Monte Carlo algorithms project: DNA methylation of cytosines

X is an i.i.d with the following density f(x). A mixture of three beta distributions.

$$f(x) = \frac{1}{3}Be(x, \alpha_1 = 1, \beta_1 = 5) + \frac{1}{3}Be(x, \alpha_2 = 3, \beta_2 = 5) + \frac{1}{3}Be(x, \alpha_3 = 10, \beta_3 = 5)$$

1. Plot the density f(x).

The density f(x) can be plotted as follows in Figure 1.
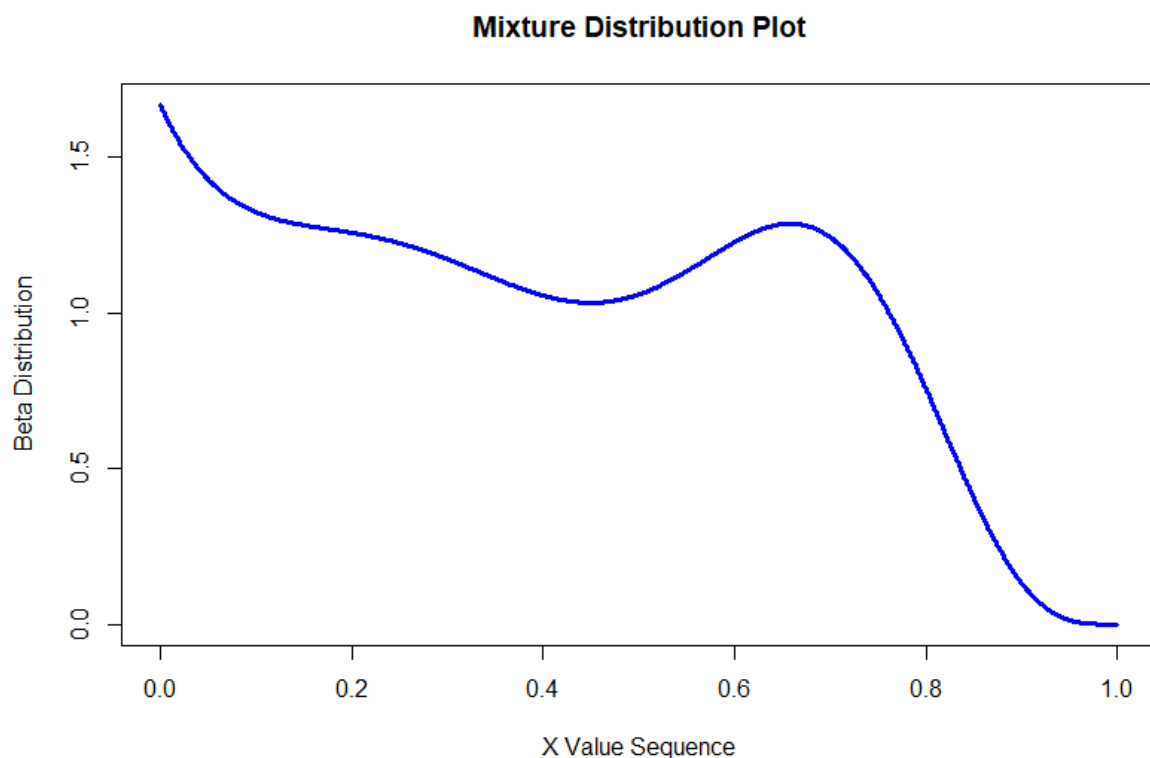

Figure 1

This was done by creating an array of x values between 0 and 1 and substituting the x values into the function and plotting it.

2. Implement an accept/reject algorithm, with Unif(0,1) proposal distribution.

K occurs at where the uniform distribution is maximum, this is to ensure that we have the tightest distribution and not reject excessive values. Looking at Figure 1 it can be seen that the highest point in the distribution occurs at x = 0 therefore K can be calculated by substituting 0 into f(x). Resulting in a value of K = 1.667.

Plotting this as a hard limit (shown by the dotted orange line in Figure 2.) allows us to define the bounds on Unif(0,1) such that we can have points in the tightest distribution.

Figure 2 shows a simulation of 10,000 values for the uniform distribution. Accepting values that have a probability of less than

$$\frac{\frac{1}{3}Be(x, \alpha_1 = 1, \beta_1 = 5) + \frac{1}{3}Be(x, \alpha_2 = 3, \beta_2 = 5) + \frac{1}{3}Be(x, \alpha_3 = 10, \beta_3 = 5)}{K}$$

