} (\frac{1}{\sigma_k^2})^{\gamma+1} e^{-\frac{\delta}{\sigma_k^2}} \Big) \Big(\prod_{k=1}^{K} \prod_{j=1}^{N} \lambda_k^{z_{jk}} N(y_j \mid \mu_k, \sigma_k^2)^{z_{jk}} \Big)
\end{aligned}
$$

Now we derive the conditionals in a manner similar to Problems 5 and 6.

$$
\begin{aligned}
p(\lambda \mid \mu, \sigma^2, y, z) &\propto p(z \mid \lambda) p(\lambda) \\
&\propto \Big(\prod_{k=1}^{K} \prod_{j=1}^{N} \lambda_k^{z_{jk}} \Big) \\
&= \Big(\prod_{k=1}^{K} \lambda_k^{\sum_{j=1}^{n_k} z_{jk}} \Big) \\
&\text{where } n_k = \sum_{z_j k=1} z_{jk} \\
&= \Big(\prod_{k=1}^{K} \lambda_k^{(1+\sum_{j=1}^{n_k} z_{jk})-1} \Big) \\
\lambda \mid \mu, \sigma^2, y, z &\sim Dirichlet(n_1 + 1, \ldots, n_K + 1)
\end{aligned}
$$

$$
\begin{aligned}
p(\mu_j \mid \mu_{-j}, \sigma^2, \lambda, y, z) &\propto p(y \mid z, \mu, \sigma^2) p(\mu) \\
&\propto \Big(\prod_{k=1}^{K} e^{-\frac{(\mu_k - \epsilon)^2}{2\alpha^2}} \Big) \Big(\prod_{k=1}^{K} \prod_{j=1}^{N} N(y_j \mid \mu_k, \sigma_k^2)^{z_{jk}} \Big) \\
&\propto \Big(e^{-\frac{(\mu_k - \epsilon)^2}{2\alpha^2}} \Big) \Big(\prod_{j=1}^{N} e^{-\frac{z_{jk}(y_j - \mu_k)^2}{2\sigma_k^2}} \Big) \\
&\propto \Big(e^{-\frac{(\mu_k - \epsilon)^2}{2\alpha^2}} \Big) \Big(e^{-\frac{\sum_{j=1}^{N} z_{jk}(y_j - \mu_k)^2}{2\sigma_k^2}} \Big) \\
&\propto \Big(e^{-\frac{(\mu_k - \epsilon)^2}{2\alpha^2}} \Big) \Big(e^{-\frac{\sum_{\{j: z_{jk}=1\}} (y_j - \mu_k)^2}{2\sigma_k^2}} \Big) \\
\mu_j \mid \mu_{-j}, \sigma^2, \lambda, y, z &\sim N(\frac{\frac{\epsilon}{\alpha^2} + \frac{1}{\sigma_k^2} \sum_{\{j: z_{jk}=1\}} y_j}{\frac{1}{\alpha^2} + \frac{n_k}{\sigma_k^2}}, \frac{1}{\frac{1}{\alpha^2} + \frac{n_k}{\sigma_k^2}})
\end{aligned}
$$

$$
\begin{aligned}
p(\sigma_j^2 \mid \mu, \sigma_{-j}^2, \lambda, y, z) &\propto p(y \mid z, \mu, \sigma^2) p(\sigma^2) \\
&\propto \Big(\prod_{k=1}^{K} (\frac{1}{\sigma_k^2})^{\eta+1} e^{-\frac{\nu}{\sigma_k^2}} \Big) \Big(\prod_{k=1}^{K} \prod_{j=1}^{N} N(y_j \mid \mu_k, \sigma_k^2)^{z_{jk}} \Big) \\
&\propto \Big((\frac{1}{\sigma_k^2})^{\eta+1} e^{-\frac{\nu}{\sigma_k^2}} \Big) \Big(\prod_{j=1}^{N} (\frac{1}{\sigma_k^2})^{\frac{1}{2}} e^{-\frac{z_{jk}(y_j-\mu_k)^2}{2\sigma_k^2}} \Big) \\
&\propto \Big((\frac{1}{\sigma_k^2})^{\eta+1} e^{-\frac{\nu}{\sigma_k^2}} \Big) \Big((\frac{1}{\sigma_k^2})^{\frac{n_k}{2}} e^{-\frac{\sum_{j=1}^{N} z_{jk}(y_j-\mu_k)^2}{2\sigma_k^2}} \Big) \\
\sigma_j^2 \mid \mu, \sigma_{-j}^2, \lambda, y, z &\sim Inv - Gamma(\eta + \frac{n_k}{2}, \nu + \frac{1}{2} \sum_{\{j: z_{jk}=1\}} (y_j - \mu_k)^2)
\end{aligned}
$$

Now we have our three posteriors from which to sample from:

$$
\begin{aligned}
\lambda \mid \mu, \sigma^2, y, z &\sim Dirichlet(n_1 + 1, \ldots, n_K + 1) \\
\mu_j \mid \mu_{-j}, \sigma^2, \lambda, y, z &\sim N(\frac{\frac{\epsilon}{\alpha^2} + \frac{1}{\sigma_k^2} \sum_{\{j: z_{jk}=1\}} y_j}{\frac{1}{\alpha^2} + \frac{n_k}{\sigma_k^2}}, \frac{1}{\frac{1}{\alpha^2} + \frac{n_k}{\sigma_k^2}}) \\
\sigma_j^2 \mid \mu, \sigma_{-j}^2, \lambda, y, z &\sim Inv - Gamma(\eta + \frac{n_k}{2}, \nu + \frac{1}{2} \sum_{\{j: z_{jk}=1\}} (y_j - \mu_k)^2)
\end{aligned}
$$

