Homework 6, STA 360

Austin Kao

1. Lab Component

```
# Code for tasks 1-3
knitr::opts_chunk$set(cache=TRUE)
library(MASS)
data <- read.csv("data/data-exponential.csv", header = FALSE)
set.seed(12345)
########################################
# This function is a Gibbs sampler
#
# Args
#   start.a: initial value for a
#   start.b: initial value for b
#   n.sims: number of iterations to run
#   data:  observed data, should be in a
            # data frame with one column
#
# Returns:
#   A two column matrix with samples
     #   for a in first column and
# samples for b in second column
########################################
```

```
knitr::opts_chunk$set(cache=TRUE)
sampleGibbs <- function(start.a, start.b, n.sims, data){
  # get sum, which is sufficient statistic. note: sum(x) = n*x_bar.
  x_sum <- sum(data)
  # get n
  n <- nrow(data)
  # create empty matrix, allocate memory for efficiency
  res <- matrix(NA, nrow = n.sims, ncol = 2)
  res[1,] <- c(start.a,start.b)
  for (i in 2:n.sims){
    # sample the values
    res[i,1] <- rgamma(1, shape = n+1,
                       rate = res[i-1,2]*x_sum+1)
    res[i,2] <- rgamma(1, shape = n+1,
                       rate = res[i,1]*x_sum+1)
  }
  return(res)
}
```

```r
# run Gibbs sampler
n.sims <- 10000
res <- sampleGibbs(.25,.25,n.sims,data)
head(res)
```

```
##          [,1]      [,2]
## [1,] 0.250000 0.2500000
## [2,] 2.101044 0.2701421
## [3,] 1.727011 0.2908984
## [4,] 2.404129 0.2410490
## [5,] 1.861079 0.3559962
## [6,] 1.247769 0.4267478
```

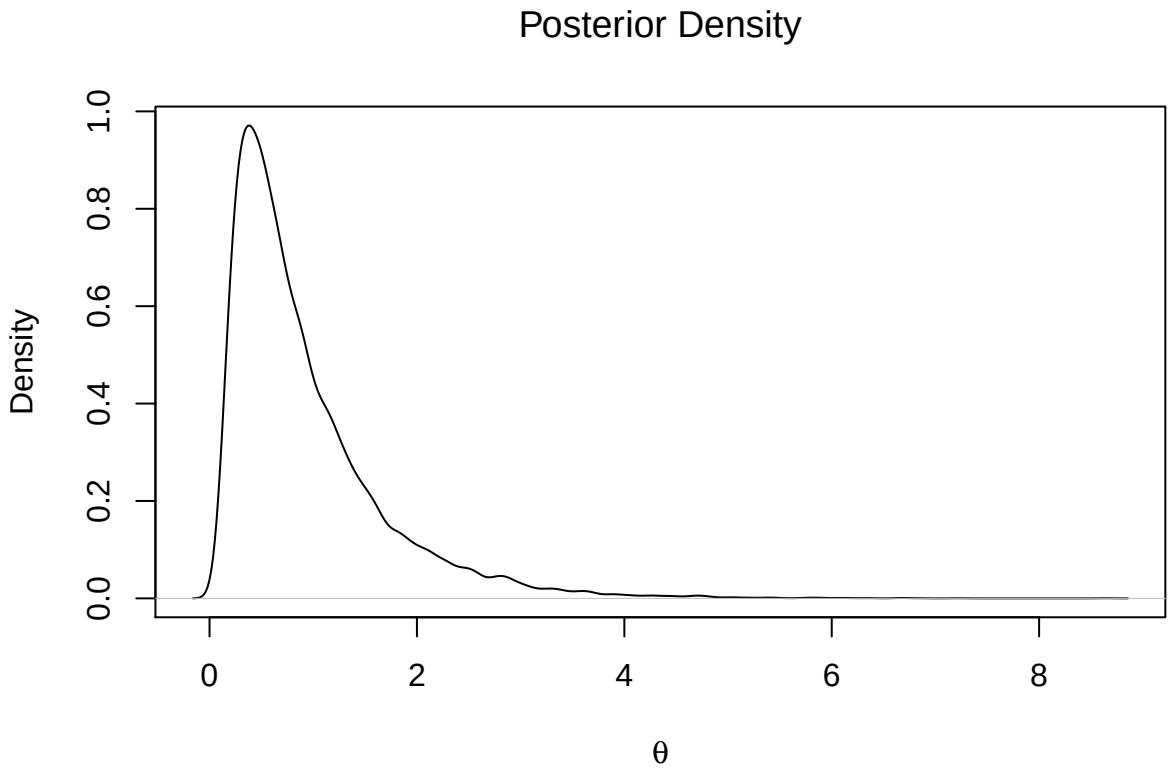
```r
dim(res)
```

```
## [1] 10000     2
```

```r
res[1,1]
```

```
## [1] 0.25
```

```r
# Code for task 4
plot(density(res), xlab = expression(theta),
     main = expression(paste("Posterior Density")))
```



