**Final Project for MA5770 - Bayesian Statistics**

**Name:** Prateek Singh, Xinyun Liu, Manoj Kumar Surabhi, Xiaolin Liu

**Project Title:** Bayesian Mixture Normal Model of Plasma Glucose Data

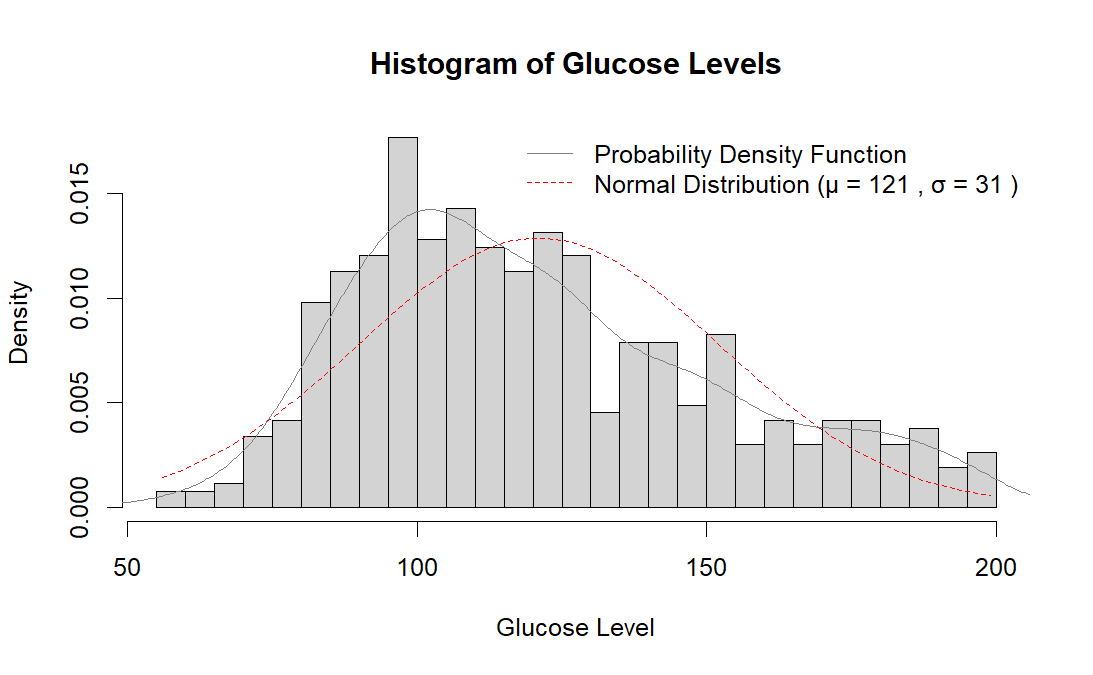
# Section 1: Introduction

The study analyzes plasma glucose levels from 532 females living near Phoenix, Arizona, who were tested for diabetes. The goal is to use a Bayesian approach with a mixture of normal distributions, each with different means and variances, to capture the data's skewness and outliers, where a single normal distribution may not adequately represent the data.

# Section 2: Statistical Analysis

The distribution of glucose levels is observed by the histogram and kernel density function of the data. The measure of central tendencies (mean, median and mode) and measure of dispersion (variance and standard deviation) are calculated. A Bayesian mixture model approximates the data distribution using a mixture of two normal models. The Gibbs and JAGS samplers are used to obtain samples to approximate the posterior distribution of the data. The analysis is performed using R Studio software and the JAGS software package.

Figure 1 shows the histogram, kernel density estimation, and normal distribution for the as 121.0 and as 31.0 for the plasma glucose concentration of 532 females. It can be seen that the data is rightly skewed. In addition, the sample mean, median, mode, variance, and standard deviation are respectively.



**Figure 1** - Histogram and Kernel Density Estimation and

# Section 3: Statistical Method

From Figure 1, we can see that a normal model should not be appropriate to model the data. So, we use a Bayesian mixture model. Let denote the plasma glucose level from the *i*th female. is assumed to have the following mixture normal distribution:

Where is the probability density function at for the normal distribution with mean and the variance with probability as , and is the probability at for the normal distribution with mean and the variance with probability as and where are unknown parameters.