This is all fine and dandy, but one potential problem we will run into here is the fact that these posteriors depend on the unobserved variable, z. Therefore in each iteration we will need to take draws of z. To do so, recall from part (a) that

$$
\begin{aligned}
p(y, z \mid \theta, \lambda) &= p(z \mid \lambda) p(y \mid z, \theta) \\
&= \prod_{j=1}^{J} \Big(\frac{\Gamma(\sum_{k=1}^{K} z_{jk})}{\prod_{k=1}^{K} z_{jk}!} \prod_{k=1}^{K} \lambda_k^{z_{jk}} \Big) \prod_{j=1}^{J} \Big(\prod_{k=1}^{K} p(y_j \mid \theta_k)^{z_{jk}} \Big) \\
&= \prod_{j=1}^{J} \prod_{k=1}^{K} (\lambda_k p(y_j \mid \theta_k))^{z_{jk}}
\end{aligned}
$$

Using this pdf, our sampling procedure will look like

$$
\begin{aligned}
z &\sim p(y, z \mid \theta, \lambda) \\
\lambda \mid \mu, \sigma^2, y, z &\sim Dirichlet(n_1 + 1, \ldots, n_K + 1) \\
\mu_j \mid \mu_{-j}, \sigma^2, \lambda, y, z &\sim N(\frac{\frac{\epsilon}{\alpha^2} + \frac{1}{\sigma_k^2} \sum_{\{j: z_{jk}=1\}} y_j}{\frac{1}{\alpha^2} + \frac{n_k}{\sigma_k^2}}, \frac{1}{\frac{1}{\alpha^2} + \frac{n_k}{\sigma_k^2}}) \\
\sigma_j^2 \mid \mu, \sigma_{-j}^2, \lambda, y, z &\sim Inv - Gamma(\eta + \frac{n_k}{2}, \nu + \frac{1}{2} \sum_{\{j: z_{jk}=1\}} (y_j - \mu_k)^2)
\end{aligned}
$$

Unlike in Problems 5 and 6 we need to initialize each variable since the distribution of the latent variable z depends on all of them.

We decide to run the Gibbs Sampler for $N = 50,000$ iterations. It takes ≈ 58 seconds to run with $n = 120$ samples. Our values are set as $\eta = 2$, $\nu = 2$, $\alpha = 3$, and $\epsilon = 0$. Below we plot Trace Plots and Autocorrelation Plots for the Gibbs Sampler to determine convergence.

Trace Plots

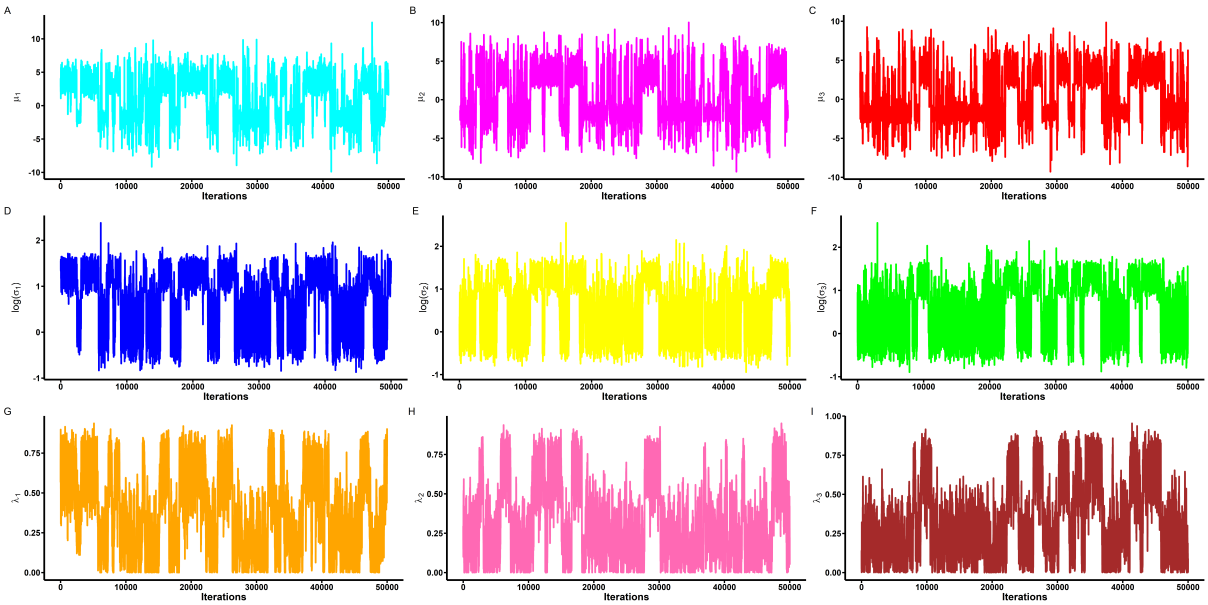

Figure 25. Trace Plots for our nine different parameters. Notice σ and τ are on the log scale.

