# INFERRING MARKOV CHAIN TRANSITION PROBABILITIES USING METROPOLIS HASTINGS AND IMPORTANCE SAMPLING

Meet Shah – i6196781

OTHER TEAM MEMBER: REBECCA WALTER – I6186657

INTRODUCTION TO SOFTWARE IN ECONOMETRICS, OPERATIONS RESEARCH AND ACTURIAL SCIENCE

EBS2043

## Introduction

Markov chains are very relevant to lives. For instance, think of the stock market. The price of a stock at any instant is dependent upon its price on the previous instant. Even the direction of change in price can be attributed to the direction of change in price in a previous period (momentum theory).

Yet, it can be difficult to estimate the transition probability parameters for a Markov Chain which is observed. One method of estimation is Bayesian inference: updating prior beliefs given a chain of state transitions. Bayesian inference can be done in several ways, for example by using Markov chain Monte Carlo methods.

This paper develops and studies the effectiveness of the Metropolis-Hastings algorithm and the Importance Sampling Algorithm in estimating the transition probabilities of a first order, two state Markov Chain. It also compares the relative performance of the two algorithms.

The first section is dedicated to explaining the simulated Markov chain. The second section focusses upon the Metropolis Hastings algorithm. The third section elaborates on the Importance Sampling algorithm. Then, the two algorithms are compared. The paper is then concluded.

## 1. Data and Purpose

*1.1 Markov Chain*

A Markov chain is a chain of states where the probability of transitioning into a given state is conditional on the states preceding it. In this simulation, a first order, two state Markov chain is simulated with the following transition probability Matrix:

$$\begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.2 & 0.8 \\ 0.45 & 0.55 \end{bmatrix} \end{array}$$

$$Let\ p_{ij} = P(State = j|Previous\ State = i)$$

Therefore, the probability of remaining in state 1 ($p_{11}$) is 0.2, and the probability of going from state 2 to state 1 ($p_{21}$) is 0.45. 1000 observations are simulated.

This leads to the following transitions:

*1-1:* 72; *1-2:* 287; *2-1:*286; *2-2:* 354.

The initial state is 1, which has been randomly generated using the steady state probabilities given the set transition probabilities.

*1.2 Likelihood Function*

For a first order two state Markov chain:

$$Let\ n_{ij}\ be\ the\ number\ of\ transitions\ from\ state\ i\ to\ j$$

$$Let\ p_{ij}\ be\ the\ probability\ of\ transitioning\ from\ state\ i\ to\ state\ j$$

Then, the likelihood function is as follows:

$$Steady\ State\ Probability_{First\ State} * \prod_{\{i,j\}\in S} p_{ij}^{n_{ij}}$$

$$Where\ S = \{\{1,1\},\{1,2\},\{2,1\},\{2,2\}\}$$

### *1.3 Problem and Prior Belief*

Given a Markov chain x, what are the transition probabilities?

Under frequentist statistics, one could get an estimator, and analyse it using confidence intervals and hypothesis tests. In Bayesian statistics, one could think of the transition probabilities as random variables on which one has a belief, and then try to update that belief based on the data collected.

In this simulation, the prior probabilities are assumed to **be independently uniformly distributed** on an interval of zero to 1. This implies that there is no preference for one probability over another before the data is obtained. The intuition behind this is that without observing the data, one can only have a wide belief about the transition probabilities of a Markov chain. This is a completely uninformative prior, which does not exclude any probability from being part of the distribution.

From a mathematical point of view, this makes the posterior distribution directly proportional to the likelihood function defined in 1.2, as the distribution of the uniform variables (assuming valid probabilities) is 1 on this interval. Hence, the likelihood function defined above shall act as the target posterior distribution for the algorithms.

## 2. Metropolis-Hastings Algorithm

The Metropolis-Hastings is a Markov Chain Monte Carlo algorithm, which uses a candidate distribution to sample from the posterior, and then calculates an acceptance probability. The acceptance probability, in this case, is based on the likelihood function defined in part 1.2, and the candidate distribution used. More elaborately, it is the ratio of the likelihood function to the candidate density for a particular draw, divided by the same ratio of the previous draw. If a draw is accepted, it is stored, otherwise the previous draw is repeated.

### *2.1 Candidate Distribution*

As a candidate distribution for the Metropolis-Hastings, a truncated normal distribution is used, on an interval from 0 to 1, with a standard deviation of 0.05, and a mean which is equal to the previously accepted value. This candidate distribution is used independently for both p11 and p21. This distribution has several favourable properties:

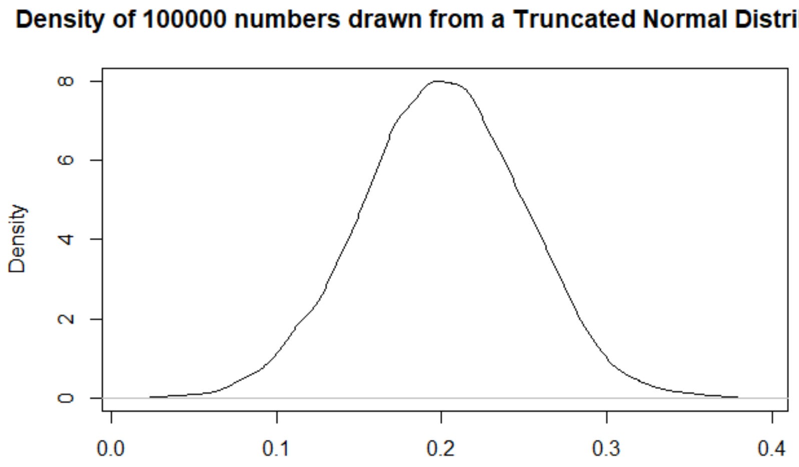

*Figure 1 Density of a Truncated Normal Distribution*

- It allows for draws only between 0 and 1, which is the support of the prior.
- When a draw is accepted, it places more emphasis on probability values near those draws and allows for dependencies in the draws.
- It allows for a convergence after a burn in as it will move closer and closer to a more likely probability over time, regardless of the initialization.
- Shifting the variance of the distribution allows for changing of the acceptance rate.

It also has some disadvantages:

- There will be autocorrelation between successive draws.
- A poor initialization may require a long burn in period to converge.

*2.2 Results:*

An acceptance ratio of **0.2657** is obtained, which is close to the optimal value of 0.234 as suggested by Gelman, Gilks and Roberts (1997). The algorithm is run for 10,000 iterations, and has the following process:

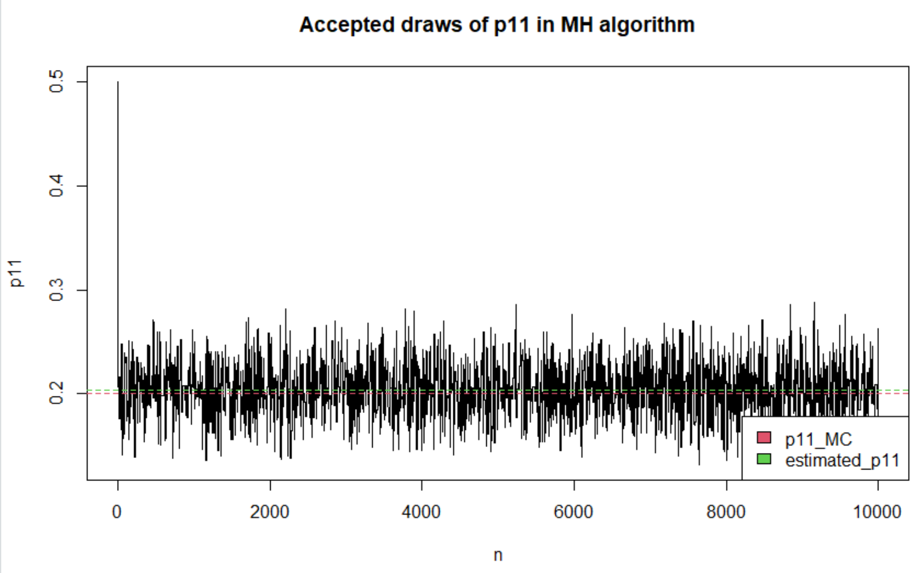

*Figure 2 10,000 P11 draws in Metropolis-Hastings*

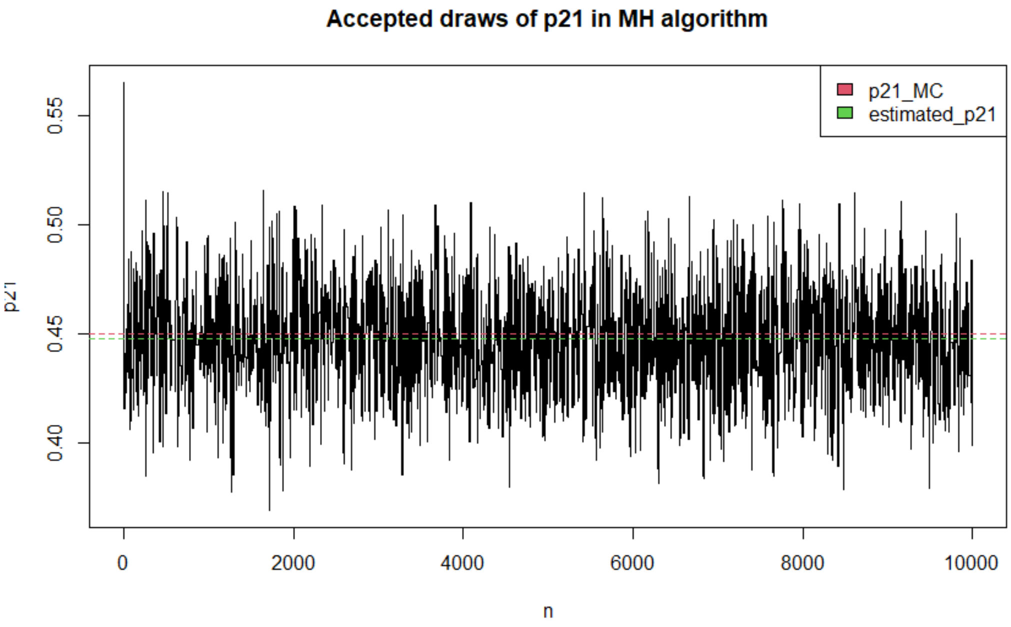

*Figure 3 10,000 P21 draws in Metropolis-Hastings*

Note the significant period before convergence. This is because of the initialization, which is at 0.5, 0.5 (expected value of the prior). This part of the data is removed from the following analysis (Burn in). This is because the algorithm requires a certain amount of draws before it starts converging near the true values. The number of draws needed is dependent on the initialization in this case. Different candidate distributions and different initializations can require shorter or longer initialization periods.

The data consists of draws 1,000 to 10,000. The required burn-in period in this case is closer to around 200 observations, however, to be on the safe side 10% of the initial data is removed. The following statistics are based on the data without the initial draws.

| Probability | P11 | P21 |
|---|---|---|
| Mean | 0.2036 | 0.4475 |
| Standard Deviation | 0.0241 | 0.0224 |
| Error of Mean from True Value | +0.0036 | -0.0025 |

Thus, the Metropolis-Hastings algorithm gives a prediction which is very close to the true value of the parameter. The root mean squared error is 0.0031. The resulting posterior distributions are shown in figures 4 and 5.

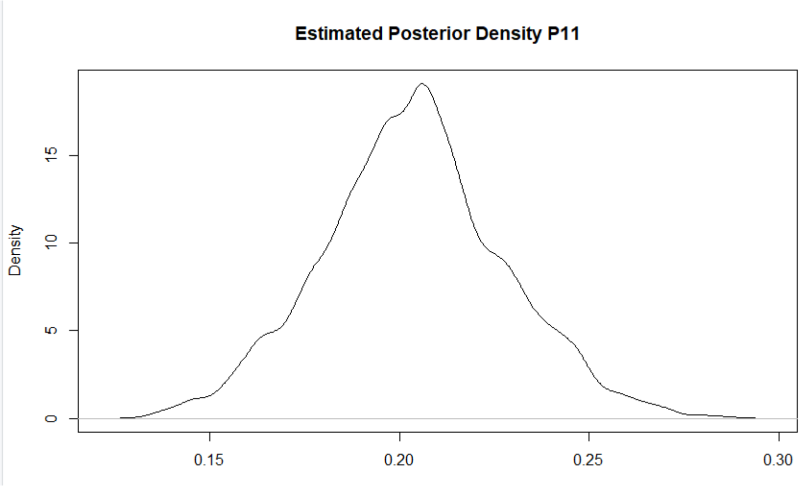

*Figure 4 Metropolis-Hastings Posterior Density P11*

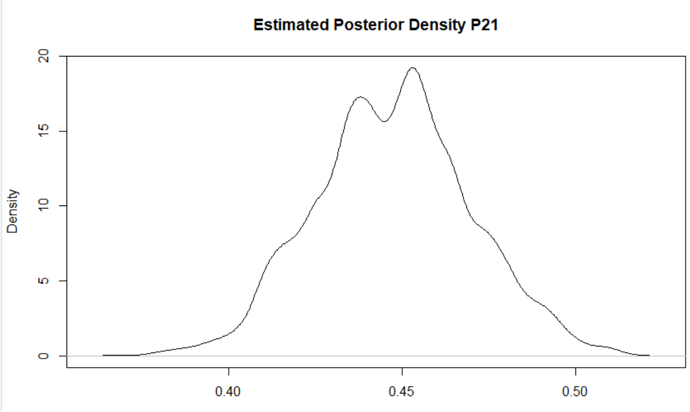

*Figure 5 Metropolis-Hastings Posterior Density P21*

For comparison, the density of a normal distribution with a mean equal to the mean of the p11 draws, and standard deviation equal to the standard deviation of p11 draws is shown below.