We further assume that there is a latent variable xi which can take value with probability and with probability . The value of xi determines the distribution of . Specifically, the conditional distribution given and the parameters is:

| { |  |
| --- | --- |
|  |

Let denote the group membership which is equal to and 2 with probability and . For and the observation and respectively. is unknown and considered as missing data in the model. Therefore our Bayesian mixture normal model can be described as followed

For the sampling distribution, we assume are conditionally independent

|  |  |  |
| --- | --- | --- |
|  |

For the prior distribution, we assume 𝜋, 𝜃1, 𝜃2, 𝜎12, and 𝜎22 are independent. Therefore,

In addition, the following prior distributions are used:

* Where which is the probability density function at 𝜋 for the beta distribution with the parameter and .
* Where and as are probability density functions of the normal distribution at with mean and precision(inverse variance) as .
* Where and as are the probability density function of the inverse gamma distribution at 𝜎12, 𝜎22 and with shape parameter and scale parameter .

Based on the mentioned models, the following full conditional distributions canbe derived and used in Gibbs sampler.

|  |  |
| --- | --- |
|  |

where is the group assignment for all the participants except participant and where is the likelihood of observing from either group respectively.

* , note that where is the number of , is the number of and
* , note that is the sample mean among with corresponding . We can use the representation where is the indicator function and .
* , note that is the sample mean among with corresponding . We can use the representation where is the indicator function and .
* where is total squared deviation of for from the mean .
* where is total squared deviation of for from the mean .

# Section 4: Gibbs Sampler and MCMC Diagnostic

It is difficult to directly obtain the posterior distribution of , so the following Gibbs Sampler is implemented to sample .

First, we set .

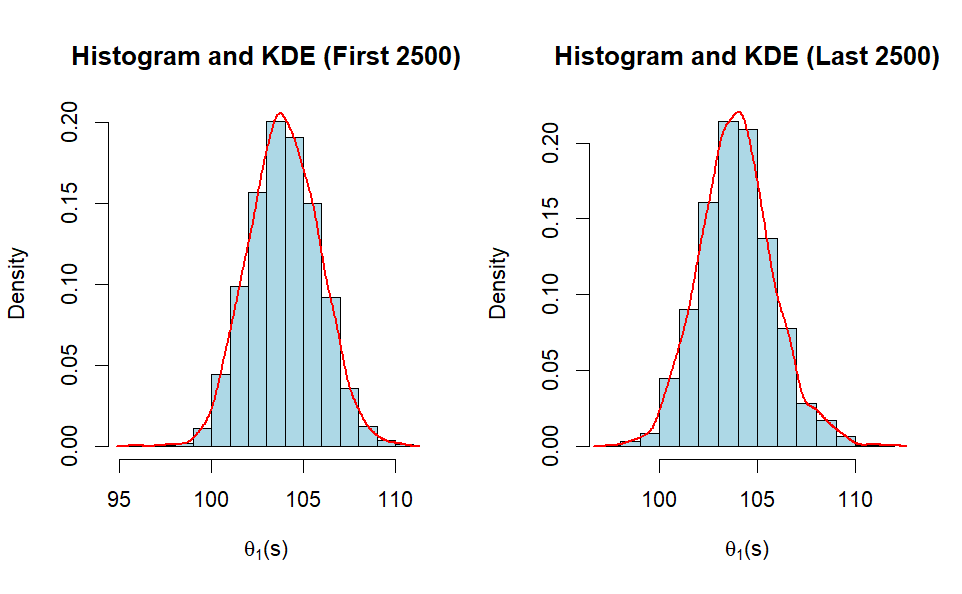
Second, we initialize according to . Specifically, we set as 1 if is less than mean of and 2 otherwise. We also set and as the sample variance of in the respective groups initialized by and .

Then for , we:

* Sample according to
* Sample according to
* Sample according to
* Sample according to . Specifically, we set as 1 with probability and as 2 with probability 1 - .
* Sample according to

To determine if the Gibbs sampler converges, we performed a number of MCMC diagnostics. In the diagnostic of . We first compare the histogram of splitted chains, examine the posterior distribution by calculating the posterior mean and 95% Confidence Interval, generate trace plots to visualize mixing, and finally compute autocorrelation factor and Effective sample size values generated from Gibbs sampler.

Firstly, comparing the histogram of and the histogram respectively and then subsequently, examining the posterior distributions by computing the posterior mean and the 95% Confidence Interval.