Autocorrelation Plots

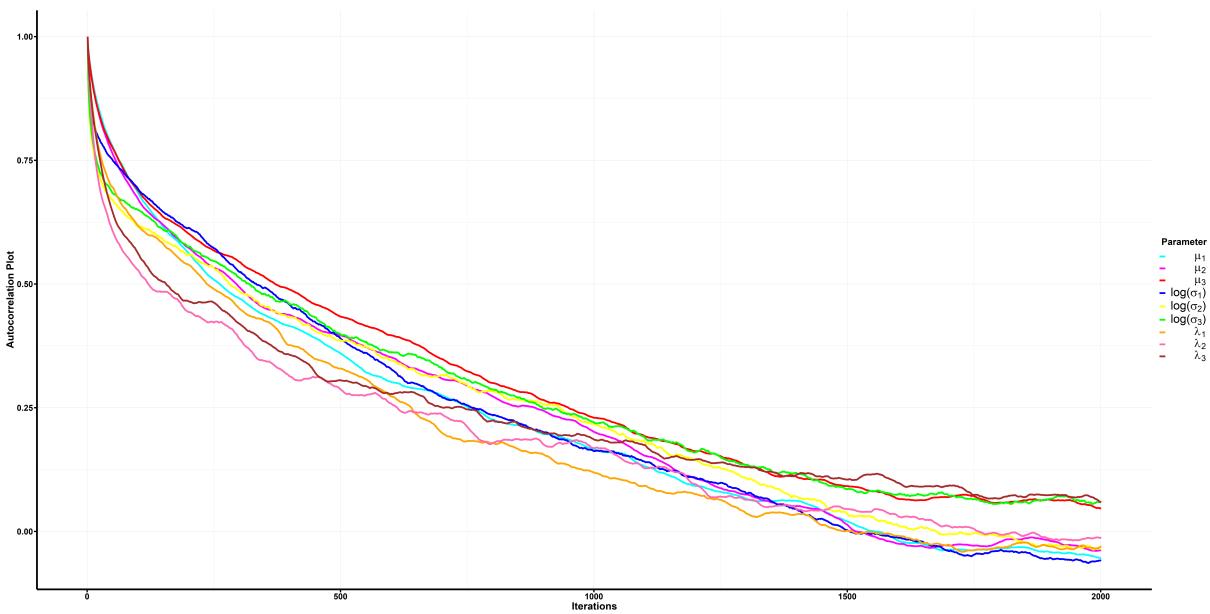

Figure 26. Autocorrelation Plots for our nine different parameters. Notice σ and τ are on the log scale.

Overall convergence for our Gibbs Sampler is extremely poor. From the trace plots one can clearly see that mixing is extremely poor since the chain continuously jumps around the distribution and stays in those areas for prolonged periods of time. This is in sharp contrasts to the previously good trace plots that we have been deriving where the distribution stays in areas of high probability, but jumps to areas of low probability occasionally for brief periods of time.The poor mixing is compounded by the autocorrelation plots which show that even after 500 iterations, our starting values are still having undue influence on where the chain is. In other words we're probably seeing the trace plots that we're observing due to the fact that once we occasionally jump to a different place in the distribution, that new point influences the subsequent iterations to a substantial degree where it essentially locks the chain in place until we jump again.

There are a myriad of reasons for why this is the case. Firstly, from part a we noticed that these Gaussian's are hard to distinguish in the first place which makes it difficult to pinpoint the exact distributions of the mixture components. Not only that but we're only given 120 observations to work with, and 40% of those observations come from the hardest to distinguish Gaussians. Since such a low amount of observations contribute to the first Gaussian, which is difficult to distinguish from the second, this can lead to no observations being assigned to that particular category meaning the posteriors essentially get all their information from the priors, which are weakly informative.

(c) Summarize the results and compare between the two approaches, any problems observed during the sampling or maximum likelihood estimator.

Solution:

The results have been duly summarized in parts (a) and (b) so there is nothing to discuss here.

Problem 9: BDA 3rd Ed. 15.3 (Extra Credit)

Regression with many explanatory variables: Table 15.2 displays data from a designed experiment for a chemical process. In using these data to illustrate various approaches to selection and estimation of regression coefficients, Marquardt and Snee (1975) assume a quadratic regression form; that is, a linear relation between the expectation of the untransformed outcome, y, and the variables x_1, x_2, x_3, their two-way interactions, x_1x_2, x_1x_3, x_2x_3, and their squares, x_1^2, x_2^2, x_3^2.

Table 3. Data from a chemical experiment, from Marquardt and Snee (1975). The first three variables are experimental manipulations, and the fourth is the outcome measurement.

