Homework 2

Claire Kelling, Statistics 540

March 16, 2018

Problem 1

(Return to the previous homework problem.) Let $X_1, \ldots, X_n$ iid from Gamma(α, β) distribution. Consider Bayesian inference for α, β with prior distributions Normal(0,3) for both. The data for this problem are available here: http://personal.psu.edu/muh10/540/Rcode/hw1.dat

From Homework 1 and the description above, we see that

$$X|\alpha, \beta \sim Gamma(\alpha, \beta)$$

$$\alpha \sim N(0, 3)$$

$$\beta \sim N(0, 3)$$

Because of Bayes rule, we have that $f(\alpha, \beta|X) \propto f(X|\alpha, \beta)\pi(\alpha)\pi(\beta)$.
We know these distributions, $f(X|\alpha, \beta)$, $\pi(\alpha)$, and $\pi(\beta)$ as mentioned above. Therefore, we can calculate

$$f(\alpha, \beta|X) \propto f(X|\alpha, \beta)\pi(\alpha)\pi(\beta)$$

$$\propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} x_i^{\alpha-1} exp(-\beta \sum_{i=1}^{n} x_i) * exp(\frac{\alpha^2}{6}) * exp(\frac{\beta^2}{6})$$

$$\propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} x_i^{\alpha-1} exp(-\beta \sum_{i=1}^{n} x_i + \frac{\alpha^2}{6} + \frac{\beta^2}{6})$$

So, for the log likelihood, we have,

$$log(f(\alpha, \beta|X)) \propto n\alpha log(\beta) - nlog(\Gamma(\alpha)) + \alpha log(\sum_{i=1}^{n} x_i) - \beta \sum_{i=1}^{n} x_i - \frac{\alpha^2}{6} - \frac{\beta^2}{6}$$

Before I begin, I will note that for most of these problems, I use an exponential proposal distribution for the parameter, so that the support of the distribution matches the support of the parameter. However, I tried a couple other distributions, including the gamma distribution, and found that the exponential distribution gave the best results. Therefore, that is what is included in this report.

(a) Use importance sampling to approximate the expectations. Provide pseudo-code, including relevant distributions. Is the variability of your Monte Carlo approximation guaranteed to be finite? Explain your answer. Provide Monte Carlo standard errors for your estimates.

Following the notes from class, we know that we want to estimate $\mu = E_f(g(x))$ which is an expectation of a function g with respect to the probability distribution f(x). We let q(x) be a distribution from which we can draw iid samples. We require that q(x)>0 whenever g(x)f(x)>0. We then see that

$$E_f(g(x)) = \int f(x)g(x)dx = \int \frac{g(x)f(x)}{q(x)} q(x)dx = E_q(g(x)\frac{f(x)}{q(x)}).$$

So, the importance sampling estimator based on iid sample $X_1, ..., X_n \sim q$ is therefore $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)f(X_i)}{q(X_i)}$.

Pseudo-code:

1. Set initial values for the importance sampling as the estimates from the Laplace sampling technique.
2. First, we will use importance sampling for α.

(a) Compute the log-posterior, shown above, log.post. Use the value of β that was the last value in the MC samples.
(b) Compute the log proposal density. For this case, we use an **exponential** distribution, as this has the same support as the posterior, log.prop. For the rate, we use the value of the last MC sample for α.
(c) Compute the log of the importance function, log.imp = log.post - log.prop.
(d) Generate 1,000 samples from the proposal distribution (exponential). Call these samples U. We use the rate as the last Monte Carlo sample, beginning with the initial value.
(e) Calculate the vector of values of LP = log.imp(U).
(f) Then, the importance sampling estimate is $\frac{mean(exp(LP)*U)}{mean(exp(LP))} = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)f(X_i)}{q(X_i)}$.
(g) Repeat this for the number of samples used in all exercises, n.samp.

3. Repeat this calculation for β. Use the value of α for the posterior as the last value of α in the MC samples. We also use an exponential distribution as the proposal for β.
4. Assess diagnostics post burnin and adjust initial values if needed.

For this problem, I will assess some diagnostics and then report the MC estimates and standard errors for importance sampling.

First, I plot the trace plots and assess if I need to remove some values as burnin. I plot the trace plots and I see there is pretty good mixing, and I see the same in the ACF plots. I remove a relatively small burnin of 500 values for both α and β samples. In Figure 1, we see the trace plots for both α and β after this small burnin has been removed. We see that quite a bit of mixing is occurring through the trace plots, and it seems to satisfy independence.

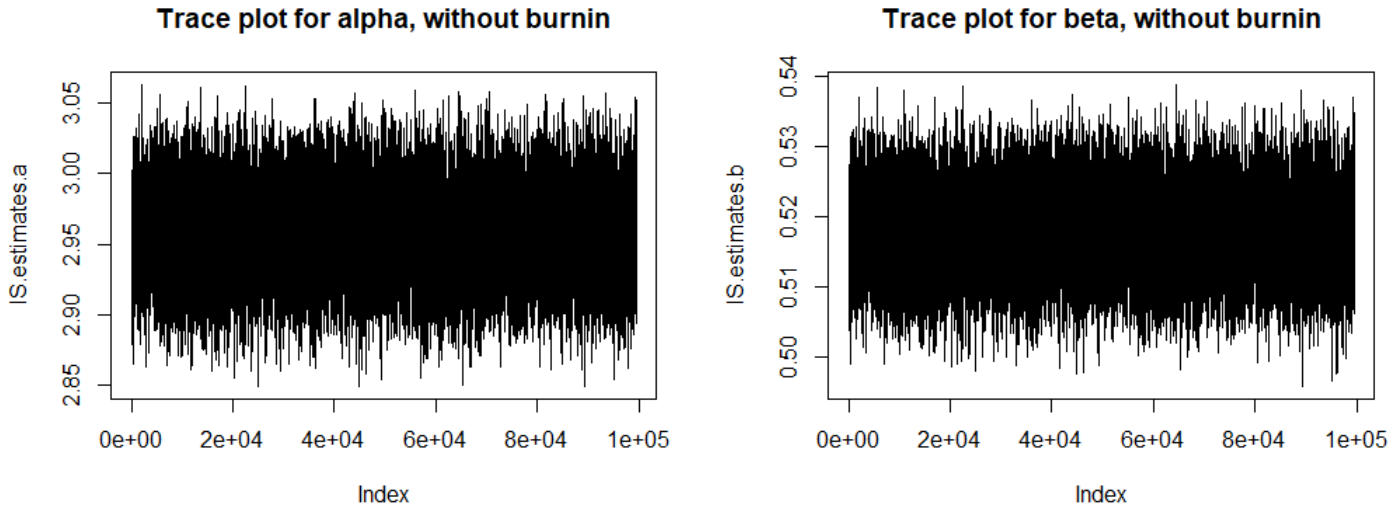

Figure 1: Trace plots for α and β, Importance Sampling

In Figure 2, we see that the ACF for both α and β decay pretty quickly. Therefore, there is no large concern with the ACF plots.