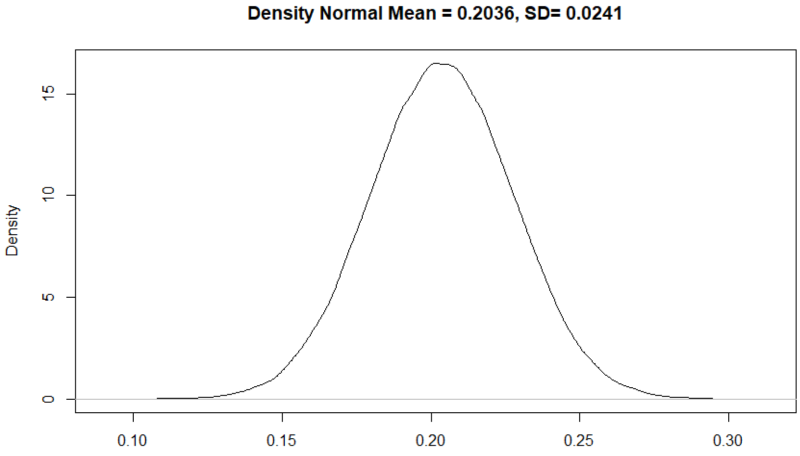

*Figure 6 Normal Distribution Density*

Hence, the beliefs are updated from all probabilities being equally likely to a much smaller subset with a non-flat density. This posterior distribution implies that probabilities around 0.2(0.45) are more likely than other probabilities in the state space. Indeed, by conditioning on the data, there is a much more informative belief about the parameters.

### *2.3 Further Discussion on Choice of Standard Deviation*

Currently, a standard deviation of 0.05 is being used, which lends itself to the 0.2657 acceptance rate. Using a larger standard deviation would decrease the acceptance rate, while using a lower one would increase autocorrelation tremendously.

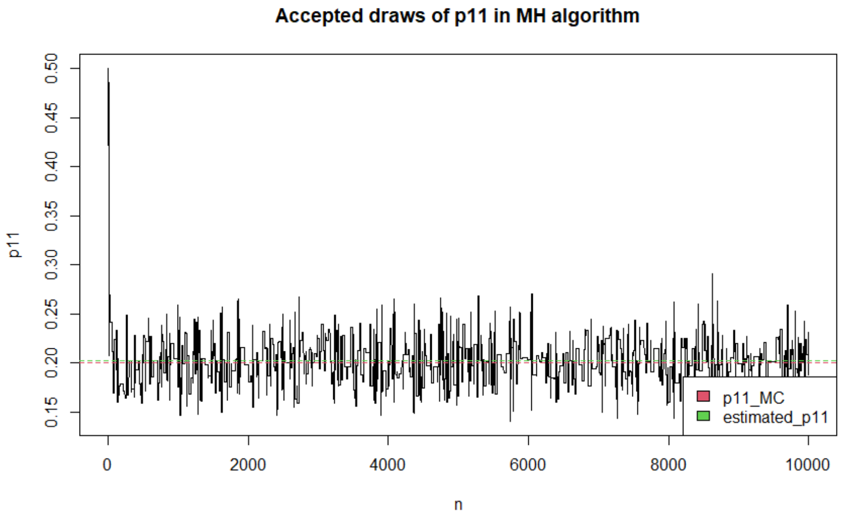

*Figure 7 MH Process with Truncated Normal sd=0.1*

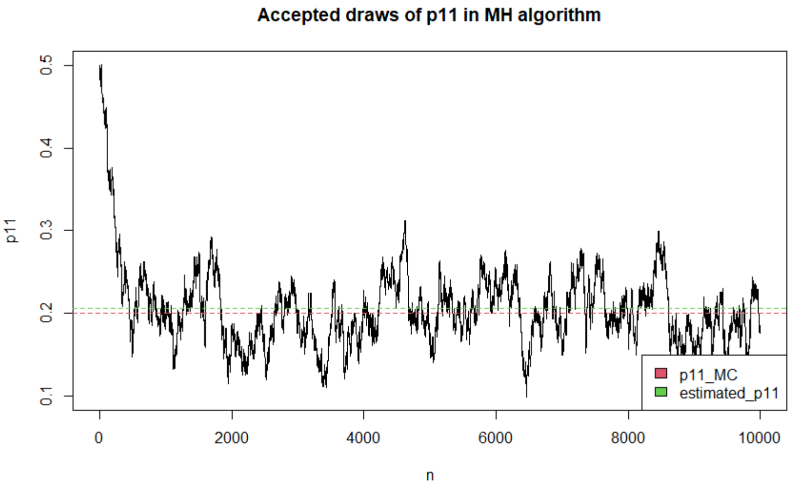

*Figure 8 MH Process with Truncated Normal sd=0.005*

Figure 7 refers to a candidate distribution with a standard deviation of 0.1. The acceptance rate is extremely low (0.08). Because several draws are repeated, there are several straight lines in the process. The low acceptance rate means that only a few observations dictate the mean, and thus is not appropriate as an estimator for the posterior distribution. There is poor mixing in this case.

Figure 8 shows the candidate truncated normal with a standard deviation of 0.005. In this case, the acceptance rate is very high (0.7562), and the sample does not appear to be independently distributed anymore. In fact, there is significant autocorrelation in the sample, which is undesirable.

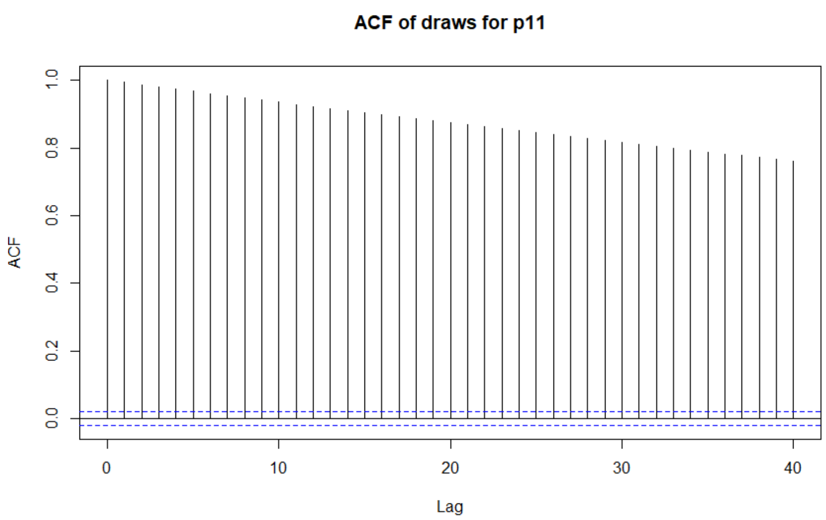

*Figure 9 Autocorrelation function of MH draws with low standard deviation in candidate*

Thus, it can be inferred that a standard deviation which provides an acceptance rate close to 25% is ideal in this scenario. Using 0.05 allows for this acceptance rate, and hence is used.

## *2.4 A note on thinning the data*

The autocorrelation function for the P11 draws is displayed below.