Reactor Temperature $(C), x_1$	Ratio of H_2 to n-Heptane (mole ratio), x_2	Contact Time (sec), x_3	Conversion of n-heptane to acetylene (%), y
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	19.5

(a) Fit an ordinary linear regression model (that is, nonhierarchical with a uniform prior distribution on the coefficients), including a constant term and the nine explanatory variables above

Solution:

An ordinary linear regression is an equation of the form:

$$y = X^T\beta + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

unlike in a frequentist setting we don't assume that β and σ^2 are fixed parameters. Instead we assume them to be random variables. Given the specifications of the problem our model at hand begins with

$$y \mid X, \beta, \sigma^2 \sim N(X^T\beta, \sigma^2 I_n)$$
$$p(\beta, \sigma^2 \mid X) \propto \frac{1}{\sigma^2}$$

This then provides us with the following posterior distribution:

$$
\begin{aligned}
p(\beta, \sigma^2 \mid y, X) &\propto p(y \mid X, \beta, \sigma^2) p(X, \beta, \sigma^2) \\
&= p(y \mid X, \beta, \sigma^2) p(\beta, \sigma^2 \mid X) p(X) \\
&\propto p(y \mid X, \beta, \sigma^2) p(\beta, \sigma^2 \mid X) \\
&\propto \frac{1}{\sigma^2} \prod_{j=1}^{N} \left(\frac{1}{\sigma^2}\right) e^{-\frac{(y_j - X_j^T\beta)(y_j - X_j^T\beta)}{2\sigma^2}} \\
&= \left(\frac{1}{\sigma^2}\right)^{N+1} e^{-\sum_{j=1}^{N} \frac{(y_j - X_j^T\beta)(y_j - X_j^T\beta)}{2\sigma^2}} \\
&= \left(\frac{1}{\sigma^2}\right)^{N+1} e^{-\frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}} \\
&= \left(\frac{1}{\sigma^2}\right)^{N+1} e^{-\frac{(y^Ty - 2y^TX\beta - \beta^TX^TX\beta)}{2\sigma^2}} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{N+1} e^{-\frac{(y^Ty - 2y^TX(X^TX)^{-1}X^TX\beta - \beta^TX^TX\beta)}{2\sigma^2}} \\
&\text{let } \hat{\beta} = (X^TX)^{-1}X^Ty \\
&= \left(\frac{1}{\sigma^2}\right)^{N+1} e^{-\frac{(y^Ty - 2\hat{\beta}X^TX\beta - \beta^TX^TX\beta + \hat{\beta}^TX^TX\hat{\beta} - \hat{\beta}^TX^TX\hat{\beta})}{2\sigma^2}} \\
&= \left(\frac{1}{\sigma^2}\right)^{N+1} e^{-\frac{(y^Ty - \hat{\beta}^TX^TX\hat{\beta} + (\hat{\beta} - \beta)^T[(X^TX)^{-1}]^{-1}(\hat{\beta} - \beta))}{2\sigma^2}} \\
&= \left(\frac{1}{\sigma^2}\right) e^{-\frac{(\hat{\beta} - \beta)^T[(X^TX)^{-1}]^{-1}(\hat{\beta} - \beta)}{2\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{N} e^{-\frac{(y^Ty - \hat{\beta}^TX^TX\hat{\beta} + 2y^TX\hat{\beta} - 2y^TX\hat{\beta})}{2\sigma^2}} \\
&k = rank(X) \\
&\propto \left(\frac{1}{\sigma^2}\right) e^{-\frac{(\hat{\beta} - \beta)^T[(X^TX)^{-1}]^{-1}(\hat{\beta} - \beta)}{2\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{N} e^{-\frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{2\sigma^2}} \\
&\propto \left(\frac{1}{\sigma^2}\right) e^{-\frac{(\hat{\beta} - \beta)^T[(X^TX)^{-1}]^{-1}(\hat{\beta} - \beta)}{2\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{N} e^{-\frac{(N-k)s^2}{2\sigma^2}} \\
&\text{where } s^2 = \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{N - k}
\end{aligned}
$$

This provides us with two posteriors:

$$
\begin{aligned}
\beta \mid y, X, \sigma^2 &\sim N(\hat{\beta}, \sigma^2(X^TX)^{-1}) \\
\sigma^2 \mid y, X &\sim Inv - \chi^2(N - k, s^2)
\end{aligned}
$$

Of course, though, this derivation makes the assumption that $(X^TX)^{-1}$ exists. If X^TX is not invertible we would then proceed using the generalized inverse.

Given that we have analytical distributions to sample from there is no need for any MCMC method as we can just sample from the posterior directly. Doing so provides the following statistics for our parameters:

	Mean	Variance	2.5%	25%	50%	75%	97.5%
β_0	35.48	0.392	34.69	35.24	35.47	35.71	36.29
β_{x_1}	932.45	670.32	-377.56	542.57	929.90	1319.55	2324.37
β_{x_2}	178.53	40.84	96.61	152.82	177.41	203.47	263.26
β_{x_3}	819.66	563.58	-281.35	489.426	808.16	1154.07	1977.45
$\beta_{x_1x_2}$	-164.48	38.61	-245.20	-187.68	-163.38	-140.28	-87.53
$\beta_{x_1x_3}$	-665.18	474.05	-1648.99	-945.38	-655.24	-389.38	251.66
$\beta_{x_2x_3}$	-22.39	6.05	-34.52	-26.09	-22.25	-18.65	-10.21
$\beta_{x_1^2}$	-836.49	627.20	-2138.34	-1191.64	-836.56	-472.46	390.15
$\beta_{x_2^2}$	-6.95	3.30	-13.75	-8.82	-6.907	-4.97	-0.34
$\beta_{x_3^2}$	-69.95	44.74	-163.77	-96.44	-69.12	-44.06	18.55
σ	2.41	2.07	0.661	1.25	1.81	2.86	7.88

(b) Fit a mixed-effects linear regression model with a uniform prior distribution on the constant term and a shared normal prior distribution on the coefficients of the nine variables above. If you use iterative simulation in your computations, be sure to use multiple sequences and monitor their joint convergence.