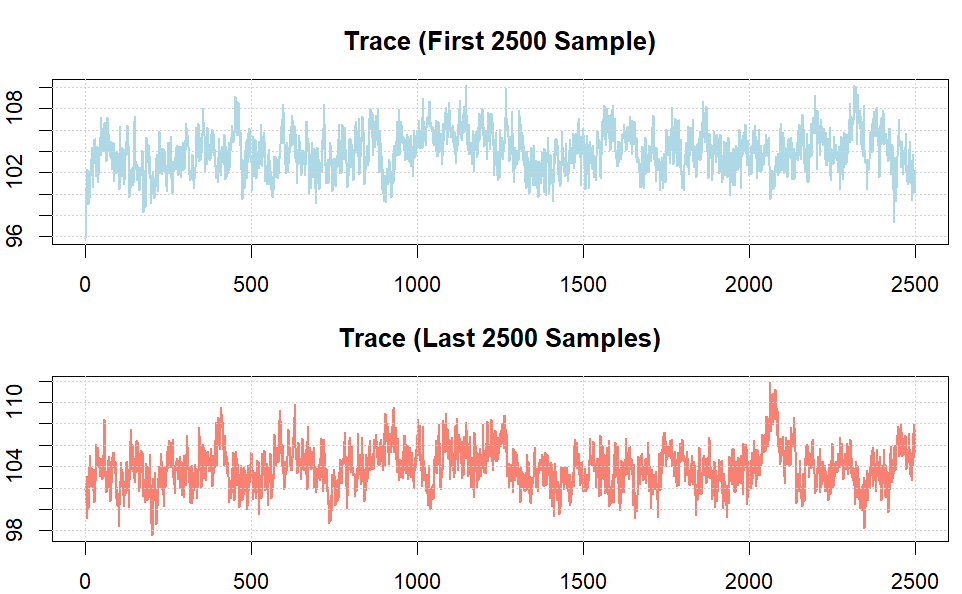


**Figure 2** - Histogram and Kernel Density Estimation for ,

|  |  |  |
| --- | --- | --- |
| **Mean** | 104.0 | 103.9 |
| **95% CI (Lower)** | 100.4 | 100.4 |
| **95% CI (Upper)** | 107.7 | 108.0 |

**Table 1** - Summary of Mean and 95% Confidence Interval for

From Table 1, it is evident that the posterior mean and 95% confidence intervals for the two chains align closely, indicating that the Gibbs sampler has likely reached convergence. The stability of these estimates suggests that sufficient mixing has occurred across iterations. To further ensure convergence, trace plots for , ​ for splitted samples. These plots are stable and show good mixing, as the chains explore the parameter space without evidence of autocorrelation or trends. Figure 3 illustrates the trace plot for .



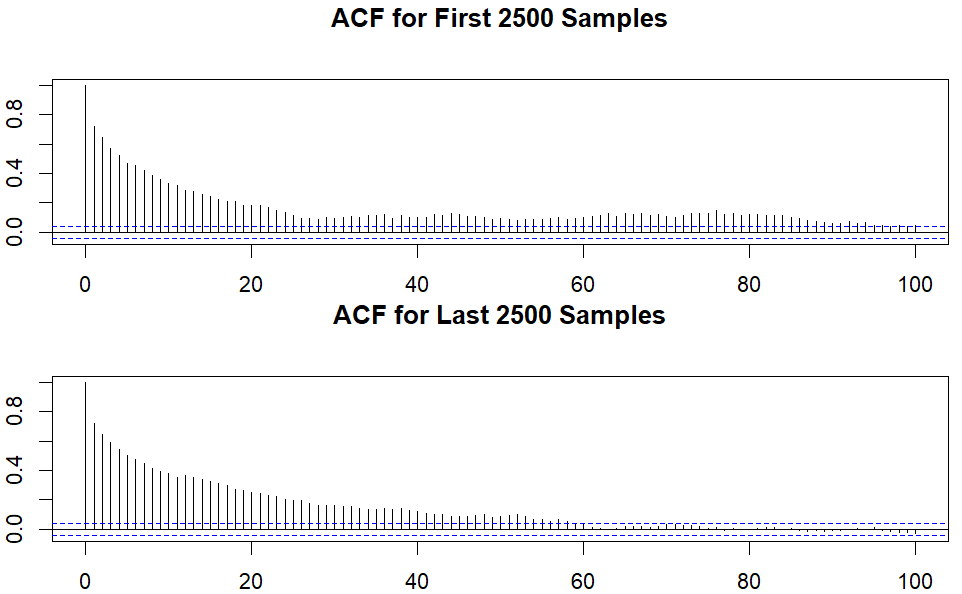
**Figure 3** - Trace plot of in blue and in red.