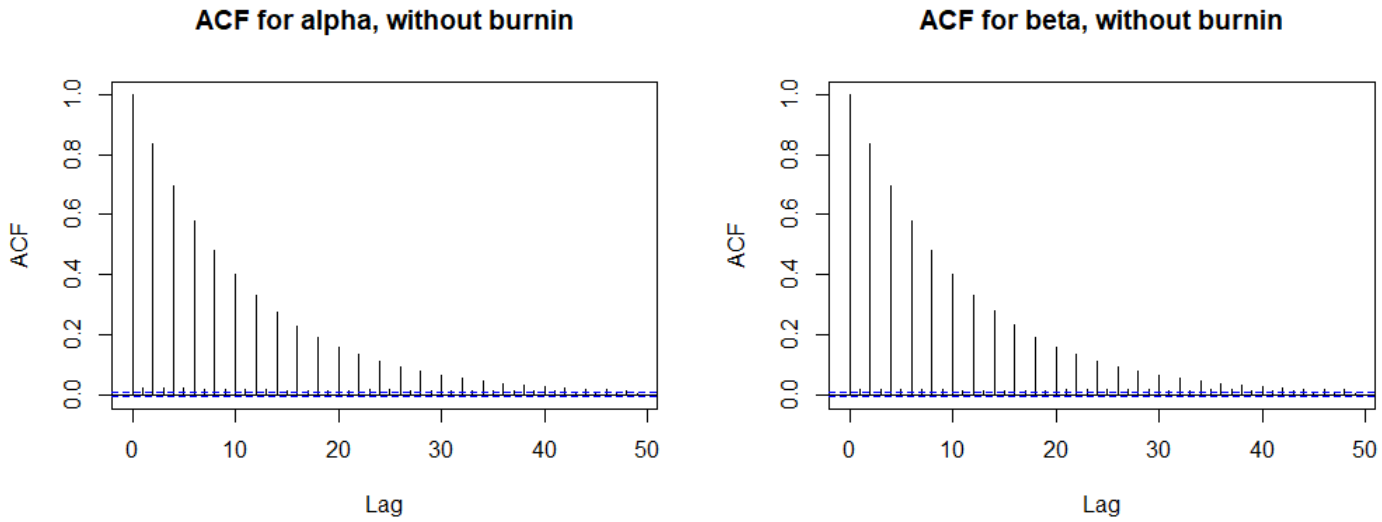

Figure 2: ACF plots for α and β, Importance Sampling

We see in Table 1 the estimates for α and β from importance sampling, as well as the Monte Carlo Standard Errors. We notice that the estimates are pretty close to the estimates found through the Laplace approximation. If I change the starting values, my estimates do not change a lot.

	α	β
MC mean	2.9572051	0.5177729
MC std error	0.0002927	0.0000556

Table 1: Importance Sampling Estimates and MC Standard Errors

Diagnostics on efficiency such as Effective Sample Size and Effective Sample Size/second are included later in section 1e as a full comparison.

A sufficient condition for the variability of my Monte Carlo approximation to be guaranteed to be finite is that the proposal has heavier tail than the target. I believe that this is true for importance sampling with our proposal distribution of exponential and our target distribution of the posterior described above.

(b) Construct an all-at-once Metropolis-Hastings (AMH) algorithm to approximate the posterior expectation of α, β. Provide pseudocode for your algorithm. This should include any distributions you had to derive. How did you determine the Monte Carlo approximations were good enough? That is, discuss stopping criteria (how you determined chain length), the number of starting values you tried, how you obtained initial values etc.

For this problem, we will construct an algorithm that will sample for α and β at the same time, through all-at-once Metropolis Hastings.

Pseudo-Code

1. Specify proposal distribution as two independent exponential random variables. This is so that the support of these distributions agree with the support of the posterior distribution, and α and β are both strictly positive. For both β and α in this problem, I also tried a Gamma distribution, but the results were not as promising.
2. Specify my target distribution as the product of the given distributions, or my posterior $log(f(\alpha, \beta|X))$, which was derived above. Return the log density.
3. Set initial values, which we will use as, once again, the values found through the Laplace Approximations.
4. Run Metropolis Hastings
   (a) Draw $\alpha*$ and $\beta*$ from the independent exponential proposals
   (b) Calculate the ratio of the target$(\alpha*, \beta*)$/target(α_i, β_i)
   (c) Accept with this probability (I do this by calculating a random uniform variable)
   (d) Store these values of α and β in a vector (either the previous α_i or β_i if the new one was rejected or use $\alpha*$ and $\beta*$ if accepted)
5. Calculate acceptance rate and assess diagnostics of post-burning samples

6. Adjust initial value if needed

Once again, I will assess some diagnostics and then report the MC estimates and standard errors for importance sampling. I determined if the Monte Carlo approximations were good enough through analysis of these plots, to see if assumptions were satisfied and if my samples were converging. I determined the starting values through the Laplace approximations. I also determined if my chain was running long enough through the trace plots.

First, I plot the trace plots in Figure 3 and assess if I need to remove some values as burnin. I plot the trace plots and I see there is pretty good mixing, but we should remove some values at the beginning as burnin. I remove 5000 values for both α and β samples as burnin. In Figure 4, we see the trace plots for both α and β after this burnin has been removed. We see that quite a bit of mixing is occurring through the trace plots, and it seems to satisfy independence. However, this assumption is not nearly as well satisfied as in the example of importance sampling.