Solution:

As discussed in section 15.2 of Bayesian Data Analysis, 3rd Edition, given your J+1 parameters, your intercept and J covariates, a linear mixed effects model is a model where one puts an improper prior on the first J_1 parameters (we assume independence amongst the improper priors) and we then assume exhangeability on the remaining J_2 parameters where they share a prior with mean μ and standard deviation τ. Obviously we have that $J+1 = J_1 + J_2$. Written mathematically our model looks like

$$\begin{aligned} y \mid X, \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I_{J+1}) \\ p(\sigma^2) &\propto \frac{1}{\sigma^2} \\ p(\beta_{0:J_1}) &\propto 1 \quad \text{for } j = 1, \ldots, J_1 \\ \beta_{-0:J_1} \mid \mu, \tau &\sim N(\vec{1}_{J+1-J_1}\mu, \tau^2 I_{J_2}) \quad \text{for } j = J_1+1, \ldots, J+1 \\ p(\mu, \tau) &\propto 1 \end{aligned}$$

$\beta_{0:J_1}$ refers to the parameters from the intercept to the J_1^{th} covariate. $\beta_{-0:J_1}$ refers to all parameters on the covariates exlcuding $\beta_{0:J_1}$. We assume our design matrix X is $N \times (J+1)$ where the first column is just a column of 1's, corresponding to the intercept term. The great thing about this model is that we can derive analytical conditional posteriors for this model! To begin let's start with the β parameters. Our only concern will be with the likelihood

Previously in part (a) we saw that we can separate β from X and get $(\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta})$. This is the only term from the likelihood that we'll care about. Notice that β here is $(J+1) \times 1$. The term $\beta_{-0:J_1}$ in $(\beta_{-0:J_1} - \vec{1}_{J_2}\mu)^T(\beta_{-0:J_1} - \vec{1}_{J_2}\mu)$ is $J_2 \times 1$ so we need to fix some things before continuing. Define the design matrix $\Sigma_{(J_2-J_1)}$

$$\Sigma_{J_2} = \begin{pmatrix} 0 & \vec{0} \\ \vec{0}^T & I_{J_2} \end{pmatrix}.$$

This is a $J \times J$ matrix where everything is zero, except the bottom right which is a $J_2 \times J_2$ identity matrix. It is easy to see that

$$(\beta_{-0:J_1} - \vec{1}_{J_2}\mu)^T(\beta_{-0:J_1} - \vec{1}_{J_2}\mu) = (\beta - \vec{1}_J\mu)^T\Sigma_{(J_2)}(\beta - \vec{1}_J\mu)$$

Completing the square will result in the following conditional posterior on the β parameters

$$\beta \mid \sigma^2, \mu, \tau^2, y, X \sim N(\tilde{\beta}, V_\beta)$$
$$\tilde{\beta} = (\frac{\mu}{\tau^2}\vec{1_J}^T\Sigma_{(J_2-J_1)} + \frac{1}{\sigma^2}y^TX)V_\beta$$
$$V_\beta = (\frac{1}{\tau^2}\Sigma_{(J_2-J_1)} + \frac{1}{\sigma^2}X^TX)^{-1}$$

One might question how we know that V_β is invertible in the first place since $\Sigma_{(J_2-J_1)}$ is obviously not invertible. To that end one only needs to notice that we just need to require that X^TX have full rank. Here is a short proof. We know if am $n \times n$ matrix A is positive definite, then $\forall x \in \mathbb{R}^n$, $x^TAx > 0$. Our first claim is that X^TX is positive semi-definite. Take any $v \in \mathbb{R}^n$, then

$$v^T(X^TX)v = (Xv)^T(Xv) = y^Ty = \|y\|^2 = \sum_{j=1}^{n} y_j^2 \geq 0$$

If X has full rank this inequality is strict. Now take $v \in \mathbb{R}^n$

$$v^T(\frac{1}{\tau^2}\Sigma_{J_2} + \frac{1}{\sigma^2}X^TX)v = v^T\frac{1}{\tau^2}\Sigma_{J_2}v + v^T\frac{1}{\sigma^2}X^TXv$$
$$> 0$$

This follows since due to the structure of Σ, $v^T\frac{1}{\tau^2}\Sigma_{J_2}v \geq 0$. Of course, though, if X doesn't have full rank then this obviously fails and other methods have to be employed. Now we derive the conditional posterior for σ.
We will begin with σ. This is pretty straightforward,

$$p(\sigma^2 \mid \beta, \mu, \tau^2, y, X) \propto p(y \mid X, \beta, \sigma^2)p(\sigma^2)$$
$$\propto \left(\frac{1}{\sigma^2}\right)\left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{\|y-X\beta\|^2}{2\sigma^2}}$$
$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}+1} e^{-\frac{Ns^2_{\sigma^2}}{2\sigma^2}}$$
$$\text{where } s^2_{\sigma^2} = \frac{\|y - X\beta\|^2}{N}$$