I know that that the estimated posterior in (3) is reliable.

2. Censored Gibbs sampling:

Researchers are studying the length of life (lifetime) following a particular medical intervention, such as a new surgical treatment for heart disease, where the study consists of 12 patients. Specifically, the number of years before death for each is

$$3.4, 2.9, 1.2+, 1.4, 3.2, 1.8, 4.6, 1.7+, 2.0+, 1.4+, 2.8, 0.6+$$

where the + indicates that the patient was alive after x years, but the researchers lost contact with the patient after that point in time.

One way we can model this data is in the following way:

$$X_i = \begin{cases} Z_i & \text{if } Z_i \leq c_i \\ c_i & \text{if } Z_i > c_i \end{cases} \tag{1}$$

$$Z_1, \ldots, Z_n | \theta \overset{iid}{\sim} \text{Gamma}(r, \theta) \tag{2}$$

$$\theta \sim \text{Gamma}(a, b) \tag{3}$$

where a, b, and r are known. In addition, we know:

- c_i is the censoring time for patient i, which is fixed, but known only if censoring occurs.
- X_i is the observation
  - if the lifetime is less than c_i then we get to observe it ($X_i = Z_i$),
  - otherwise all we know is the lifetime is greater than c_i ($X_i = c_i$).
- θ is the parameter of interest—the rate parameter for the lifetime distribution.
- Z_i is the lifetime for patient i, however, this is not directly observed.

The probability density function (pdf) associated consists of two point masses:one at Z_i and one at c_i. The formula is

$$p(x_i|z_i) = \mathbf{1}(x_i = z_i)\mathbf{1}(z_i \leq c_i) + \mathbf{1}(x_i = c_i)\mathbf{1}(z_i > c_i).$$

.

Now we can easily find the full conditionals (derived in class and reproduced below). Notice that z_i is conditionally independent of z_j given θ for $i \neq j$. This implies that x_i is conditionally independent of x_j given z_i for $i \neq j$. Now we have

$$\begin{aligned} p(z_i|z_{-i}, x_{1:n}, \theta) &= p(z_i|x_i, \theta) \\ &\underset{z_i}{\propto} p(z_i, x_i, \theta) \\ &= p(\theta)p(z_i|\theta)p(x_i|z_i, \theta) \\ &\underset{z_i}{\propto} p(z_i|\theta)p(x_i|z_i, \theta) \\ &= p(z_i|\theta)p(x_i|z_i). \end{aligned}$$

There are now two cases to consider. If $x_i \neq c_i$, then $p(z_i|\theta)p(x_i|z_i)$ is only non-zero when $z_i = x_i$. The density devolves to a point mass at x_i. This corresponds to the case where z_i is observed, so x_i is the observed value and we should always sample this value. Practically speaking, we do not sample this value when running the Gibbs sampler.