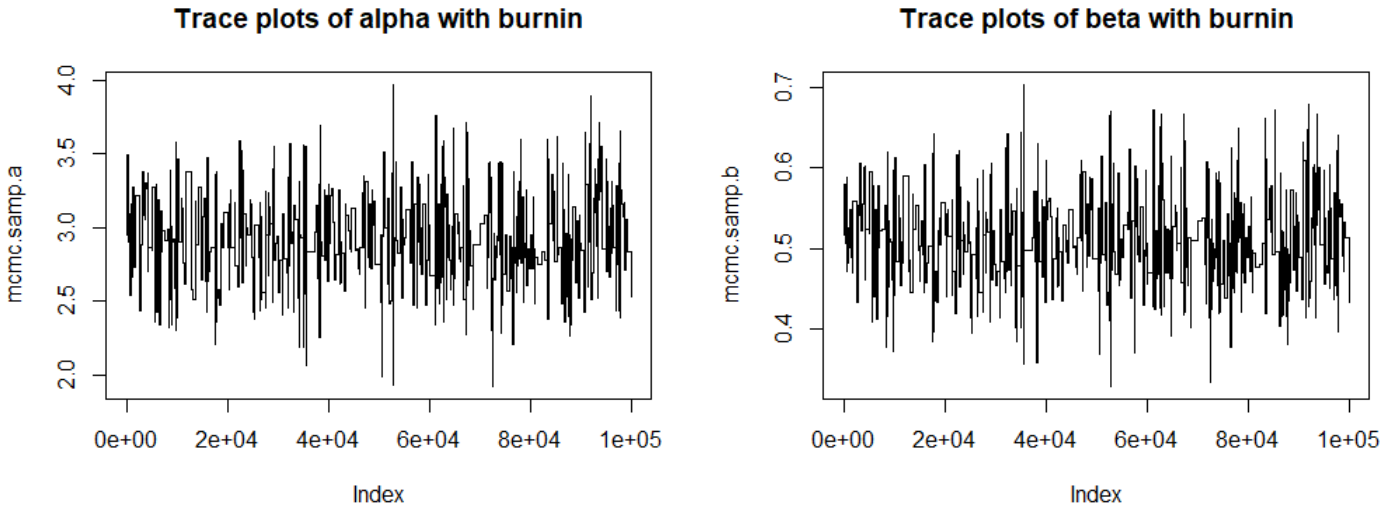

Figure 3: Trace plots for α and β, AMH with burnin

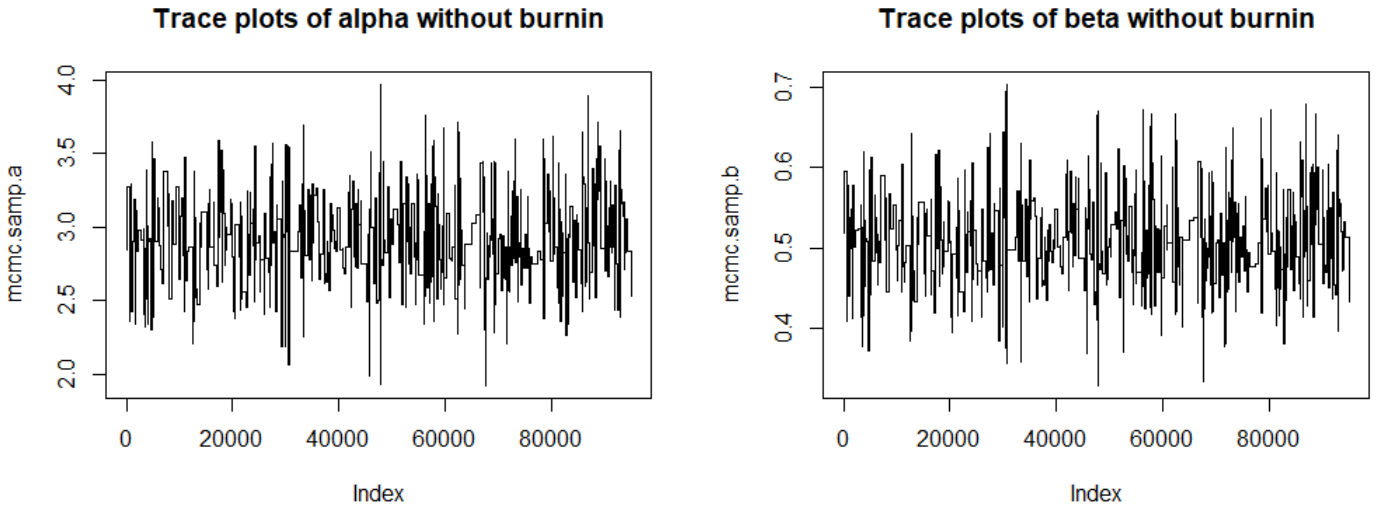

Figure 4: Trace plots for α and β, AMH without burnin

In Figure 5, we see that the ACF for both α and β decay, but not as quickly as in the case of importance sampling. There is some concern with the acf plots, as they do not decay very quickly, but the plots do not change much as we change the initial values or the distributions.

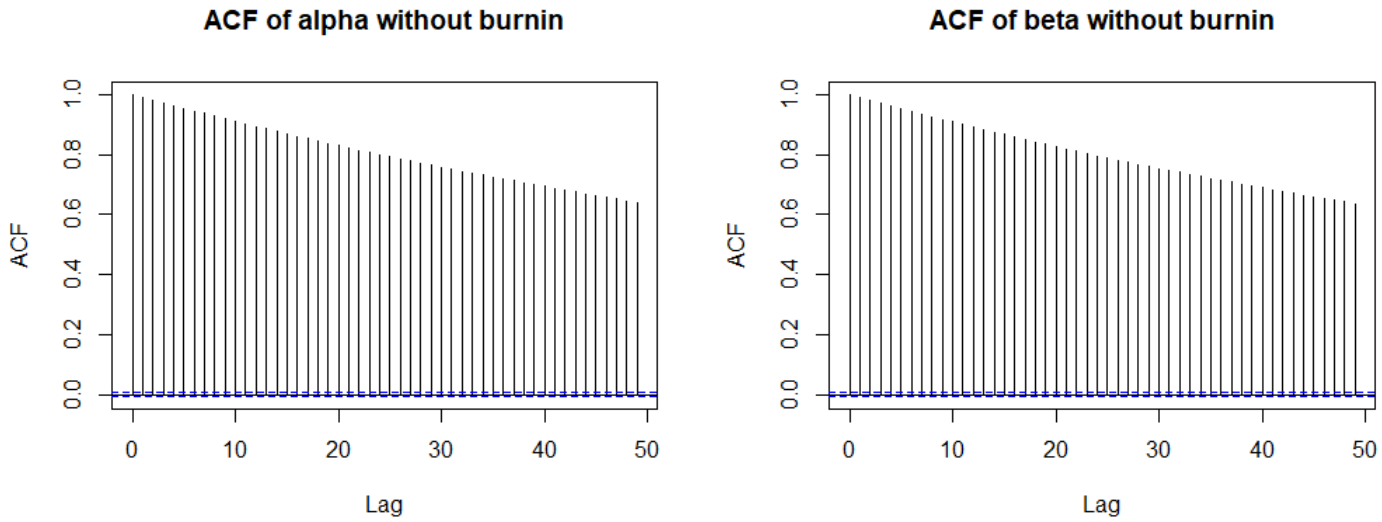

Figure 5: ACF plots for α and β, AMH

We see in Table 2 the estimates for α and β from all-at-once Metropolis Hastings, as well as the Monte Carlo Standard Errors. We notice that the estimates are pretty close to the estimates found through the Laplace approximation and as the importance sampling above. If I change the starting values, my estimates do not change a lot, as with importance sampling. However, the Monte Carlo Standard errors are higher than in importance sampling, as we will see later.

	α	β
MC mean	2.9150056	0.5088584
MC std error	0.0111121	0.0020651

Table 2: AMH Estimates and MC Standard Errors

Once again, diagnostics on efficiency such as Effective Sample Size and Effective Sample Size/second are included later in section 1e as a full comparison.

(c) Construct a variable-at-a-time Metropolis-Hastings (VMH) algorithm. You need to provide the same level of detail here as you did for the previous algorithm.

As shown in part a, we can calculate

$$f(\alpha|X,\beta) \propto f(X|\alpha,\beta)\pi(\alpha)$$

$$\propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\prod_{i=1}^{n} x_i^{\alpha-1} exp(-\beta\sum_{i=1}^{n} x_i) * exp(\frac{\alpha^2}{6})$$

$$\propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\prod_{i=1}^{n} x_i^{\alpha-1} exp(-\beta\sum_{i=1}^{n} x_i + \frac{\alpha^2}{6})$$

$$\text{So, } log(f(\alpha|X,\beta)) \propto log(\frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}) + log(\prod_{i=1}^{n} x_i^{\alpha-1}) - \beta\sum_{i=1}^{n} x_i + \frac{\alpha^2}{6}$$

and

$$f(\beta|X,\alpha) \propto f(X|\alpha,\beta)\pi(\beta)$$

$$\propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\prod_{i=1}^{n} x_i^{\alpha-1} exp(-\beta\sum_{i=1}^{n} x_i) * exp(\frac{\beta^2}{6})$$

$$\propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\prod_{i=1}^{n} x_i^{\alpha-1} exp(-\beta\sum_{i=1}^{n} x_i + \frac{\beta^2}{6})$$

$$\text{So, } log(f(\beta|X,\alpha)) \propto log(\frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}) + log(\prod_{i=1}^{n} x_i^{\alpha-1}) - \beta\sum_{i=1}^{n} x_i + \frac{\beta^2}{6}$$

Pseudo-Code

1. Specify proposal distribution for β as an exponential random variable, mean β. For both β and α in this problem, I also tried a Gamma distribution, but the results were not as promising.
2. Specify my target distribution as the prior times the likelihood (posterior): $f(\beta|\mathbf{Y},\alpha) \propto f(\mathbf{Y}|\alpha,\beta)f(\beta)$, and return the log density, shown above
3. Set initial values and variables
4. Run Metropolis Hastings
   (a) Draw $\beta*$ from the proposal
   (b) Calculate the ratio of the target($\beta*$)/target(β_i)
   (c) Accept with this probability (I do this by calculating a random uniform variable)
   (d) Store this value of β
5. Specify proposal distribution for α as an exponential random variable, mean α
6. Specify my target distribution as the prior times the likelihood (posterior): $f(\alpha|\mathbf{Y},\beta) \propto f(\mathbf{Y}|\alpha,\beta)f(\alpha)$, and return the log density, shown above
7. Set initial values and variables
8. Run Metropolis Hastings
   (a) Draw $\alpha*$ from the proposal
   (b) Calculate the ratio of the target($\alpha*$)/target(α_i)
   (c) Accept with this probability (I do this by calculating a random uniform variable)
   (d) Store this value of α
9. Calculate acceptance rate and assess diagnostics of post-burning samples
10. Adjust tuning parameter if needed

For this problem, I will once again assess some diagnostics and then report the MC estimates and standard errors for variable-at-a-time Metropolis-Hastings.