$$\frac{\frac{1}{3}Be(x, \alpha_1 = 1, \beta_1 = 5) + \frac{1}{3}Be(x, \alpha_2 = 3, \beta_2 = 5) + \frac{1}{3}Be(x, \alpha_3 = 10, \beta_3 = 5)}{1.667}$$
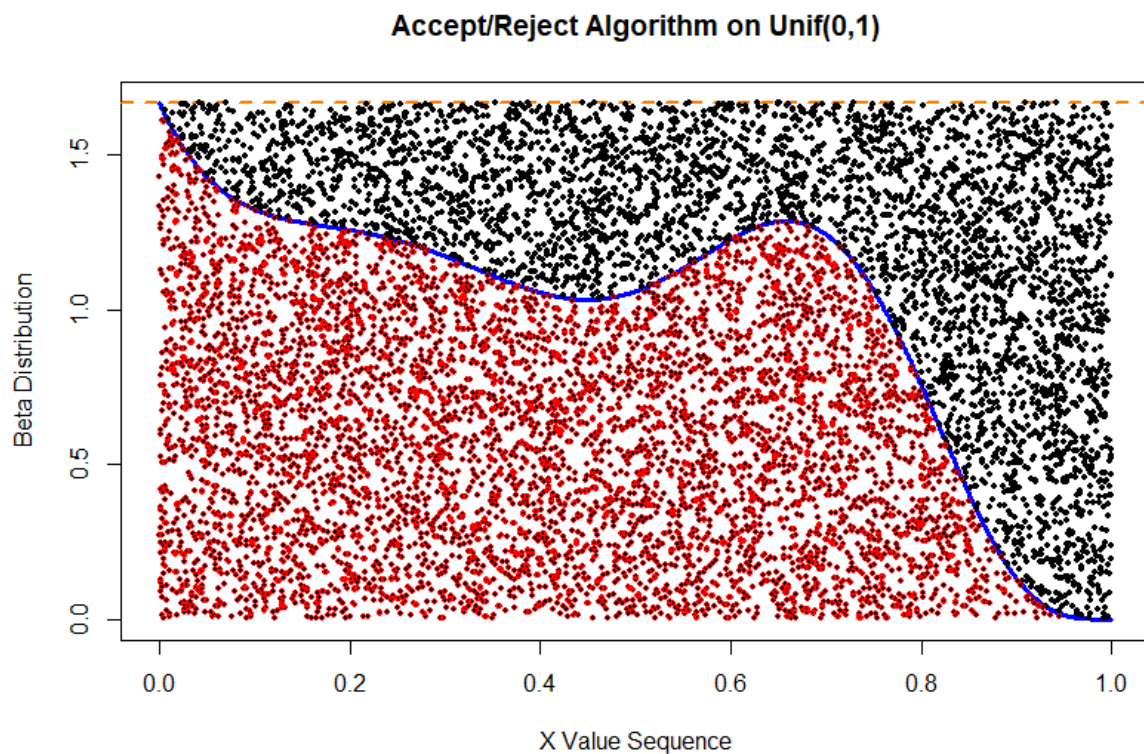

Figure 2

3. Compute the observed acceptance rate and compare it with the theoretical one

The observed acceptance rate, can be done by determining the total accepted values, and divide that by the original 10,000 values simulated. This results in an accepted rate of 0.5978. On the other hand the theoretical acceptance rate is

$$\frac{1}{K} = \frac{1}{1.667}$$

Which is calculated to be 0.6. Both values are very close to each other.

4. Implement an importance sampling algorithm.

The importance weights can be determined from the following equation

$$w(\theta) = \frac{\frac{1}{3}Be(x, \alpha_1 = 1, \beta_1 = 5) + \frac{1}{3}Be(x, \alpha_2 = 3, \beta_2 = 5) + \frac{1}{3}Be(x, \alpha_3 = 10, \beta_3 = 5)}{Unif(\theta)}$$

Where the Importance Sampling acceptance is the sum of the weights, and is calculated in R to be each quartile to be.

Table 1

25%	50%	75%
0.178	0.388	0.615

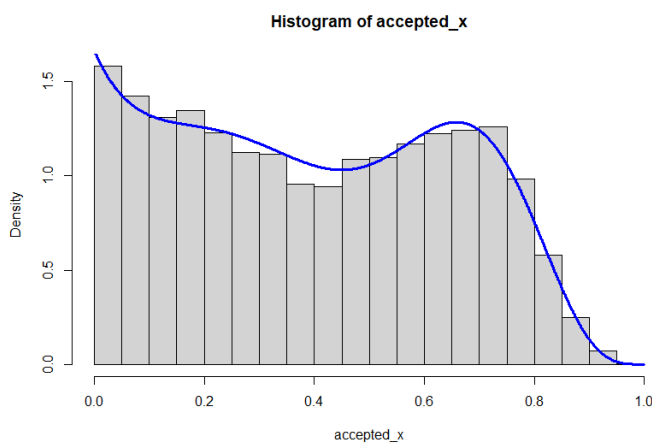## 5. Comparison of Histograms Distribution
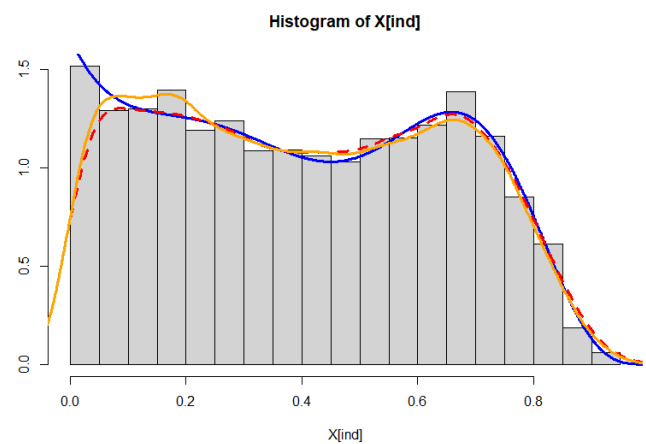

Figure 3   *Figure 4*

Figure 3, shows the histogram of the density of the accept reject algorithm, the density follows the distribution a lot closer, in places the distribution peaks, however the histogram of the importance sampling, shown in Figure 5, is a lot more closer to the distribution in the troughs of the function.

6. Which of the two algorithms seems to approximate the target distribution better? Which one you would use and why?

The Importance Sampling algorithm targets the distribution better. However, as this is a distribution that is bounded between 0 and 1, it is simpler to use the accept reject method, especially since only 40% of the simulations are rejected. Similarly it was very straightforward to calculate the K constant as the maximum point was clearly visible on the plot. If the situation was unbounded, I would have however picked the Importance sampling as that performs better with unbounded parameter space.