If $x_i = c_i$, then the density becomes $p(x_i|z_i) = \mathbf{1}(z_i > c_i)$, so

$$p(z_i|\ldots) \propto p(z_i|\theta)\mathbf{1}(z_i > c_i),$$

which is a truncated Gamma.

For the Gibbs sampler, we will use the current value of θ to impute the censored data. We will sample from the truncated gamma using a modified version of the iverse CDF trick. For the censored values of Z_i we know c_i. If we know θ (which we will in a Gibbs' sampler), we know the distribution of $Z_i|\theta \sim Gamma(r, \theta)$. Let F be the CDF of this distribution. Suppose we truncate this distribution to (c, ∞). The new CDF is

$$P(Z_i < z) = \frac{F(z) - F(c)}{1 - F(c)}.$$

Therefore Y is a sample from the truncated Gamma, as desired.

In the actual code for the Gibbs' sampler we do not sample the observed values. We simply impute the censored values using the method above.

You will find code below (that is also taken from class) that will help you with the remainder of the problem.

1. (5 points) Write code to produce trace plots and running average plots for the censored values for 200 iterations. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.

```
knitr::opts_chunk$set(cache=TRUE)
library(xtable)
# set seed for reproducibility
set.seed(123)
# Samples from a truncated gamma with
# truncation (t, infty), shape a, and rate b
# Input: t,a,b
# Output: truncated Gamma(a,b)
sampleTrunGamma <- function(t, a, b){
  # This function samples from a truncated gamma with
  # truncation (t, infty), shape a, and rate b
  p0 <- pgamma(t, shape = a, rate = b)
  x <- runif(1, min = p0, max = 1)
  y <- qgamma(x, shape = a, rate = b)
  return(y)
}
# Gibbs sampler for censored data
# Inputs:
  # this function is a Gibbs sampler
  # z is the fully observe data
  # c is censored data
  # n.iter is number of iterations
  # init.theta and init.miss are initial values for sampler
  # r,a, and b are parameters
  # burnin is number of iterations to use as burnin
# Output: theta, z
sampleGibbs <- function(z, c, n.iter, init.theta, init.miss, r, a, b, burnin = 1){
  z.sum <- sum(z)
  m <- length(c)
  n <- length(z) + m
  miss.vals <- init.miss
  res <- matrix(NA, nrow = n.iter, ncol = 1 + m)
```

```
  for (i in 1:n.iter){
    var.sum <- z.sum + sum(miss.vals)
    theta <- rgamma(1, shape = a + n*r, rate = b + var.sum)
    miss.vals <- sapply(c, function(x) {sampleTrunGamma(x, r, theta)})
    res[i,] <- c(theta, miss.vals)
  }
  return(res[burnin:n.iter,])
}
# set parameter values
r <- 10
a <- 1
b <- 1
# input data
z <- c(3.4,2.9,1.4,3.2,1.8,4.6,2.8)
c <- c(1.2,1.7,2.0,1.4,0.6)
n.iter <- 200
init.theta <- 1
init.missing <- rgamma(length(c), shape = r, rate = init.theta)
# run sampler
res <- sampleGibbs(z, c, n.iter, init.theta, init.missing, r, a, b)
```

Below, we see traceplots for 200 iterations of the Gibbs sampler. It is difficult to tell whether or not the sampler has failed to converge, thus, we turn to running average plots.

```
plot(1:n.iter, res[,1], pch = 16, cex = .35,
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Traceplot of ", theta)))
```

```
missing.index <- c(3,8,9,10,12)
par(mfrow=c(2,3))
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:n.iter, res[,which(missing.index == ind)], pch = 16, cex = .35,
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Traceplot of ", .(x.lab))))
}
plot.new()
```