Additionally, we calculate the autocorrelation with a lag value of 100 to assess dependence between samples and determine the effective sample size. We visualize the ACF plot of correlation with its lag of 100.

|  |  |  |
| --- | --- | --- |
| **Autocorrelation (Lag 100)** | 0.1628 | 0.1419 |
| **Effective Sample Size** | 134 | 94 |

**Table 2** - Summary Autocorrelation and Effective Samples size for

From Table 2, it is evident that the autocorrelation at lag 100 for the two splits is relatively low (0.1628 and 0.1419), indicating moderate dependence between samples. The effective sample sizes (134 and 94) suggest that the Gibbs sampler is generating a reasonable number of independent samples.These results indicate that the chains are adequately mixed and have sufficient independent information for posterior inference.

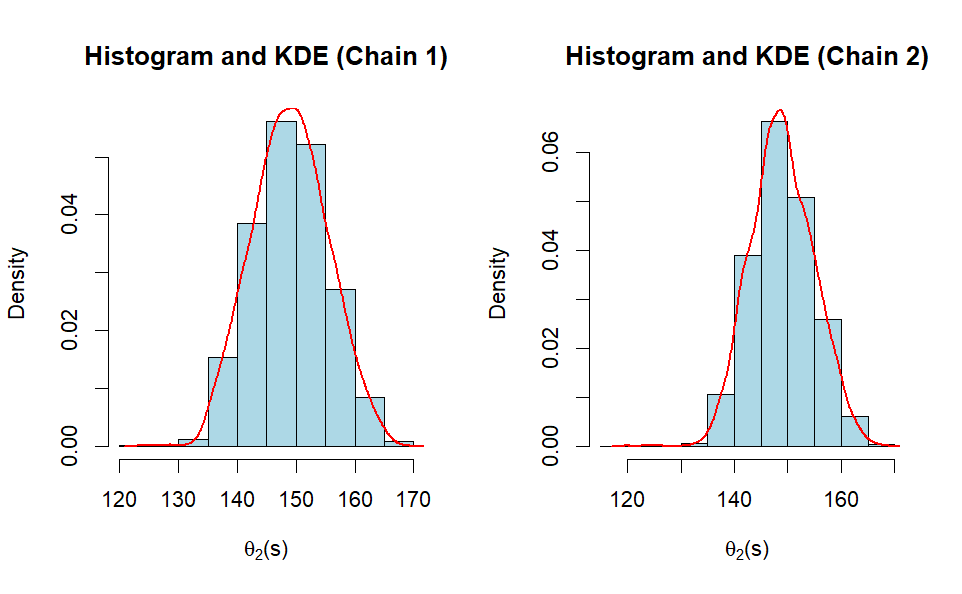


**Figure 4** - ACF plot for with a Lag of 100

To further evaluate dependence, the ACF plot for 1s with a lag of 100 (Figure 4) was analyzed. This plot demonstrates a gradual decline in autocorrelation across lags, confirming that the chains are exploring the parameter space effectively without excessive autocorrelation.

Alternatively, we can use Multiple chains to perform MCMC Diagnostics as well. For this We use is set as 1 or 2 with probability 0.5 to sample 5,000 and the initial value and . Similarly, for each chain and , we sample according to . For of ;

Firstly, comparing the histogram of from independent chains and then subsequently, examining the posterior distributions by computing the posterior mean and the 95% Confidence Interval.