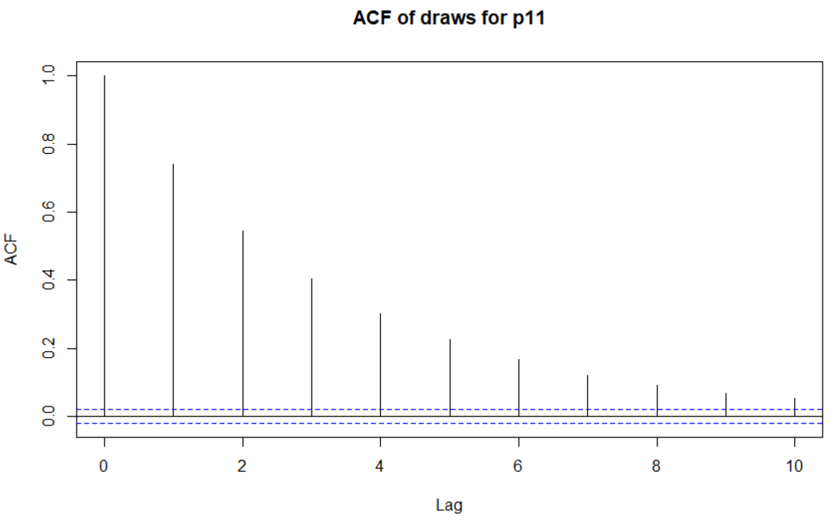

*Figure 10 Autocorrelation function for Original MH Process*

There is some autocorrelation, and hence thinning may be beneficial in increasing the robustness of the estimated mean. Taking every 6th draw from the same sample and increasing the number of iterations to 61,000, the following autocorrelation function is obtained:

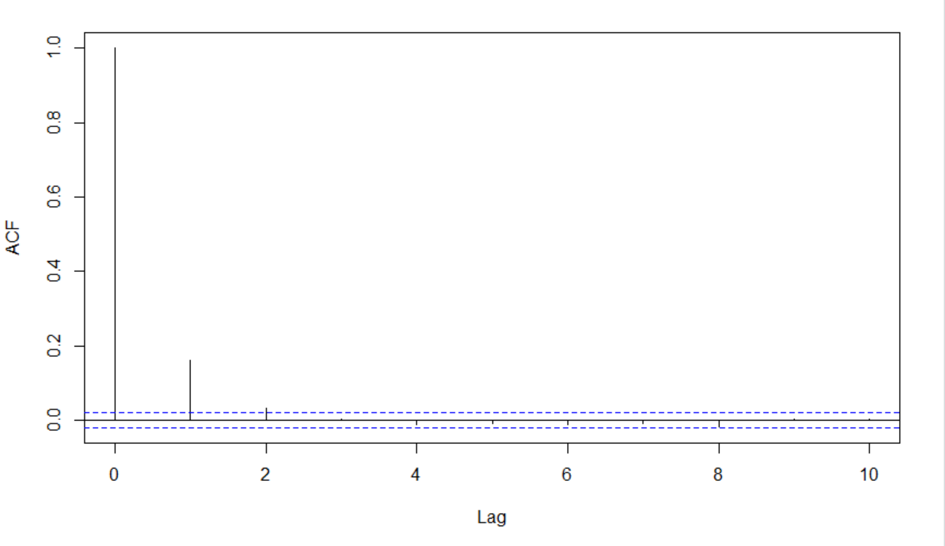

*Figure 11 Autocorrelation Function for MH process with Thinned Draws*

The following series is obtained:

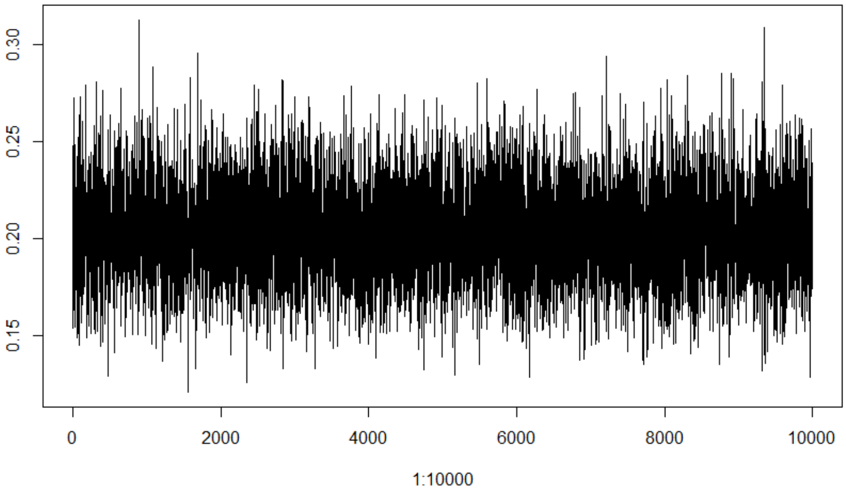

*Figure 12 Thinned draws in MH process*

These set of draws have mean 0.2035 and standard deviation 0.0244. It is not very different from the results with autocorrelated draws. Hence, the results reported without thinning are like results obtained with thinning, and hence ignoring it in the analysis does not yield spurious results and results in computational efficiency for obtaining a sample of a given size. However, it does result in smoother posterior densities.

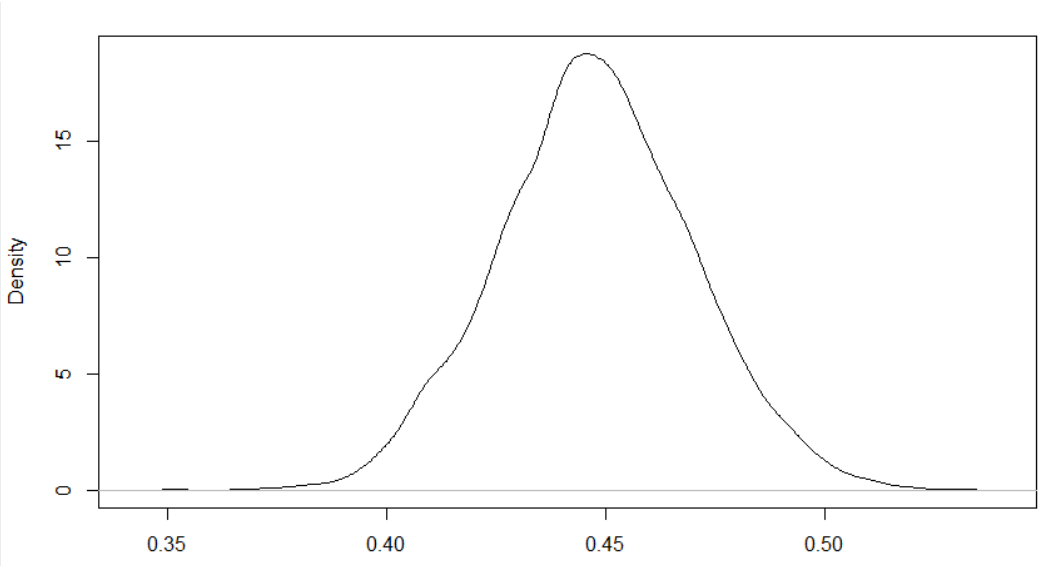

*Figure 13 Density of P21 with Thinned Draws*

### *2.5 A note on other candidate distributions*

The only real constraint in choice of candidate distribution is that they should yield numbers on an interval from 0 to 1. Two such distributions are the uniform distribution on an interval of 0 to 1 and the beta distribution.

The problem with the uniform distribution is that it shall have a horrible acceptance rate in this scenario. The likelihood that both values drawn are close to the true value is very slim. One way to avoid this is to use an estimate for the true probability and use the uniform distribution on an interval around it (say P11/P21 Estimate ± 0.1). This would result in successive draws being independent and identical, and acceptance rate can be changed by changing the wideness of the interval. However, it is reliant on the probability estimate. The truncated normal distribution is more likely to explore the whole parameter space and make use of the learning process embedded in Metropolis-Hastings.

There are two ways to use the beta distribution. First, one can use independent successive draws from the beta distribution, perhaps centred on a transition probability estimate. Then, draws would be independently and identically distributed. Acceptance rates can be adjusted by adjusting the shape parameters. This would be one possible independent candidate in the Metropolis-Hastings algorithm. However, it is again very dependent on the estimate for the probability taken, and adjusting the variance is not as precise as that in the truncated normal distribution.