This gives us

$$\sigma^2 \mid \beta, \mu, \tau, \lambda^2, y, X \sim Inv - \chi^2(N, s^2_{\sigma^2})$$

The conditional posteriors for μ and τ^2 will follow relatively simply from this as well, in a form akin to the conditional posteriors in Problem 5. We will start with μ. Our only concern is the term $(\beta_{-0:J_1} - \vec{1}_{J_2}\mu)^T(\beta_{-0:J_1} - \vec{1}_{J_2}\mu)$. When re-arranged we get that

$$(\beta_{-0:J_1} - \vec{1}_{J_2}\mu)^T(\beta_{-0:J_1} - \vec{1}_{J_2}\mu) = \frac{1}{2\tau^2}\Big(\vec{1}_{J_2}^T\vec{1}_{J_2}\mu^2 - 2\vec{1}_{J_2}\beta_{-0:J_1})\mu + \text{ other terms}\Big)$$

This gives us that the conditional posterior distribution will be

$$\mu \mid \beta, \sigma, \tau, y \sim N(\frac{1}{J_2}\sum_{j=J_1+1}^{J}\beta_j, \frac{\tau^2}{J_2})$$

Likewise for τ^2 we can go through a similar derivation to obtain

$$\tau^2 \mid \beta, \mu, \sigma, y \sim Inv - \chi^2(J_2, \frac{\sum_{j=J_1+1}^{J}(\beta_j - \mu)^2}{J_2})$$

Since we only consider the intercept fixed, $J_1 = 1$ and $J_2 = J$ in this case. Likewise given our data, $N = 16$, $J = 9$. As mentioned in Problem's 5 and 6, we can be smart in our smapling procedure by arranging the conditional posteriors in a specific way. e can simply initialize β and μ to go through with our sampling procedure which just amounts to initializing β since we can assign the mean of our initial β to μ.

$$\sigma^2 \mid \beta, \mu, \tau, \lambda^2, y, X \sim Inv - \chi^2(N, s_{\sigma^2}^2)$$
$$\tau^2 \mid \beta, \mu, \sigma, y, X \sim Inv - \chi^2(J-1, \frac{\sum_{j=1}^{J}(\beta_j - \mu)^2}{J-1})$$
$$\mu \mid \beta, \sigma, \tau, y, X \sim N(\frac{1}{J}\sum_{j=1}^{J}\beta_j, \frac{\tau^2}{J})$$
$$\beta \mid \sigma^2, \mu, \tau^2, y, X \sim N(\tilde{\beta}, V_\beta)$$

The results from the regression are below in the table:

	Mean	Standard Deviation	2.5%	25%	50%	75%	97.5%
β_0	35.48	0.69	34.11	35.05	35.49	35.92	36.87
β_{x_1}	5.27	4.83	-3.72	2.59	5.08	7.73	15.29
β_{x_2}	2.26	5.11	-6.21	-0.60	1.86	4.55	13.10
β_{x_3}	-2.16	5.34	-13.26	-4.88	-2.05	0.74	8.09
$\beta_{x_1x_2}$	0.18	4.93	-10.00	-2.15	0.42	2.85	9.04
$\beta_{x_1x_3}$	-1.81	5.00	-11.69	-4.60	-1.89	0.90	8.48
$\beta_{x_2x_3}$	2.57	1.87	-1.15	1.40	2.60	3.78	6.23
$\beta_{x_1^2}$	4.86	4.62	-4.42	2.47	4.85	7.35	13.90
$\beta_{x_2^2}$	-2.72	3.05	-9.10	-4.59	-2.61	-0.73	3.04
$\beta_{x_3^2}$	-0.06	3.33	-6.62	-2.07	-0.12	1.88	6.80
σ	2.69	0.68	1.73	2.22	2.57	3.02	4.33
μ	0.92	2.19	-3.36	-0.26	0.90	2.08	5.34
τ	5.56	2.99	2.37	3.72	4.82	6.51	13.03

(c) Discuss the differences between the inferences in (a) and (b). Interpret the differences in terms of the hierarchical variance parameter. Do you agree with Marquardt and Snee that the inferences from (a) are unacceptable?

Solution:

The largest difference between the results of the two models is the magnitude of the distributions attached to each. The ordinary Bayesian Linear Regression model assumes noninformative uniform priors on each of the covariates which has the added effect of causing extraordinarily large variances attached to the parameters leading to larger 95% credible intervals. The linear mixed effects regression, on the other hand, attaches more reasonable distributions to the distributions due to the shared normal prior which draws the posterior means closer together, which is evident by the fact that all the posterior means lie between -3 and 6.

(d) Repeat (a), but with a t_4 prior distribution on the nine variables.

Solution:

We repeat the analysis from part (a), but now we assume $\beta_j \sim t_4$ for j = 1, . . . , 9.

$$
\begin{aligned}
y \mid X, \beta, \sigma^2 &\sim N(X^T\beta, \sigma^2 I_n) \\
p(\beta_j) &\propto 1 \quad \text{for } j = 0 \\
\beta_j \sim t_4 &\quad \text{for } j = 1, \ldots, 9 \\
p(\sigma^2) &\propto \frac{1}{\sigma^2}
\end{aligned}
$$

The interesting part about this model is that we actually don't need to use external software. Like parts (a) and (b) we can get analytical forms for the posteriors and do a simple Gibbs sampling. For that purpose I think it might be fruitful to derive it for the general case when using a Generalized t-distribution for ν degrees of freedom. Since this is also an extra credit problem I think it will be fine if I don't get straight to the point.