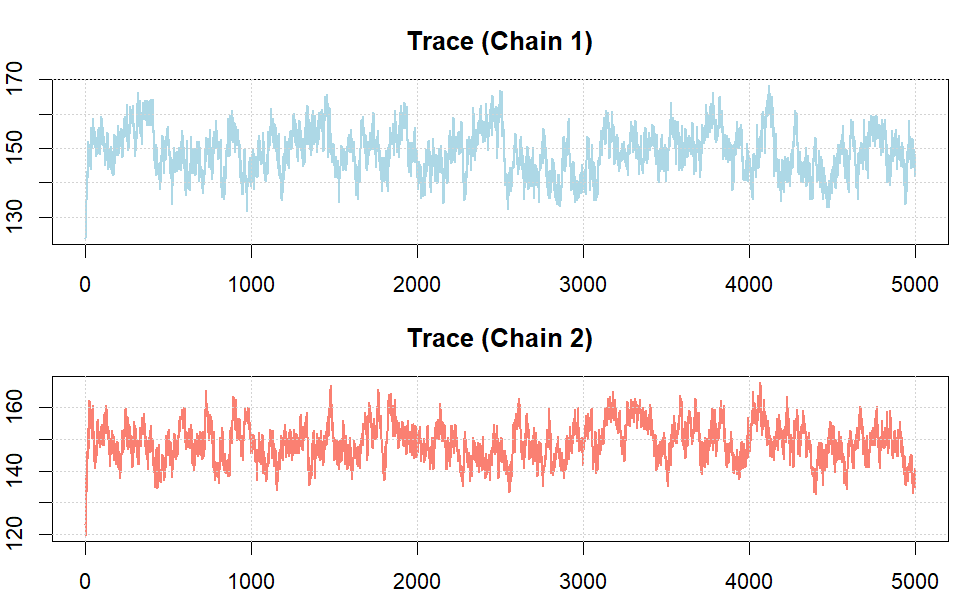


**Figure 5** - Histogram and Kernel Density Estimation for from two different chains

|  | **First chain** | **Second chain** |
| --- | --- | --- |
| **Mean** | 149.0 | 149.0 |
| **95% CI (Lower)** | 136.8 | 138.0 |
| **95% CI (Upper)** | 161.7 | 160.5 |

**Table 3** - Summary of Mean and 95% Confidence Interval for

From Table 3, it is evident that the posterior mean and 95% confidence intervals for the two chains align closely, indicating that the Gibbs sampler has likely reached convergence. The stability of these estimates suggests that sufficient mixing has occurred across iterations. To further ensure convergence, trace plots for across different chains. These plots are stable and show good mixing, as the chains explore the parameter space without evidence of autocorrelation or trends.



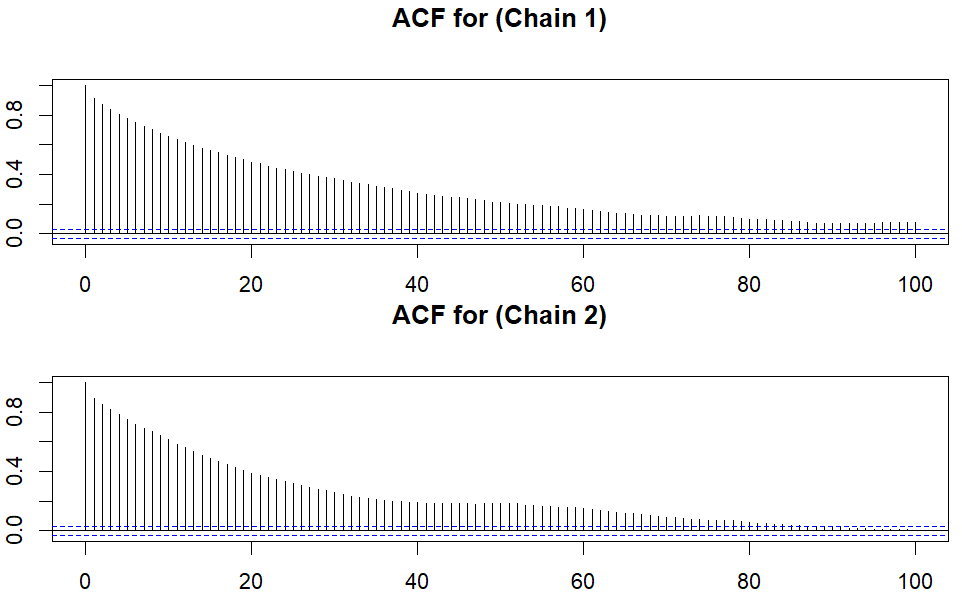
**Figure 6** - Trace plot of for Chain 1 in blue and Chain 2 in red respectively.

Additionally, we calculate the autocorrelation with a lag value of 100 to assess dependence between samples and determine the effective sample size. We visualize the ACF plot of correlation with its lag of 100.

|  | **First Chain** | **Second Chain** |
| --- | --- | --- |
| **Autocorrelation (Lag 100)** | 0.2958 | 0.2414 |
| **Effective Sample Size** | 98 | 118 |

**Table 4** - Summary Autocorrelation and Effective Samples size for

From Table 4, it is evident that the autocorrelation at lag 100 for the two splits is relatively low (0.2958 and 0.2415), indicating moderate dependence between samples. The effective sample sizes (98 and 118) suggest that the Gibbs sampler is generating a reasonable number of independent samples. These results indicate that the chains are adequately mixed and have sufficient independent information for posterior inference.