The second way to use the beta distribution is to use accepted probabilities to change the candidate distribution. One way of doing this is to adjust the shape parameters so that the mean of the beta distribution from which the new draw is being made is the previous draw. However, this would typically imply a non-constant variance in the Beta distribution, and hence it would be difficult to adjust the acceptance rate. The truncated normal distribution is, in that sense, easier to manipulate.

## 3. Importance Sampling

Importance sampling takes random draws within the support, assigns a weight to each draw, and then takes a weighted average of the resulting draws to estimate the mean of the posterior distribution.

### *3.1 Sampling Distribution*

A truncated normal distribution is used again, however, in this case the mean of the distribution is dependent on the number of state transitions. For p11, for example, the mean is equal to

$$\frac{n11}{n11 + n12}$$

This is done to mostly have draws near the expected probability of the chain, as otherwise the weights are too insignificant. One could explore the whole parameter space, but in this case, it seems to be more useful to explore only around the probability estimate for a greater number of significant weights. To that end, a standard deviation of 0.05 is used.

### *3.2 Weights*

The weight is calculated by finding the likelihood of a given draw, divided by the density of the respective truncated normal distribution. The weights used for each probability are the same, as they are dependent on both the draws, but the estimates are calculated separately. The weights are scaled by a factor of 10^274, however this does not affect the mean calculated (It does affect the weighted variance of the draws and the variance of the weights).

The weights calculated for 10,000 draws are shown below.

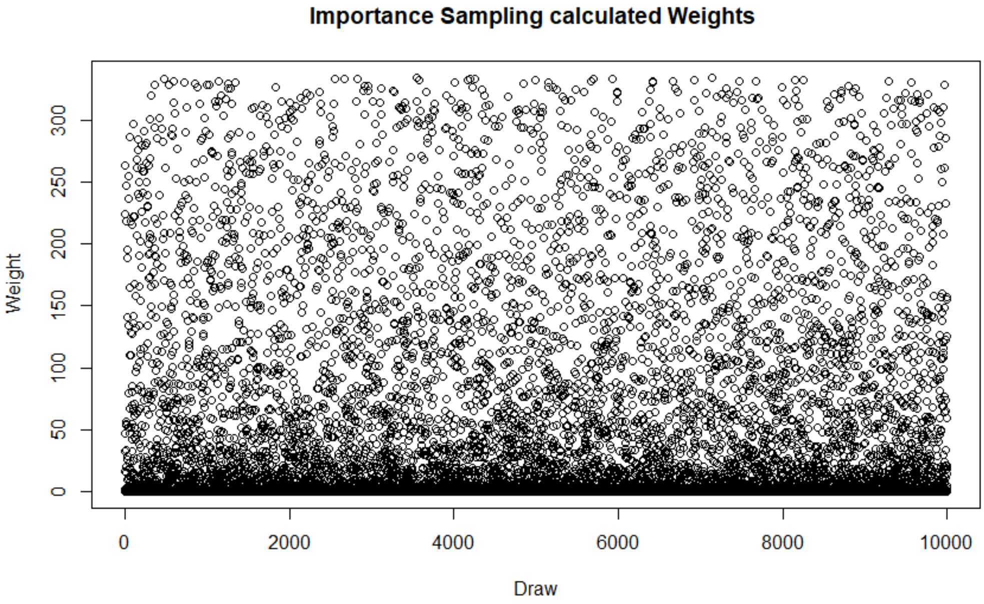

*Figure 14 Importance Sampling Weights Using Truncated Normal Sampling (Scaled)*

There are a significant number of observations at a high weight, and thus the estimate is not influenced by only one or two observations.

### *3.3 Results*

In this case, we do not burn in any draws, as an unlikely candidate will have a very low weight, and there is no initialization which prevents convergence. Moreover, we do not need to trim draws as they are uncorrelated.

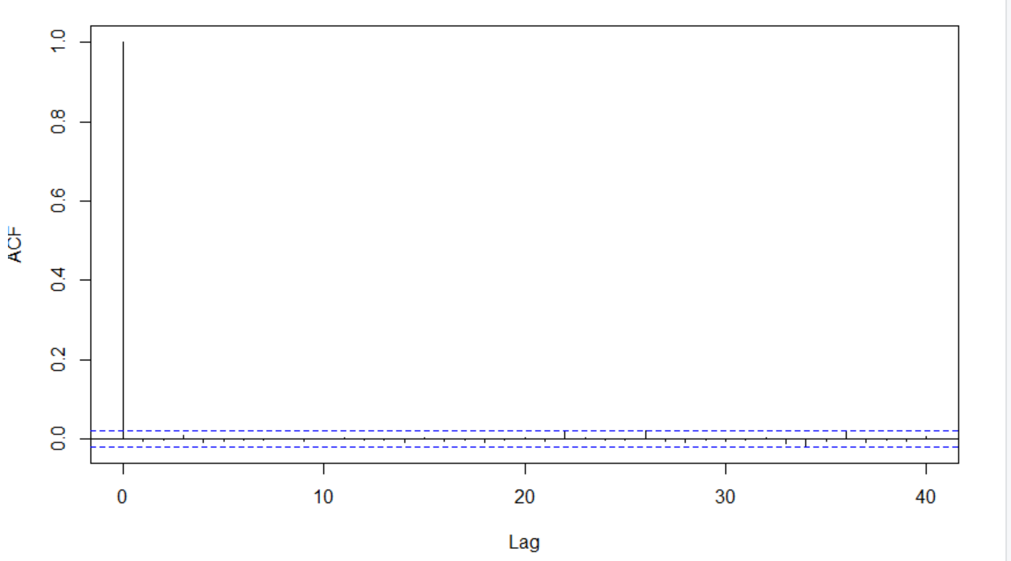

*Figure 15 Autocorrelation function for P11 draws in Importance Sampling*

For the weighted average of 10000 draws, the calculated p11 is 0.2027, and the calculated p21 is 0.4477, both of which are quite close to the respective true parameters. The root mean squared error of the mean estimates is 0.0031. The graphs below show the convergence of the average with more draws from the sampling distribution.