One interesting fact about $t_\nu(\mu, \tau^2)$ is that we can express it in the following way:

$$
\begin{aligned}
t_\nu(\mu, \tau^2) &= \mu + \tau\sqrt{\frac{\nu}{V}}Z \\
Z &\sim N(0, 1) \\
V &\sim \chi^2_\nu
\end{aligned}
$$

Essentially we can turn this into a missing data problem and derive the conditional posteriors in a similar manner as in problem 8. Essentially $t_\nu(\mu, \tau^2) \sim N(\mu, \tau^2\frac{\nu}{V}) = W$ so denote $t_\nu(\mu, \tau^2) = W \mid \mu, \tau, V$. One quick thing to note is that we can also introduce a new variable $\lambda^2 \sim Inv - \chi^2(\nu, \tau^2)$ due to the fact that $\tau^2\frac{\nu}{V}$ has a familiar distribution. Given this formulation we can actually take a quick cue from Problem 6 and simplify our analysis substantially.

$$
\begin{aligned}
y \mid X, \beta, \sigma^2 &\sim N(X^T\beta, \sigma^2 I_n) \\
p(\beta_0) &\propto 1 \\
p(\sigma^2) &\propto \frac{1}{\sigma^2} \\
\beta &\sim N(\mu\vec{1}, \lambda^2 I) \\
p(\mu, \tau) &\propto 1 \\
\lambda^2 &\sim Inv - \chi^2(\nu, \tau^2)
\end{aligned}
$$

The nice part about this model is that most of the conditional posteriors have been derived in part (b). For instance, the posterior of σ^2 will be the same, and with slight tweaking (i.e. replacing τ^2 with λ^2) our posteriors for μ and β are the same. We list them below:

$$
\begin{aligned}
\beta \mid \sigma^2, \mu, \tau^2, y, X &\sim N(\tilde{\beta}, V_\beta) \\
\tilde{\beta} &= (\frac{\mu}{\lambda^2}\vec{1_J}^T\Sigma_{(J_2-J_1)} + \frac{1}{\sigma^2}y^T X)V_\beta \\
V_\beta &= (\frac{1}{\lambda^2}\Sigma_{(J_2-J_1)} + \frac{1}{\sigma^2}X^T X)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2 \mid \beta, \mu, \tau, \lambda^2, y, X) &\sim Inv - \chi^2(N, s^2_{\sigma^2}) \\
\text{where } s^2_{\sigma^2} &= \frac{\|y - X\beta\|^2}{N}
\end{aligned}
$$

$$
\mu \mid \beta, \sigma, \tau, y \sim N(\frac{1}{J_2 - J_1}\sum_{j=J_1+1}^{J}\beta_j, \frac{\lambda^2}{J_2 - J_1})
$$

The only elbow grease that we need to put in is for the posteriors of λ^2 and τ^2.

$$
\begin{aligned}
p(\lambda^2 \mid \beta, \sigma^2, \mu, \tau^2, y, X) &\propto p(\beta_{-0} \mid \mu, \lambda^2)p(\lambda^2 \mid \tau^2) \\
&\propto \frac{1}{(\lambda^2)^{J/2}}e^{-\frac{\|\beta_{-0}-\mu\vec{1}_{-0}\|^2}{2\lambda^2}}\frac{1}{(\lambda^2)^{1+\nu/2}}e^{-\frac{\nu\tau^2}{2\lambda^2}} \\
&\propto \frac{1}{(\lambda^2)^{1+(J+\nu)/2}}e^{-\frac{(J+\nu)}{2\sigma_j^2}\frac{1}{J+\nu}(\|\beta_{-0}-\mu\vec{1}_{-0}\|^2+\nu\tau^2)}
\end{aligned}
$$

$$
\lambda^2 \mid \beta, \sigma^2, \mu, \tau^2, y, X \sim Inv - \chi^2(\nu + J, \frac{1}{\nu + J}\Big(\|\beta_{-0} - \mu\vec{1}_{-0}\|^2 + \nu\tau^2\Big))
$$

As for τ, since the prior on τ is flat we find that the conditional posterior for τ is just the parts of the inverse χ^2 distribution of λ^2. We'll have to do grid sampling since the conditional posterior is of the form,

$$p(\tau^2 \mid \beta, \sigma^2, \mu, \lambda^2, y, X) \propto (\tau^2)^{\nu/2} e^{-\frac{\nu\tau^2}{2\lambda^2}}$$