**Figure 7** - ACF plot for with a Lag of 100

After performing MCMC diagnostics using both sample splitting for and multiple chains for , the results show that the posterior means and 95% confidence intervals are consistent. Trace plots are stable, and the chains do not exhibit poor mixing. Autocorrelation values are low, indicating good exploration of the parameter space. However, the effective sample size (ESS) is relatively low compared to the total sample size, suggesting that increasing the number of iterations may be beneficial. Despite this, with 5000 samples, the results appear stable, and the chains seem to have converged, indicating that further increases in ESS may be limited.

# Section 5: JAGS and MCMC Diagnostic

In this section, we perform the same diagnostic analysis but with using JAGS.

model {

for (i in 1:n) {

X[i] ~ dcat(pi\_vals[])

y[i] ~ dnorm(

ifelse(X[i] == 1, theta1, theta2),

ifelse(X[i] == 1, 1 / s21, 1 / s22))

}

pi ~ dbeta(alpha, beta)

theta1 ~ dnorm(mu0, 1 / s21)

theta2 ~ dnorm(mu0, 1 / s22)

tau1 ~ dgamma(v0/2, v0 \* s20 / 2)

tau2 ~ dgamma(v0/2, v0 \* s20 / 2)

s21 <- 1 / tau1

s22 <- 1 / tau2

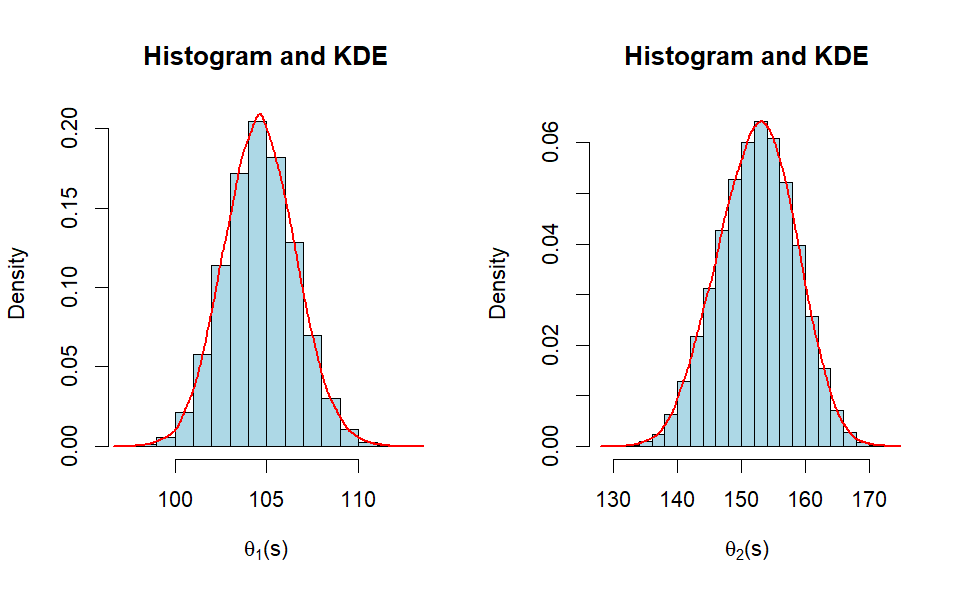
}

The JAGS model assumes that each observation is generated from one of the two normal distribution, determined by a latent variable which is modelled as a discrete distribution, with the probabilities taken from . The observed data is modeled as coming from one of the two distributions depending on the value of . If , then is drawn from normal distribution with mean and variance ; if , then is drawn from a normal distribution with mean and variance .

The parameters ​ and ​ represent the means of the two normal distributions and are assigned normal priors with mean and variance as reciprocal of and respectively. The variances and are modelled as inverse of precision parameters and which are assigned gamma prior with shape and rate parameters as and .

Additionally the mixing proportion for the latent variable is modelled using a Beta distribution with parameters and . The Beta prior reflex the initial beliefs about the probability distribution of .

Performing the same analysis as in the previous section, histograms of the posterior distributions for ​ and ​, alongside their posterior means and 95% credible intervals. Additionally, we display trace plots and autocorrelation function (ACF) plots to assess the mixing and convergence of the chains, and compute the effective sample size (ESS) to ensure sufficient sampling for reliable inference.