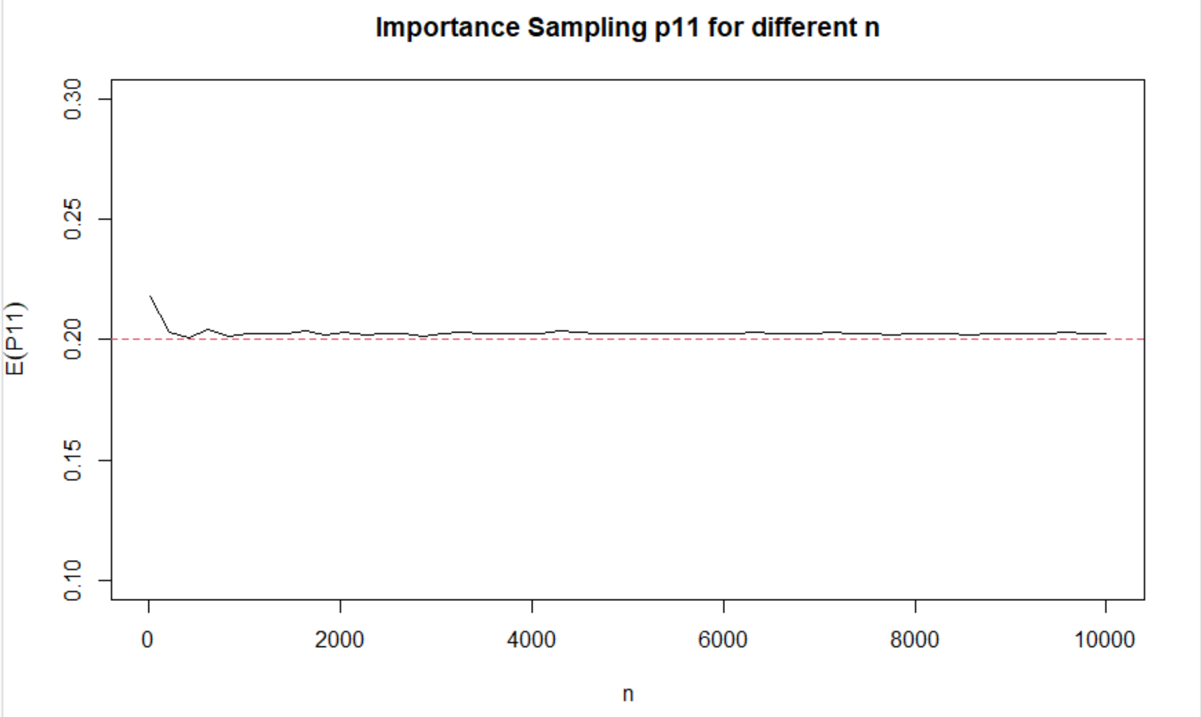

*Figure 16 Estimated Mean of Posterior with Different Number of Draws P11*

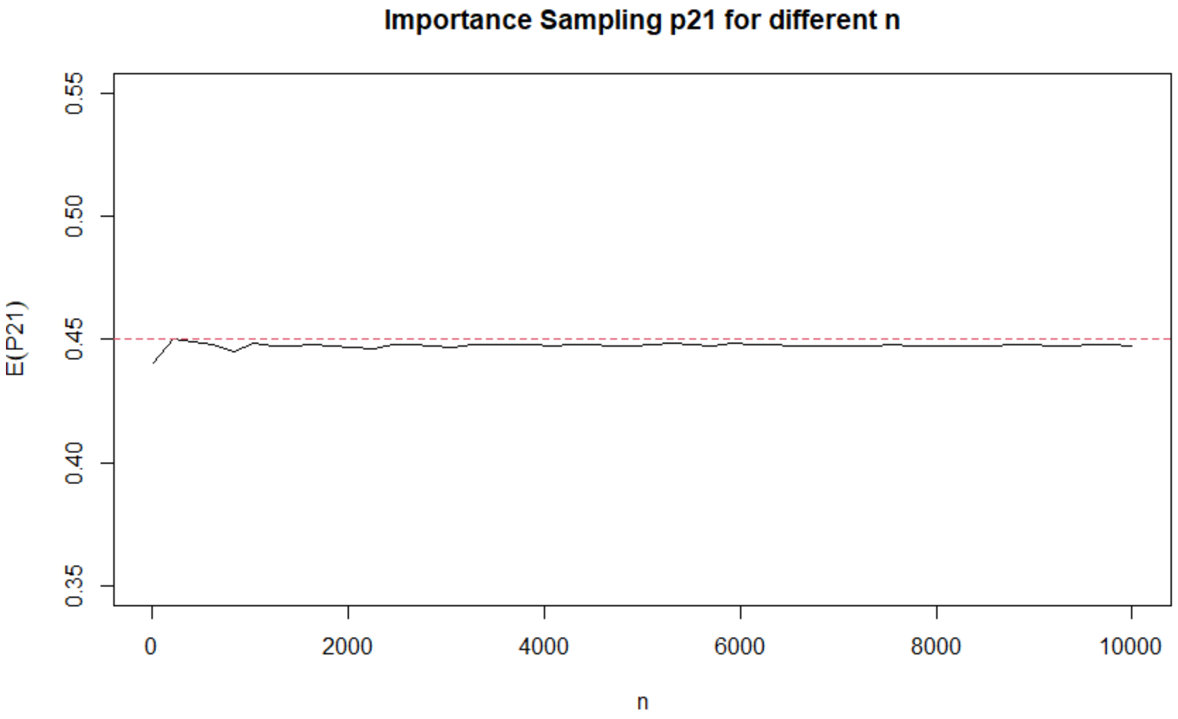

*Figure 17 Estimated Mean of Posterior with Different Number of Draws P21*

### *3.4 Elaboration on choice of sampling distribution*

In the importance sampling, one can choose several sampling distributions. The only constraint in this instance is that the sampling distribution should return a probability, or more simply a number between 0 and 1. This implies that distributions like a uniform distribution on the interval of 0 to 1, or a Beta distribution are valid candidates.

*For the uniform distribution:*

Given that there is no information about the transition probabilities except for the state transitions, it would be very plausible to use a uniform distribution to sample from. Such a distribution would explore the entire parameter space and weight would only be given to those parameter values which are close to the true parameter values. The only problem is that this likelihood function requires probabilities which are very close to the real values to get significant weights. While this is not a problem for any single probability, both being close to their true value in one draw is unlikely.

To that end, the uniform distribution does give decent results, but requires a very large number of draws relative to the truncated normal distribution in estimating close to the true probability. Moreover, the weights are only high for some draws, which makes them more important than any other draws to estimate the posterior distribution mean.

The following figures show the draws and the weights for a uniform distribution importance sample with 10000 draws.