First, I plot the trace plots in Figure 6 and assess if I need to remove some values as burnin. I plot the trace plots and I see there is pretty good mixing, but we should remove some values at the beginning as burnin. I remove 2000 values for both α and β samples as burnin. In Figure 7, we see the trace plots for both α and β after this burnin has been removed. We see that quite a bit of mixing is occurring through the trace plots, and it seems to satisfy independence. However, this assumption is not nearly as well satisfied as in the example of importance sampling, as with all-at-once Metropolis Hastings.

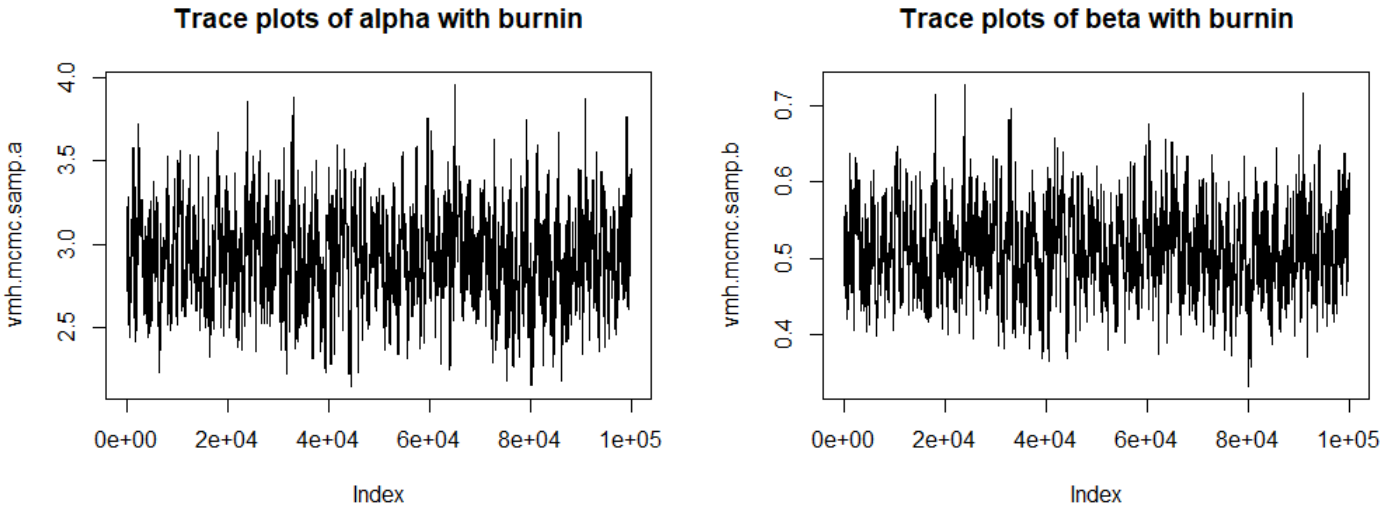

Figure 6: Trace plots for α and β, VMH with burnin

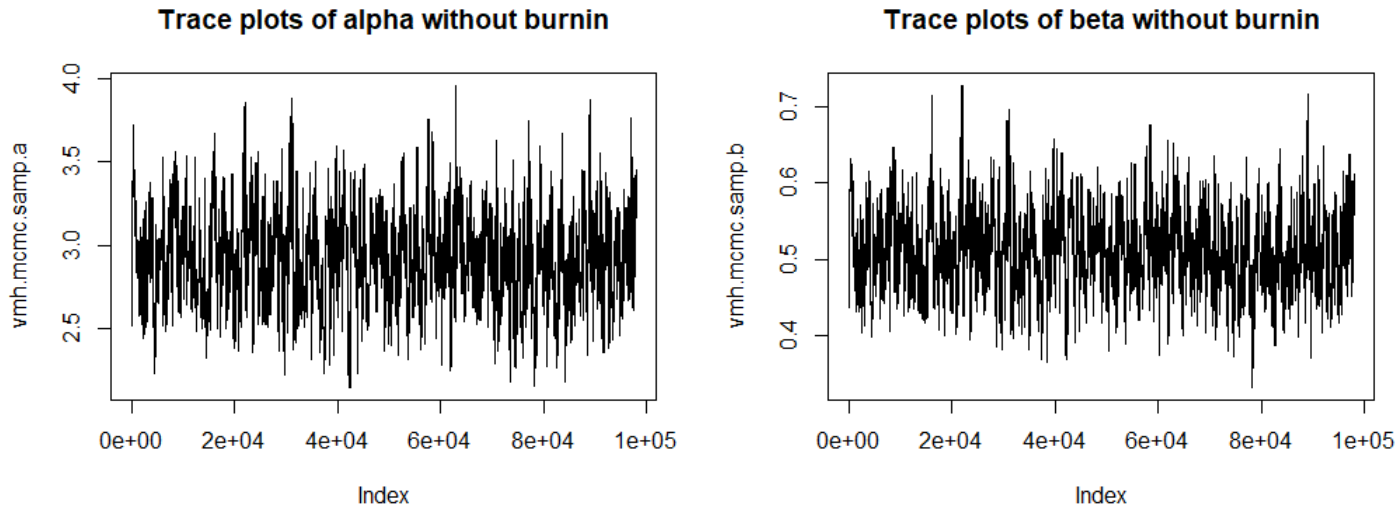

Figure 7: Trace plots for α and β, VMH without burnin

In Figure 8, we see that the ACF for both α and β decay, but not as quickly as in the case of importance sampling. There is some concern with the acf plots, as they do not decay very quickly, but the plots do not change much as we change the initial values or the distributions. This is very similar to the case of all-at-once MH.

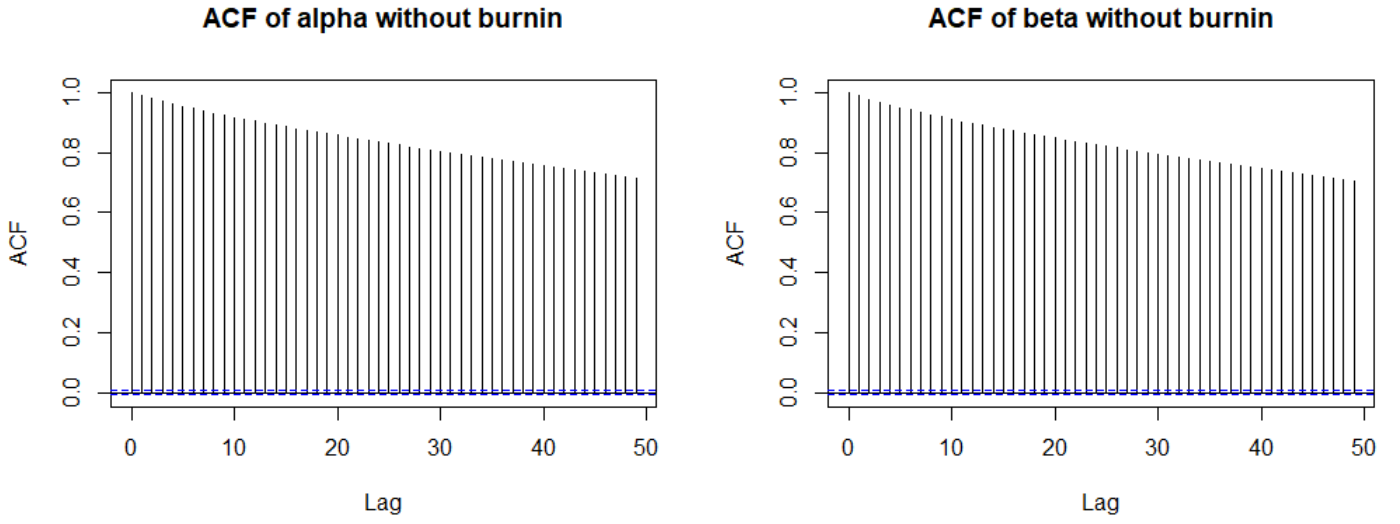

Figure 8: ACF plots for α and β, VMH

Finally, we see in Table 3 the estimates for α and β from importance sampling, as well as the Monte Carlo Standard Errors. We notice that the estimates are pretty close to the estimates found through the Laplace approximation. If I change the starting values, my estimates do not change a lot, as with AMH. The estimates are also pretty similar to AMH, as we will see in the next section.