****

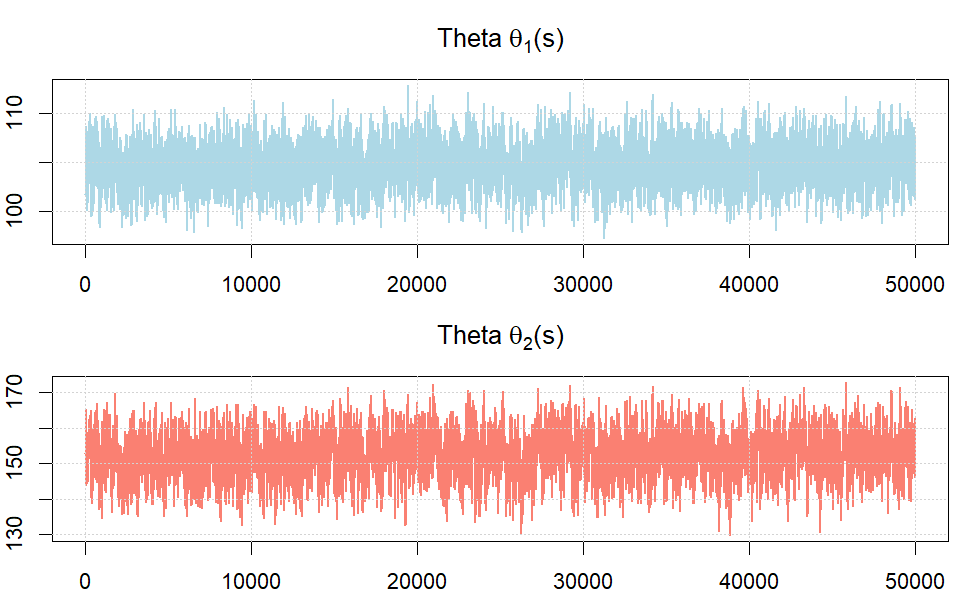
**Figure 8** - Histogram and Kernel Density Estimation for and

|  |  |  |
| --- | --- | --- |
| **Mean** | 104.7 | 152.6 |
| **95% CI (Lower)** | 100.9 | 140.5 |
| **95% CI (Upper)** | 108.6 | 163.6 |

**Table 5** - Summary of Mean and 95% Confidence Interval for and

From Table 5, it is evident that the mixture model distinguishes between two distinct normal distributions. The posterior mean for is 104.7, while for is 152.6. The non-overlapping 95% confidence intervals for [100.9, 108.6] and [140.5, 163.6] further supports the separation of these two components. This clear distinction in means and intervals highlights the adequacy of the mixture normal model. To further ensure convergence, trace plots for and across different chains. These plots are stable and show good mixing, as the chains explore the parameter space without evidence of autocorrelation or trends.

Additionally, we calculate the autocorrelation with a lag value of 100 to assess dependence between samples and determine the effective sample size. We visualize the ACF plot of correlation with its lag of 100.



**Figure 9** - Trace plot of in blue and in red

Additionally, we calculate the autocorrelation with a lag value of 100 to assess dependence between samples and determine the effective sample size. We visualize the ACF plot of correlation with its lag of 100.

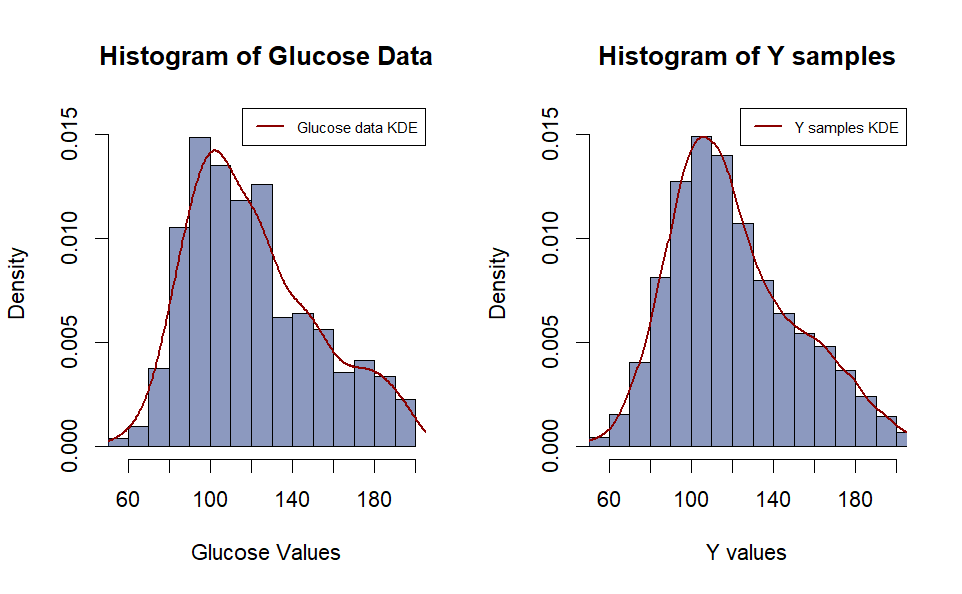
|  |  |  |
| --- | --- | --- |
| **Autocorrelation (Lag 100)** | 0.1767 | 0.2566 |
| **Effective Sample Size** | 1453 | 969 |