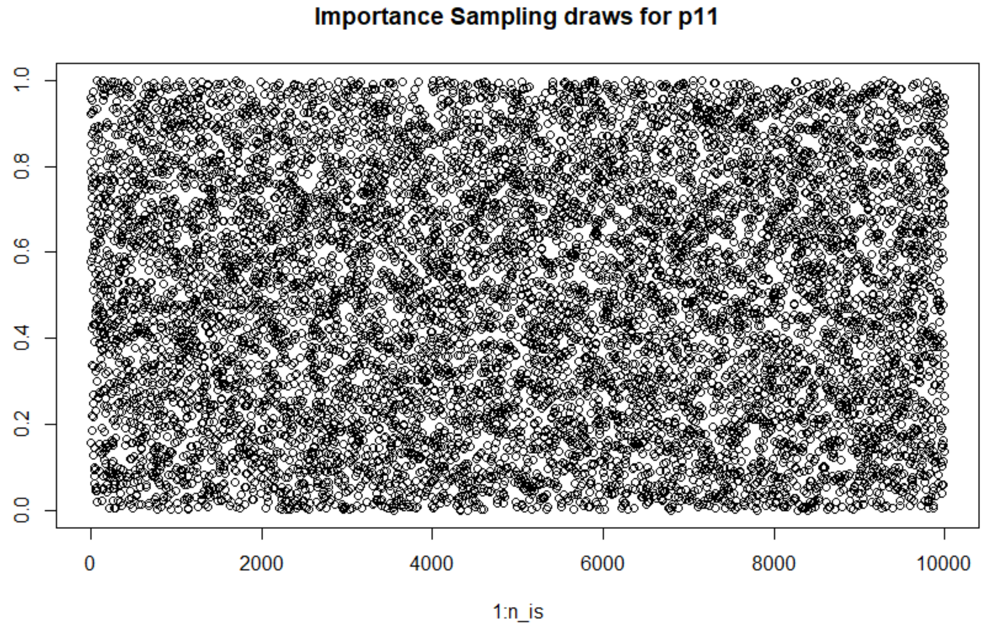

*Figure 18 Uniform distribution P11 Draws*

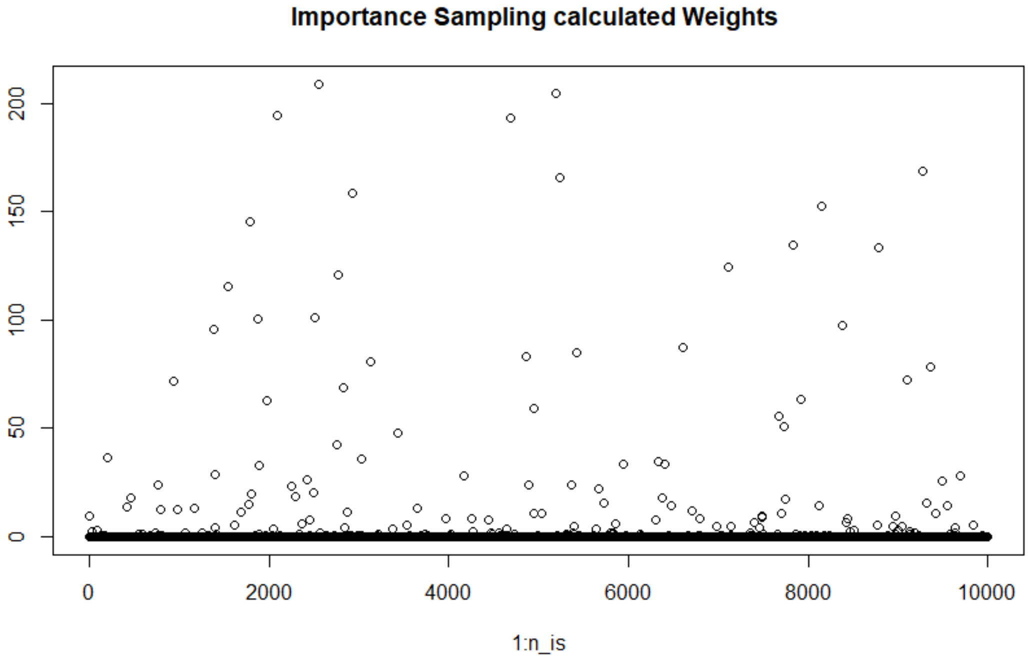

*Figure 19 Uniform Distribution Importance Sampling Weights*

Therefore, the uniform distribution is not the preferred distribution to sample from.

*For the Beta distribution:*

In using a Beta distribution, the main challenge is in selecting the shape parameters alpha and beta. An obvious choice is (for p11) to use the $n_{11}$ as alpha and $n_{12}$ as beta, as that would make the mean of the Beta distribution the same as the estimate for the transition probability. The only worry would be that the variance would not be constant across different Markov Chains, and cannot be adjusted as easily as in case of the Truncated Normal distribution (as alpha and beta can only be scaled by the same factor to preserve the mean). Moreover, adjusting the variance can skew the distribution. The following figure shows the density of a beta distribution with alpha = $n_{11}$/100 and beta = $n_{12}$/100.

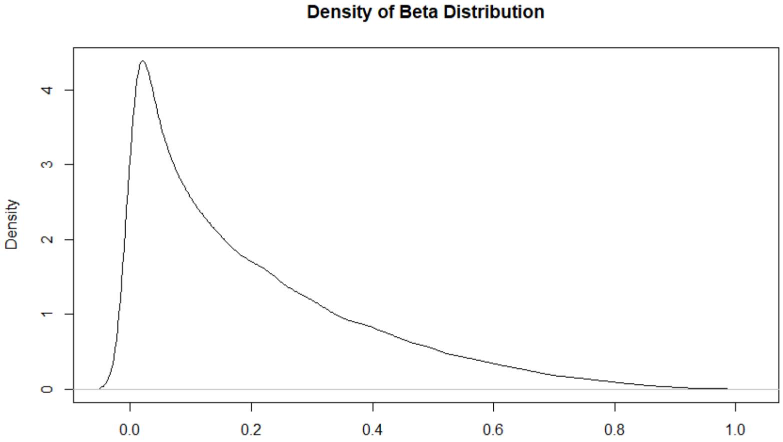

*Figure 20 Density of a Beta distribution with alpha = $n_{11}$/100 and beta = $n_{12}$/100*

While this is not inherently an issue (as the weighing function would adjust for the relative density of different values), it could potentially increase the number of draws needed for a ‘good’ estimate of the mean of the posterior. Moreover, it is easier to change the standard deviation of the truncated normal to observe a specific wider interval around the mean.

Hence, the truncated normal is used over the beta distribution for convenience in setting the variance. The truncated normal distribution used does not suffer from any drawbacks which the beta distribution would solve.  However, the beta distribution still provides a reasonable estimate for the mean of the posterior distribution in 10000 draws (E(P11) = 0.2023, E(P21) =0.4470) and can be used instead. The following figures show, for P11, the weights (scaled by a factor of 10^273) and draws for 10000 draws taken and different mean estimates for different number of draws for a beta distribution with alpha = $n_{11}$/10 and beta = $n_{12}$/10.