	α	β
MC mean	2.9186986	0.5104147
MC std err	0.0115759	0.0022150

Table 3: VMH Estimates and MC Standard Errors

Once again, diagnostics on efficiency such as Effective Sample Size and Effective Sample Size/second are included later in section 1e as a full comparison.

(d) Provide a table with all approximations along with any error approximations, as well as the computational time taken by the algorithms.

Below, we see Table 4 with all of the approximations for α and β along with the Monte Carlo error approximations for our estimates, for each of the algorithms (importance sampling, VMH, and AMH).

	Laplace	Imp Samp	AMH	VMH
α est	2.9556237	2.9572051	2.9150056	2.9186986
β est	0.5174864	0.5177729	0.5088584	0.5104147
α MC error	-	0.0002927	0.0111121	0.0115759
β MC error	-	0.0000556	0.0020651	0.0022150
approx time	0.17	91.73	3.40	4.19

Table 4: Estimates and MC Error for each algorithm

We see that the estimates for α and β are pretty similar but they are not close to identical between the algorithms. There is only 1 decimal places in common across all estimates of α and β. We see that the importance sampling has the lowest monte carlo standard error, but the highest computation time. That is why it will be important to study efficiency, in the following section.

(e) Compare the efficiency of the four algorithms {importance sampling, AMH, VMH, and the Laplace approximation from the previous homework } using the methodology discussed in class, e.g. effective sample size, effective samples per second etc. The Monte Carlo algorithms are easier to compare than the comparison between them and the Laplace approximation; you may have to think carefully about this. Be sure to clearly explain why your approaches for comparing the algorithms are reasonable.

Below, in Table 9 we can see the estimates for the effective sample size as well as the effective sample size per second for our three Monte Carlo algorithms: importance sampling, VMH, and AMH. We see that importance sampling definitely has the highest effective sample size for both α and β. All-at-once MH follows, with variable-at-a-time MH having the smallest effective sample size for both α and β. However, when I simulated the importance sampling, it took considerably longer than the importance sampling. Therefore, when we analyze effective samples per second (ES/sec), we notice that AMH has the highest number of effective samples per second, followed by importance sampling and then once again by VMH.

	Imp Samp	AMH	VMH
ESS for α	8562.909	418.899	335.627
ESS for β	8611.336	419.341	339.223
seconds	91.73	3.4	4.19
ES/sec for α	93.349	123.206	80.102
ES/sec for β	93.877	123.336	80.960

Table 5: Efficiency Comparison across Algorithms

As mentioned in the problem statement, these are the easiest to compare. For the Laplace approximation, our result will be the same if we were to iterate this again, due to the use of the optim function, so a Monte Carlo framework is not helpful. I will note that it took a very small amount of time, 0.17 seconds, to run this optimization to find estimates of α and β for the Laplace approximation. I believe that the Laplace approximation is extremely efficient, as it was very fast to come up with our estimate, but perhaps it is not as useful in creating a final estimate of α and β. For the Monte Carlo techniques, it is very helpful to have an idea of error around our estimate, and this is not possible with the Laplace approximation. Therefore, we believe that it may be more appropriate to use Laplace approximation as a method to find starting values, as we did throughout the length of this homework. Then, we can proceed from there and compare our estimates to those found in the Laplace approximation.

It is reasonable to compare our algorithms in this manner because all of the Monte Carlo techniques can be compared through Effective Sample Size and ES/sec. These comparison tools show the tradeoff between effective sample size and time in producing those effective samples. It is also reasonable to compare the Laplace approximation by using it as a starting value and then comparing our results of the Monte Carlo estimates to these values. Then, we can come up with an idea to see if the Laplace approximation was close to our other values.

(f) Which algorithm would you recommend for this problem? How would you order the algorithms in terms of ease of implementation?

For this problem, I would recommend, as stated above, a dual approach. First, I would use Laplace approximation as it is very quick and does not need to be iterated repeatedly. This will give some good initial values for our Monte Carlo approaches. Out of the Monte Carlo approaches, I would suggest the Importance Sampling algorithm. Even though it has a smaller number of effective samples per second, it has a much higher number of effective samples, and the requirements of independence (seen through the trace and ACF plots) are much better satisfied in this algorithm. Therefore, with the caveat of "if we have the computing power/time to be able to do so" I would suggest importance sampling and if we do not, I would suggest all-at-once Metropolis Hastings, due to the smaller effective samples per second due to the fast computation time.

Problem 2

Lightbulbs: Assume that lightbulb lifetimes for a lightbulb made by a particular company are independent and exponentially distributed with expectation θ. Suppose in an experiment, m bulbs are switched on at the same time, but are only completely observed up to time τ. Let the lifetimes of these bulbs be $A_1, A_2, \ldots, A_m$. However, since the bulbs are only observed till time τ, not all these lifetimes will be observed. Now suppose that at time τ, the experimenter observes the number of light-bulbs, W, still working at time τ, and the lifetimes of all lightbulbs that stopped working by τ. For convenience, denote these bulbs that stopped working by time τ as $A_1, A_2, \ldots A_{m-W}$. Hence, the missing information consists of the lifetimes of the lightbulbs still working at time τ, $A_{m-W+1}, \ldots, A_m$. For a particular experiment, let τbe 184 days and m = 100. The observations for this experiment are as follows. The data on the lightbulb lifetimes for the bulbs that stopped working by τ are here: http://personal.psu.edu/muh10/540/hwdir/bulbs.dat (Assume that the remaining bulbs were still working at time τ.) Let the prior for θ be Gamma(1,100) (with parameterization so it has mean 100 and variance 100^2).

(a) Using auxiliary variables for the missing lightbulb data, construct an MCMC algorithm to approximate the posterior distribution of θ. Provide the same level of detail about your MCMC algorithm as you did in the previous problem.

In Data Augmentation, we need to optimize a difficult likelihood. So, we add unobserved (latent) variables so the Maximum Likelihood Estimation process is feasible. Namely, if $L(\theta|Y)$ is the likelihood of the observed vector of variables Y, we can introduce the latent variable Z, or more than one, so that the likelihood $L(\theta|Y, Z)$ becomes easy to optimize.

In our case, we know the form of $L(Y|\theta)$, where Y is the times of the 75 lights that are no longer lit up and have complete lifetimes by time τ. So, we can add auxiliary variables, Z, for the lightbulbs that did not die by time τ. We also know W, the number of lightbulbs that have survived up to time τ. To be specific, $A_1, A_2, \ldots A_{m-W}$ is Y, our known data, and $A_{m-W+1}, \ldots, A_m$ is Z, our missing or auxiliary data that we can add into the model. We know that all of the times, the A_i are all independent and they all follow an exponential distribution with parameter θ.

Instead of the posterior as $L(\theta|Y)$, we would like to simulate from $L(\theta|Y, Z, W)$. We know the Z times that are not observed have not died up to time τ, but they are still independent exponential random variables (because of the memoryless property). However, these are now shifted exponentials, instead of the typical exponential.