**Table 6** - Summary Autocorrelation and Effective Samples size for

In conclusion, both the Gibbs Sampler and JAGS methods provide consistent results in terms of posterior means and credible intervals for the parameters ​ and ​. The histograms, posterior means, and 95% credible intervals derived from both approaches are very similar, indicating reliable inference. Trace plots demonstrate stable mixing, and autocorrelation values suggest good exploration of the parameter space. However, the effective sample size (ESS) in the Gibbs Sampler was relatively lower than in JAGS, where the ESS was significantly higher, suggesting that JAGS performs better in terms of convergence and sampling efficiency. Overall, both methods exhibit convergence, but JAGS offers a more effective solution for sampling in this case.

# Section 6: Model Checking and Summary Statistics

We generated 50,000 samples from the mixture model using Gibbs sampling and JAGS, and then compared the results through their kernel density estimates (KDE). After obtaining the posterior samples for the mixture model parameters , the latent variable and then sample . The histogram for Glucose data and samples Y values from JAGS sampler are observed and compared.



**Figure 9** - Histogram of Glucose Data and Sampled Y values from Posterior Distribution

The KDE of glucose data is skewed with the peak leaning a lot more towards lower glucose values, with a tail towards higher glucose values and a slight second peak towards those higher glucose values. The KDE of sampled Y values from Gibbs sampler follows a similar pattern with slight minor differences. Both histograms and KDE have a similar shape and distribution. So the two-component model is reasonably good to capture the overall shape and spread of our glucose data distribution. Slight differences in distribution's peak and tail can be tuned by a more complex model like hierarchical model .

| **Theta** | **Method** | **Mean** | **95% CI (Lower)** | **95% CI (Upper)** | **ACF(Lag 100)** | **ESS**  **(S=...)** |
| --- | --- | --- | --- | --- | --- | --- |
|  | Gibbs - [First Half] | 104.0 | 100.4 | 107.7 | 0.1628 | 134  (2,500) |
| Gibbs - [Last Half] | 104.0 | 100.4 | 108.0 | 0.1419 | 94  (2,500) |
| JAGS | 104.7 | 100.9 | 108.6 | 0.1767 | 1453  (50,000) |
|  | Gibbs - [Chain 1] | 149.0 | 136.8 | 161.7 | 0.2958 | 98  (5,000) |
| Gibbs - [Chain 2] | 149.0 | 138.0 | 160.5 | 0.2414 | 118  (5,000) |
| JAGS | 152.6 | 140.5 | 163.6 | 0.2566 | 969  (50,000) |

**Table 6** - Summary of and for Gibbs and JAGS

The results from JAGS further validate the effectiveness of the two-component mixture model in capturing the glucose data distribution. The posterior means for ​ and ​ from JAGS align closely with those from Gibbs sampling, with and . The 95% credible intervals for and from JAGS are slightly wider, indicating slightly greater uncertainty, but the overall distributions remain consistent. The trace plots from JAGS exhibit stable and well-converged behavior, with chains mixing efficiently and showing no evident trends, further confirming convergence. The histograms and KDE comparisons reinforce that the two-component model effectively captures the skewness and dual peaks in the glucose data distribution, and the stable trace plots indicate the reliability of the posterior estimates.

# Section 7: Conclusion and Discussion

The analysis of plasma glucose levels using a Bayesian mixture model with two normal components successfully captured the data's key features, including its inherent skewness and the presence of outliers. Unlike a single normal distribution, this approach effectively modeled the data by identifying two distinct subpopulations with unique means and variances. These results demonstrate the utility of the mixture model in providing a detailed and accurate understanding of glucose distribution patterns, underscoring the broader applicability of Bayesian methods to complex datasets.

The implementation of the Gibbs sampler showed clear evidence of convergence, supported by diagnostic metrics such as trace plots, effective sample size, and posterior summaries. Additionally, the use of JAGS significantly simplified the process of model specification and posterior sampling, ensuring efficient estimation of parameters while validating the robustness of the results. This highlights the advantages of integrating JAGS into Bayesian analyses, particularly for addressing intricate modeling challenges.