We set this problem up by examining dependencies. λ^2 is dependent upon β, τ and μ. τ is dependent on λ. μ is dependent on λ^2 and β. Likewise σ^2 depends on β.Therefore we only initialize the β and λ parameters. μ can be initialized by the mean of the β initialization. Our order of sampling will be

$$\begin{aligned}
p(\tau^2 \mid \beta, \sigma^2, \mu, \lambda^2, y, X) &\propto (\tau^2)^{\nu/2} e^{-\frac{\nu\tau^2}{2\lambda^2}} \\
\sigma^2 \mid \beta, \mu, \tau, \lambda^2, y, X &\sim Inv - \chi^2(N, s^2_{\sigma^2}) \\
\lambda^2 \mid \beta, \sigma^2, \mu, \tau^2, y, X &\sim Inv - \chi^2(\nu + J, \frac{1}{\nu + J}\Big(\|\beta_{-0} - \mu\vec{1}_{-0}\|^2 + \nu\tau^2\Big)) \\
\mu \mid \beta, \sigma, \tau, y &\sim N(\frac{1}{J}\sum_{j=1}^{J}\beta_j, \frac{\lambda^2}{J}) \\
\beta \mid \sigma^2, \mu, \tau^2, y, X &\sim N(\tilde{\beta}, V_\beta)
\end{aligned}$$

	Mean	Standard Deviation	2.5%	25%	50%	75%	97.5%
β_0	35.493	0.696	34.083	35.058	35.490	35.922	36.918
β_{x_1}	4.924	3.292	-1.494	2.834	4.910	6.925	11.318
β_{x_2}	1.516	3.232	-4.564	-0.584	1.401	3.447	8.517
β_{x_3}	-1.802	3.565	-8.830	-3.990	-1.756	0.384	5.381
$\beta_{x_1x_2}$	0.465	3.194	-6.022	-1.398	0.460	2.464	6.720
$\beta_{x_1x_3}$	-1.892	3.456	-8.910	-4.142	-1.867	0.296	4.851
$\beta_{x_2x_3}$	2.370	1.726	-0.919	1.267	2.332	3.476	5.777
$\beta_{x_1^2}$	5.030	3.192	-1.392	3.114	5.094	7.098	11.092
$\beta_{x_2^2}$	-2.174	2.533	-7.597	-3.703	-2.155	-0.528	2.604
$\beta_{x_3^2}$	-0.267	2.519	-5.200	-1.929	-0.273	1.318	4.740
σ	2.693	0.651	1.758	2.237	2.576	3.017	4.262
λ	3.930	1.252	2.271	3.095	3.714	4.514	6.838
μ	0.922	1.218	-1.518	0.169	0.932	1.679	3.210
τ	3.185	1.180	1.359	2.346	3.010	3.838	6.044

Of course, though, the above is just a generalization for the t prior. In our case we have that $\mu = 0$, $\tau = 1$, and $\nu = 4$ so our posterior for τ^2 is irrelevant. Our model is then described below:

$$\begin{aligned}
y \mid X, \beta, \sigma^2 &\sim N(X^T\beta, \sigma^2 I_n) \\
p(\beta_0) &\propto 1 \\
p(\sigma^2) &\propto \frac{1}{\sigma^2} \\
\beta &\sim N(\mu\vec{1}, \lambda^2 I) \\
\lambda^2 &\sim Inv - \chi^2(\nu = 4, \tau^2 = 1)
\end{aligned}$$

Again, we set this problem up by examining dependencies. λ^2 is dependent upon β and μ. μ is dependent on λ^2 and β. Likewise σ^2 depends on β.Therefore we only initialize the β parameters. μ can be initialized by the mean of the β initialization. Our order of sampling will be

$$\sigma^2 \mid \beta, \mu, \tau, \lambda^2, y, X) \sim Inv - \chi^2(N, s^2_{\sigma^2})$$
$$\lambda^2 \mid \beta, \sigma^2, \mu, \tau^2, y, X \sim Inv - \chi^2(\nu + J, \frac{1}{\nu + J}\Big(\|\beta_{-0} - \mu\vec{1}_{-0}\|^2 + \nu\tau^2\Big))$$
$$\beta \mid \sigma^2, \mu, \tau^2, y, X \sim N(\tilde{\beta}, V_\beta)$$

	Mean	Standard Deviation	2.5%	25%	50%	75%	97.5%
β_0	35.510	0.778	33.923	35.056	35.507	35.973	37.089
β_{x_1}	3.991	2.384	-0.398	2.459	3.919	5.453	8.845
β_{x_2}	0.973	2.343	-3.542	-0.478	0.847	2.335	5.804
β_{x_3}	-2.106	2.506	-7.311	-3.572	-2.106	-0.576	2.631
$\beta_{x_1x_2}$	0.644	2.238	-3.770	-0.749	0.669	2.001	5.335
$\beta_{x_1x_3}$	-2.143	2.460	-6.987	-3.652	-2.129	-0.617	2.594
$\beta_{x_2x_3}$	1.412	1.739	-2.161	0.243	1.473	2.595	4.701
$\beta_{x_1^2}$	4.171	2.303	-0.165	2.665	4.093	5.640	9.099
$\beta_{x_2^2}$	-1.286	2.085	-5.905	-2.510	-1.173	0.101	2.620
$\beta_{x_3^2}$	-0.443	2.014	-4.206	-1.786	-0.477	0.734	3.746
σ	2.928	0.806	1.799	2.348	2.780	3.323	4.924
λ	2.697	0.907	1.469	2.066	2.539	3.120	5.011

(e) Discuss other models for the regression coefficients.

Solution:

Other possible models for the regression coefficients could involve regularization. We could use LASSO (L1 Regularization) to shrink some of the coefficients ot 0, or use Ridge (L2 Regularization) to shrink down the sizes of the coefficients).