In general, the shifted exponential has CDF $1 - exp(\frac{1}{\lambda}(x - L))$ where $x \geq L$. So, the pdf is

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}(1 - exp(-\frac{1}{\lambda}(x - L))) = \frac{1}{\lambda}exp(-\frac{1}{\lambda}(x - L)).$$

In our case, $L = \tau = 184$, all of the unknown times obviously satisfy the requirement of being greater than or equal to τ, and our rate parameter is θ.

Due to the fact that we know that the times Y_i are less than τ, Y_i also no longer follows a normal exponential distribution. Instead, it is a truncated exponential distribution. The pdf of a truncated exponential distribution for Y is found here: http://lagrange.math.siu.edu/Olive/ch4.pdf. Specifically, for $0 < y \leq b$, where $\lambda > 0$, we have

$$f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda}exp(-\frac{y}{\lambda})}{1 - exp(-\frac{b}{\lambda})}$$

In our case, $b = \tau = 184$, and $\lambda = \theta$.

$$\text{So, } f_Y(y|\theta, b) = \frac{\frac{1}{\theta}exp(-\frac{y}{\theta})}{1 - exp(-\frac{\tau}{\theta})}$$

We also know W, the total number of lightbulbs that have survived up to time τ, is 25. We know that this is a binomial variable. For this binomial variable, n is 100 (the total number of bulbs), and p is the probability of a lightbulb surviving up to time τ. We know that the probability of surviving up until time τ is simply $p(\tau) = exp(-\frac{\tau}{\theta})$. For the binomial pmf, we have

$$f(W|\theta) \propto p^W(1 - p)^{n-W} \text{ where p is } exp(-\frac{\tau}{\theta}).$$

$$\propto exp(-\frac{\tau W}{\theta})(1 - exp(-\frac{\tau}{\theta}))^{n-W}$$

So, we have that $L(\theta, Z|Y) \propto L(Y, Z|\theta)\pi(\theta) \propto L(Y|\theta)L(Z|\theta)\pi(\theta)$ because Y and Z are independent. Therefore, the posterior has the form

$$f(\theta|Y, Z, W) \propto f(Y|\theta)f(Z|\theta)f(W|\theta)\pi(\theta)$$

$$\propto \left[\frac{(\frac{1}{\theta})^{75}exp(-\frac{1}{\theta}\sum_{i=1}^{75} y_i)}{(1-exp(-\frac{\tau}{\theta}))^{75}}\right] * \left[\theta^{-25}exp(-\frac{1}{\theta}(\sum_{i=1}^{25} z_i - 25*\tau))\right] * \left[exp(-25\frac{\tau}{\theta})(1-exp(-\frac{\tau}{\theta}))^{100-25}\right] * \left[exp(-\frac{\theta}{100})\right]$$

So, $log(f(\theta|Y, Z, W))$

$$\propto \left[-75*log(\theta) - \frac{1}{\theta}\sum_{i=1}^{75} y_i - 75*log(1-exp(-\frac{\tau}{\theta}))\right] + \left[-25*log(\theta) - \frac{1}{\theta}(\sum_{i=1}^{25} z_i - 25*\tau)\right] + \left[-25\frac{\tau}{\theta} + 75*log((1-exp(-\frac{\tau}{\theta}))\right] + \left[-\frac{\theta}{100}\right]$$

$$= \left[-100*log(\theta) - \frac{1}{\theta}\sum_{i=1}^{75} y_i - \frac{1}{\theta}\sum_{i=1}^{25} z_i - \frac{\theta}{100}\right]$$

We also know the conditional distribution of $f(Z|\theta, Y, W)$ is just an shifted exponential, as shown above. So, $log(f(Z|\theta, Y, W)) = log(f(Z|\theta)) = -25*log(\theta) - \frac{1}{\theta}(\sum_{i=1}^{25} z_i - 25*\tau) = -25*log(\theta) - \frac{1}{\theta}(\sum_{i=1}^{25} z_i - 25*184)$

Pseudo-code

1. We know the full conditional for Z, so we just draw Z from a shifted exponential distribution, with parameter θ (initially using initial value, and then previous value of θ).
2. We will use Metropolis Hastings to provide updates of our parameter of interest, θ, given its distribution displayed above.
3. Specify proposal distribution for θ as an exponential, mean θ, so that the support of the proposal matches the support of θ, strictly positive.
4. Specify my target distribution for θ as the distribution derived above, $log(f(\theta|Y, Z, W))$, and return the log density
5. Set initial values and variables
6. Run Metropolis Hastings for θ
   (a) Draw $\theta*$ from the proposal
   (b) Calculate the ratio of the target($\theta*$)/target(θ_i), using previous values of Z_i
   (c) Accept with this probability (I do this by calculating a random uniform variable)
   (d) Store this value of θ
7. Repeat many times and assess diagnostics.

For this problem, I will once again assess some diagnostics and then report the MC estimates and standard errors for the Metropolis-Hastings.

First, I plot the trace plots in Figure 9 on the left and assess if I need to remove some values as burnin. I plot the trace plots and I see there is some need to remove burnin values. I remove 500 values for the samples as burnin. In Figure 9 on the right, we see the trace plot for θ after this burnin has been removed. We see that quite a bit of mixing is occurring through the trace plot, and it seems to satisfy independence.

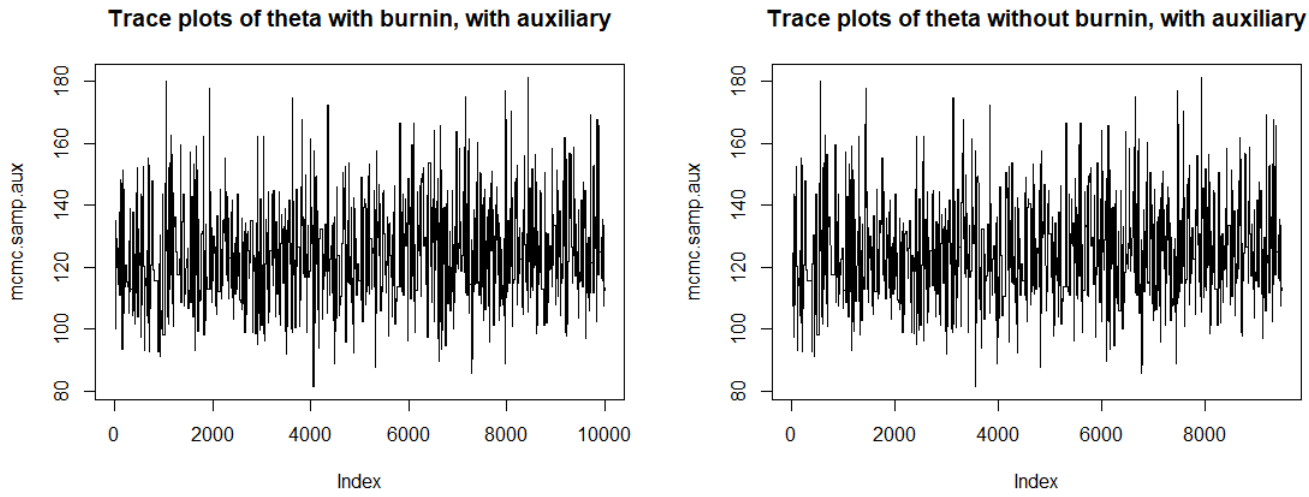

Figure 9: Trace plot for θ with and without burnin

In Figure 10, we see that the ACF for θ decays quickly. We therefore believe there is no concern presented by the ACF plot in terms of independence.