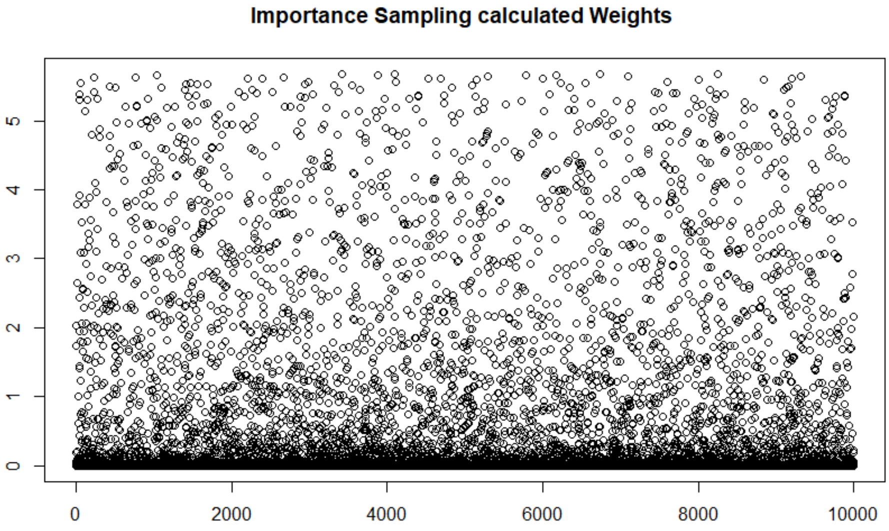

*Figure 21 Density of a Beta Distribution with alpha = n11/10 and beta = n12/10*

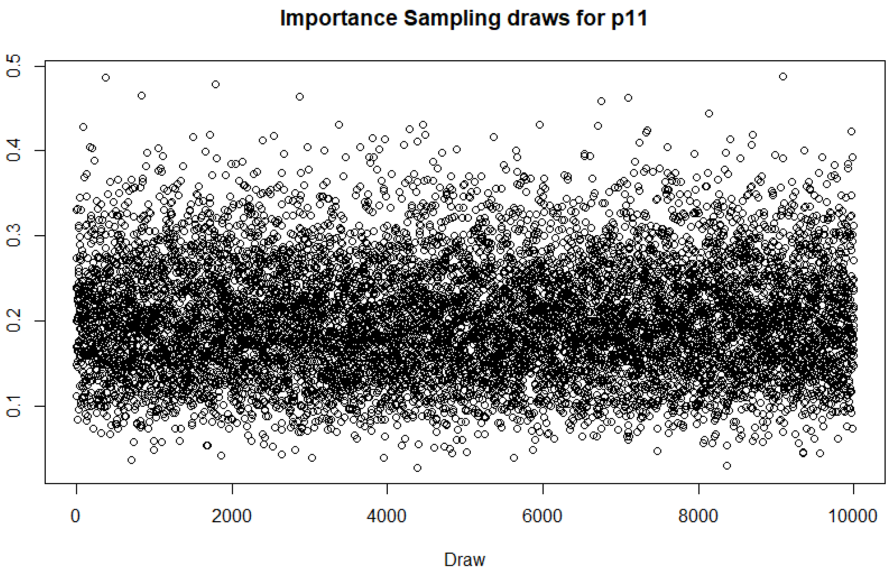

*Figure 22 Draws from Beta Distribution for P11*

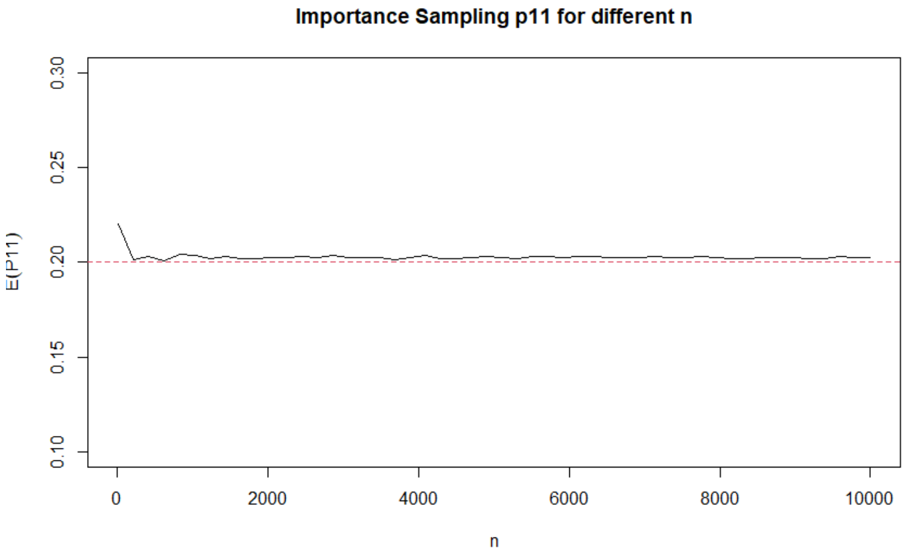

*Figure 23 Estimated Mean of Posterior with Different Number of Draws P11*

*With respect to using the log-likelihood:*

Using the log-likelihood when calculating the weights is a viable option. It increases the numerical stability of the weights, makes them more varied and increases the range. However, a similar result is achieved by the scaling done to the weights. To that end, the log-likelihood is not used while displaying results, but its viability is acknowledged.

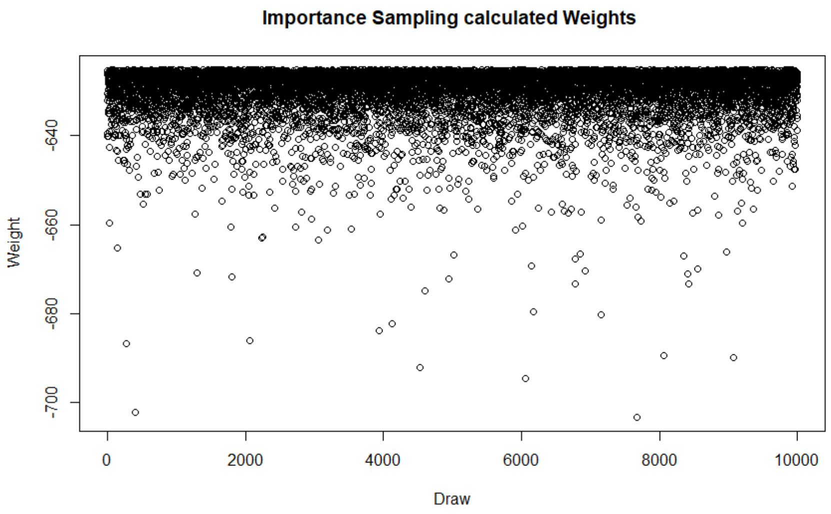

*Figure 24 Weights for Truncated Normal Importance Sampling, calculated using Log-likelihoods*

## 4. Relative Performance of the two methods

In practise, in this simulation one would expect similar expected values from both the importance sampling and Metropolis-Hastings methods. The major difference between the two is that in this implementation Metropolis-Hastings uses a candidate distribution which is reliant on previous draws, while the Importance sampler samples from the same distribution throughout. However, because of how they behave (reject/give low weight to unlikely values), the two algorithms lead to very similar results, with estimates for the posterior mean values very close to the true parameter values.

| Algorithm | Metropolis-Hastings | Importance-Sampling |
|---|---|---|
| Estimate of P11 | 0.2037 | 0.2027 |
| Estimate of P21 | 0.4475 | 0.4477 |
| Root Mean Squared Error | 0.0031 | 0.0025 |

The importance sampling algorithm has a lower root mean squared error. However, this is a minor and relatively negligible difference.

In this case, due to the simplicity in choosing the sampling distribution, the importance sampling algorithm does have a few advantages. First, it does not require a burn-in period nor does it require trimming of draws. This is because an independent and identically distributed sample is being observed, and unlikely values are simply weighed lower. Second, a wider variety of sampling distributions can be used, although their effectiveness and number of draws required may differ. Third, even a smaller sample typically leads to a credible estimate (Estimated P11 with 500 draws in importance sampling is 0.2017), while this is not the case in Metropolis-Hastings due to the burn-in.

## 5. Conclusion

This study points to the potential of Bayesian inference algorithms in estimating transition probabilities for a Markov Chain. Before conditioning on an instance of a Markov Chain, there was an uninformative, flat prior belief. This belief led to an expected value of 0.5 for both transition parameters P11 and P21. By using the Metropolis-Hastings algorithm and the Importance Sampling algorithm, it is possible to estimate the expected value of the posterior conditional distribution. This value comes very close to the true parameter estimates which were used to simulate the Markov Chain.

This paper goes over the application of Metropolis Hastings and Importance sampling on a first order, two state Markov Chain. These algorithms effectively let one sample from a posterior distribution which is difficult to sample directly from. They yield reliable and close estimates of the parameters in question for the first order, two-state Markov Chain. Future research can generalize the methods to a multi-state, multi-order Markov chain. Moreover, the impact of different prior beliefs on the posterior distribution can be studied.

## References

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120. https://doi.org/10.1214/aoap/1034625254