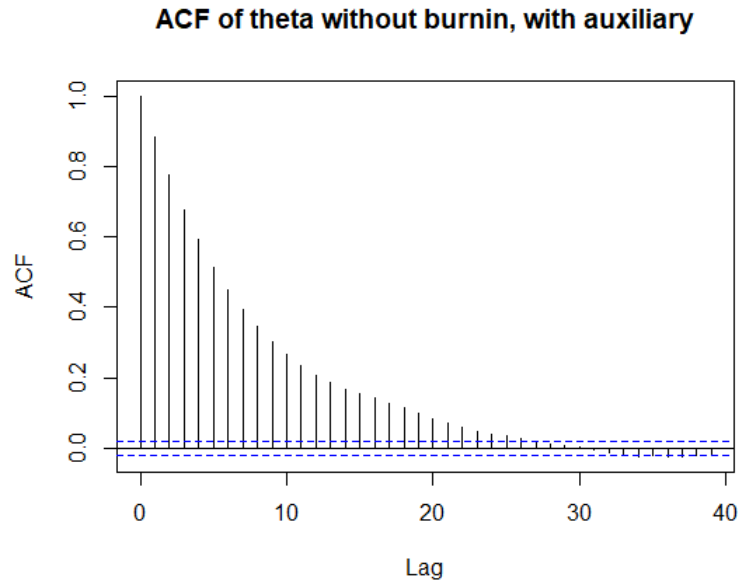

Figure 10: ACF plots for θ

Finally, we see in Table 6 the estimate for θ as approximately , with a standard error of124.2 approximately 0.566. We also present diagnostics on efficiency such as Effective Sample Size and Effective Sample Size/second which we will discuss and compare later in section 2e as a full comparison.

	value, with aux
theta estimate	124.2016391
MC std error	0.5664241
time	1.91
ESS	599.8483751
ES/sec	15.91

Table 6: Description of MCMC algorithm with aux data

(b) Construct a different MCMC algorithm, this time without using auxiliary variables/data augmentation. Again, provide details.

This time, I will sample from the posterior distribution directly to estimate the posterior distribution of θ. We know that the posterior distribution is as follows: $f(\theta|x) \propto f(x|\theta)\pi(\theta)$.

For this problem, we know that the likelihood $f(x|\theta)$ is Exponential(θ) and the posterior is Gamma(1,100). Therefore, the posterior has the form

$$f(\theta|x) \propto f(x|\theta)\pi(\theta).$$
$$\propto \frac{1}{\theta}^n exp(-\frac{1}{\theta}(\sum_{i=1}^{n} x_i) * exp(-\frac{\theta}{100})$$
$$\propto \frac{1}{\theta}^n exp(-\frac{1}{\theta}(\sum_{i=1}^{n} x_i) - \frac{\theta}{100})$$
$$\text{So, } log(f(\theta|x)) \propto -nlog(\theta) - \frac{1}{\theta}(\sum_{i=1}^{n} x_i) - \frac{\theta}{100}, \text{ where n is 75.}$$

Therefore, my MCMC algorithm will simulate from this distribution.

Pseudo-code

1. We would like to construct an MCMC algorithm to approximate the posterior distribution of θ. To do this, we will use Metropolis Hastings
2. First, specify proposal distribution for θ as a exponential, mean θ, so that the support of the proposal matches the support of θ, strictly positive
3. Specify my target distribution as the prior times the likelihood (posterior): $f(\theta|x) \propto f(x|\theta)f(\theta)$, and return the log density, shown above
4. Set initial values and variables
5. Run Metropolis Hastings
   (a) Draw $\theta*$ from the proposal
   (b) Calculate the ratio of the target($\theta*$)/target(θ_i)
   (c) Accept with this probability (I do this by calculating a random uniform variable)
   (d) Store this value of θ
6. Run many times and perform diagnostic measurements to see if initial value needs to be adjusted.

For this problem, I will once again assess some diagnostics and then report the MC estimates and standard errors for the Metropolis-Hastings.

First, I plot the trace plots in Figure 11 on the left and assess if I need to remove some values as burnin. I plot the trace plots and I see there is definite need to remove some burnin values, because my starting value of 100 was not a great fit. I remove 500 values again for the samples as burnin. In Figure 11 on the right, we see the trace plot for θ after this burnin has been removed. We see that quite a bit of mixing is occurring through the trace plot, and it seems to satisfy independence.

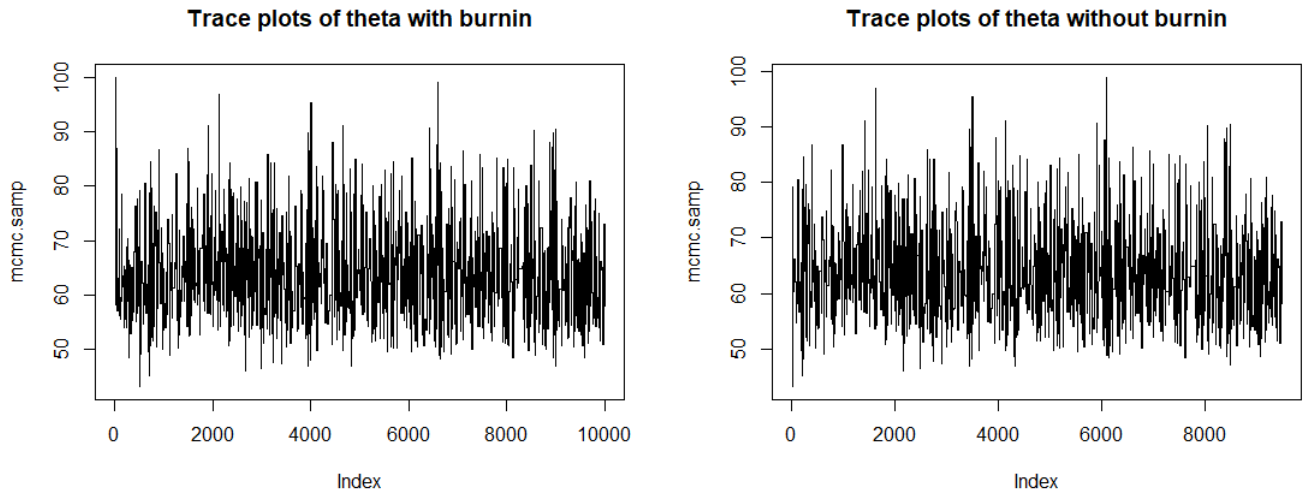

Figure 11: Trace plot for θ with and without burnin

In Figure 12, we see that the ACF for θ decays quickly. We therefore believe there is no concern presented by the ACF plot in terms of independence.

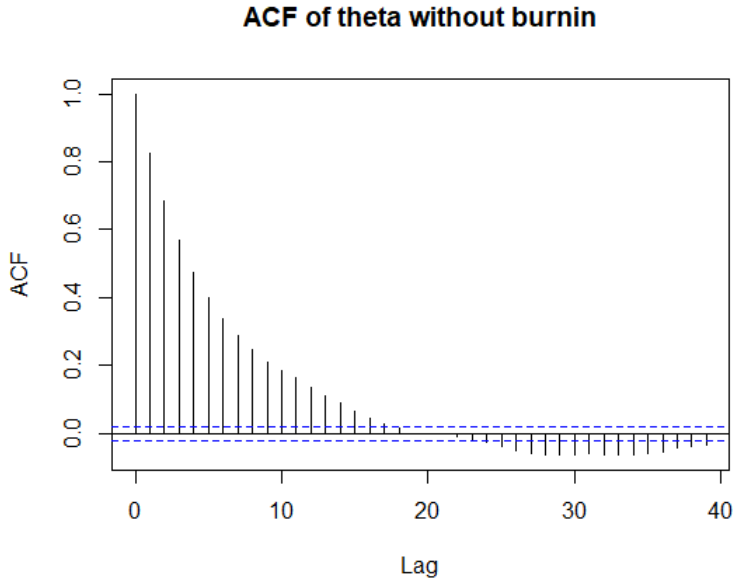

Figure 12: ACF plots for θ

Finally, we see in Table 7 the estimate for θ as approximately 63.97, with a standard error of approximately 0.244. We also present diagnostics on efficiency such as Effective Sample Size and Effective Sample Size/second which we will discuss and compare later in section 2e as a full comparison.

	value, without aux
theta estimate	63.9709443
MC std error	0.2443904
time	0.13
ESS	883.7914867
ES/sec	6798.3960517

Table 7: Description of MCMC algorithm with no aux data

(c) Overlay posterior density plot approximations for the two algorithms. Provide a table that shows the posterior mean approximations for θ along with MCMC standard errors.

In Figure 13, we overlay the posterior density approximations for the two algorithms. We notice that when we incorporate auxiliary information, specifically information on Z and W, we get a considerably higher estimate for θ. In fact, the whole distribution is shifted higher. However, we also notice that the variance of this distribution is larger than the previous distribution.

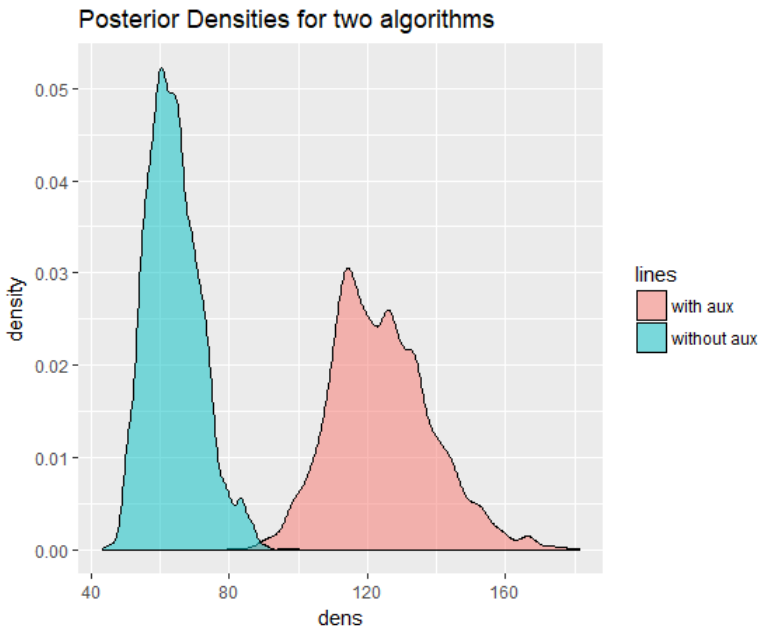

Figure 13: Overlay of Posterior Density Approximations

Below, in Table 8, we notice that, as we expect from Figure 13, that the estimate of θ is higher when we include auxiliary information than if we do not include auxiliary information. Also as we may expect from observing the density plots, the standard error is also higher when we incorporate auxiliary information.

	with aux	without aux
theta estimate	124.2016391	63.9709443
MC std error	0.5664241	0.2443904

Table 8: Posterior mean approx with std errors

Intuitively, this set of result makes sense. When I incorporated information on lightbulbs that had not yet gone out at time τ, we are going to have a higher estimate of θ. We know that θ is the expected lifetime, so of course we will have a higher estimate for θ when we incorporate data on bulbs that have not gone out yet, than if we say we have a complete dataset with the ones that have gone out up to time τ.

(d) For the auxiliary variable method, plot the approximate posterior pdf for one of the "missing" lightbulbs, then overlay the approximate posterior pdfs for a lightbulb made by the company. You should notice that they are different. Report the posterior mean estimates for each of them.

Below, in Figure 14, I plot the approximate posterior pdf for one of the missing lightbulbs, seen in red, vs one of the normal ones, seen in blue. In essence, this is just an exponential random variable vs a shifted exponential random variable. We know that the missing lightbulbs have not gone out yet at time $\tau = 184$, so the support of this density is going to be above time 184. For a normal lightbulb made by the company, the lightbulb could die at any point, so the bulk of this density lies lower than the missing lightbulb.

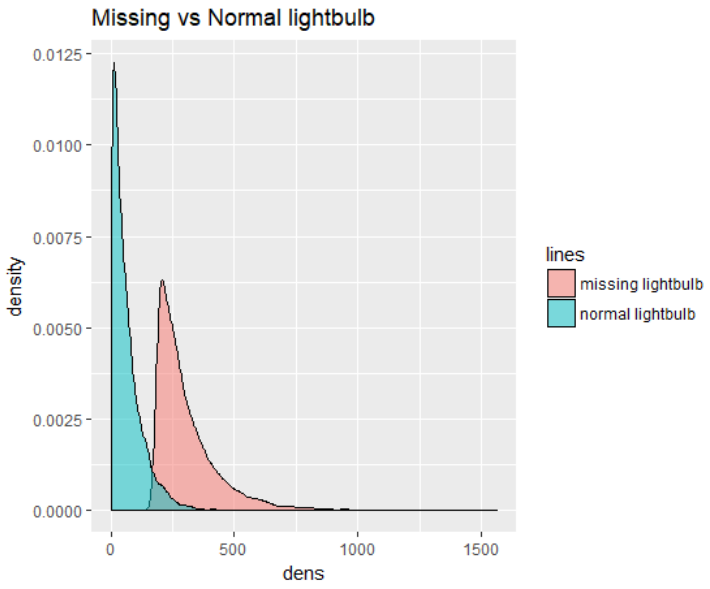

Figure 14: Densities for missing lightbulb vs lightbulb made by company

The posterior mean estimate for the posterior pdf of the "missing" lightbulbs is approximately 305 days. The posterior mean estimate for a lightbulb made by the company is approximately 63 days. This agrees with our reasoning from earlier, that the expected lifetime is going to be longer for the bulbs that we know have not gone out at time τ.

(e) Looking at just the ability to approximate the posterior per iteration of the algorithm, which of the two MCMC algorithms is more efficient? Now accounting for computing costs, which of the two MCMC algorithms is more efficient? Which algorithm would you recommend?

In Table 9, we see that the algorithm without the auxiliary information has higher effective sample size than the one that incorporates auxiliary information, though they are still in the same order of magnitude.

However, when we incorporate computing costs and instead look at the time taken to run these algorithms and look at effective samples per second, we see that the algorithm without auxiliary information is much more efficient.

	with aux	without aux
time (sec)	15.91	0.13
ESS	599.8483751	883.7914867
ES/sec	37.7026006	6798.3960517

Table 9: Efficiency comparison

Based on these results, I conclude that the algorithm without auxiliary information is the more efficient algorithm. However, I still do not believe that I can recommend this algorithm based on these results. If it were based solely on computing cost, then I would perhaps choose the algorithm without auxiliary information. However, I think that this exercise is useful in pointing out that we get drastically different results when we incorporate auxiliary information in terms of the estimate of θ. Therefore, I would say, if the computing power is available, then the algorithm that incorporates auxiliary information may be worth the extra time because of the added information and perhaps improved estimate of θ.

(f) This course is focused on computing but it is worth noting some basics about inference. Compare your results above with what would happen to inference if you ignored the missing data by overlaying the density plots.

This part is not too particular about what we are supposed to overlay. However, I will assume that it wants us to overlay the posterior estimates of θ using the two algorithms, which was presented earlier in Figure 13. As mentioned above, this plot gives quite meaningful insight as to the potential problems with inference about θ if we use the missing data vs if we do not use the missing data. If we do use the missing data, we would conclude that θ is much higher than if we were to ignore the missing data in our inference. We would get drastically different results for hypothesis testing. Therefore, this is one reason why I would think that the extra computing cost is worth adding this additional information into our estimation of θ.