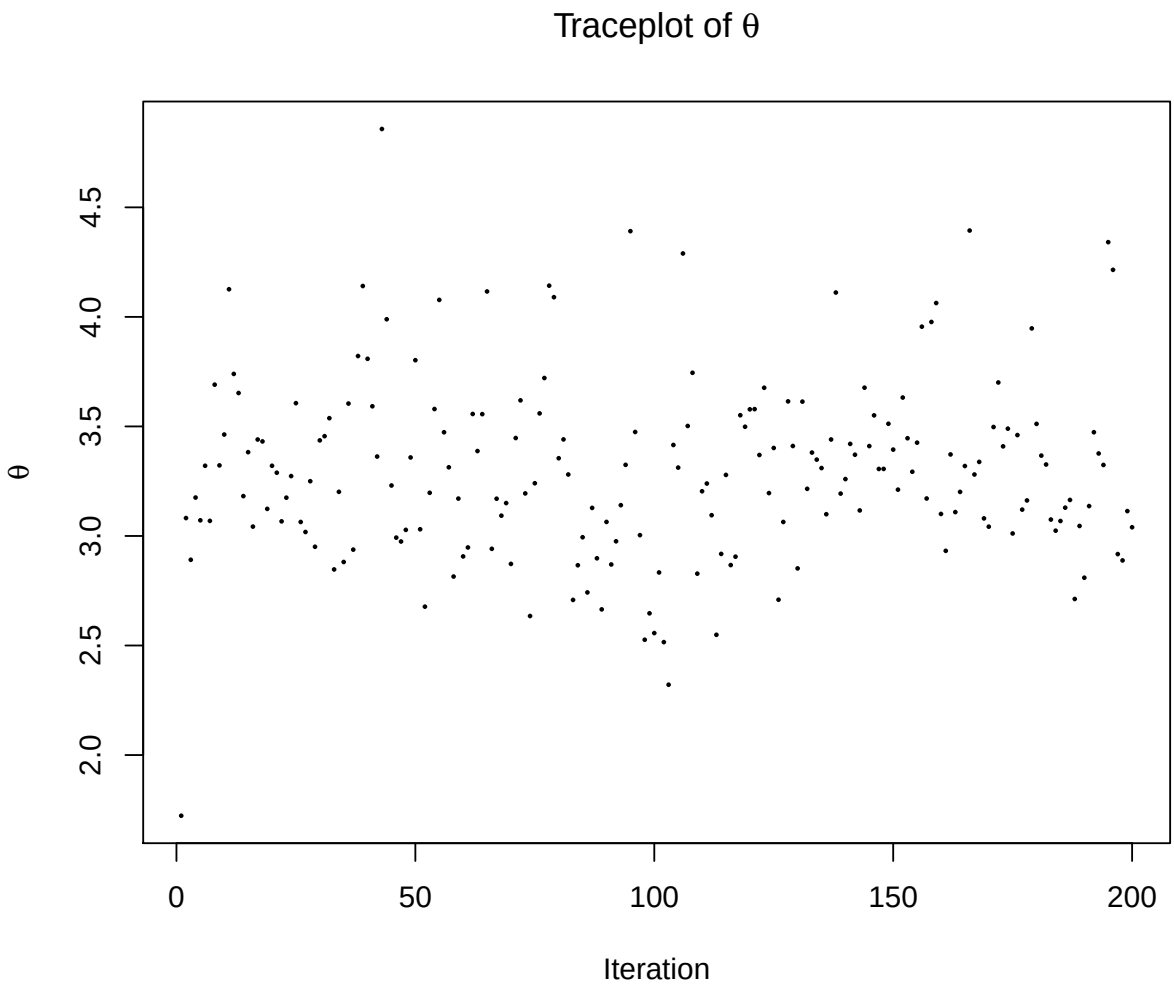
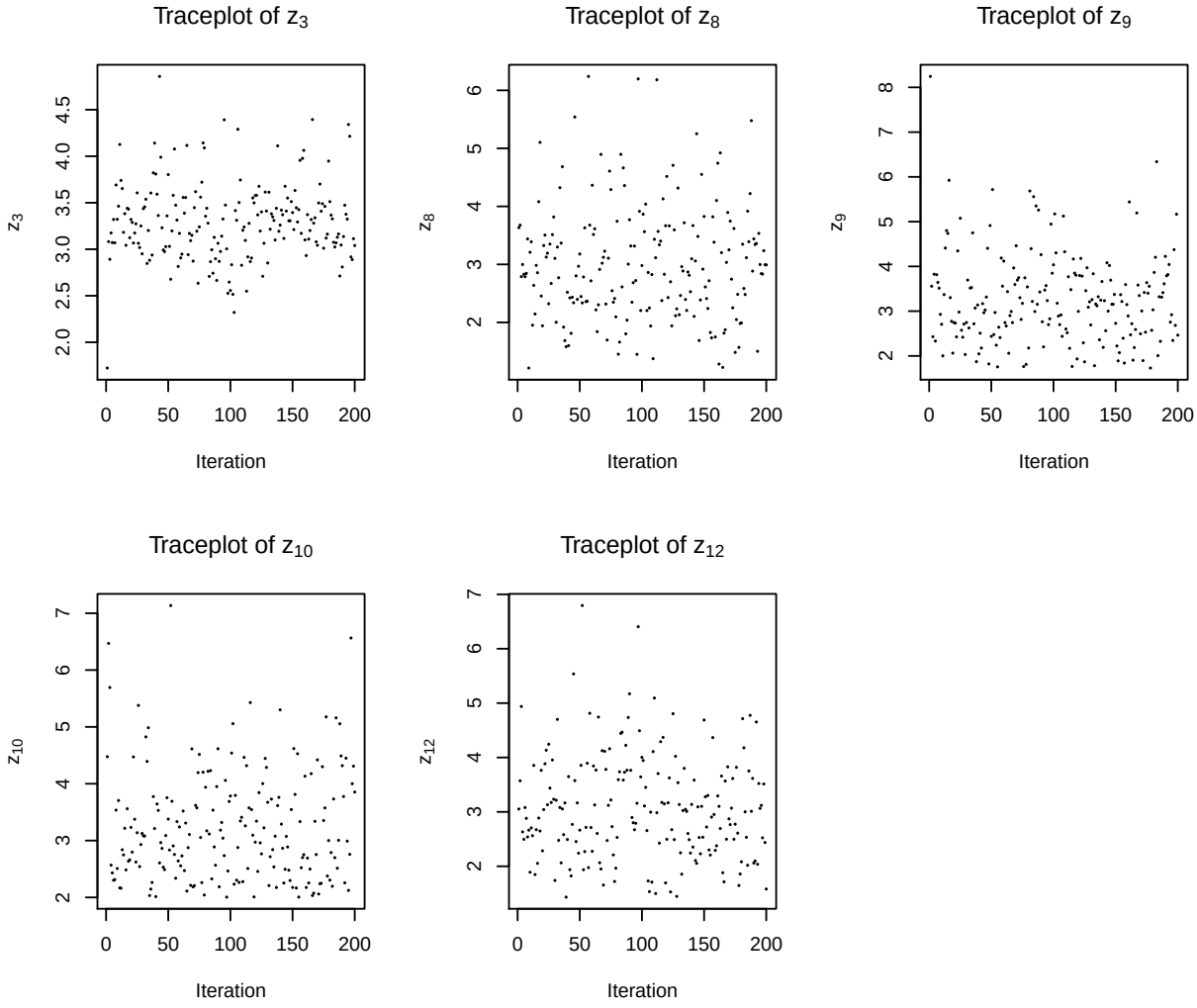


Figure 1: Traceplot of theta, 200 iterations



From these trace plots, we don't notice any particular "clumping" of data points that may suggest a problem with our sampler.

Below, we see running average plots for 200 iterations of the Gibbs sampler, where from all of these it is clear that after 200 iterations the sampler is having mixing issues, and should be run for long to check that "it has not failed to converge."

```
# get running averages
run.avg <- apply(res, 2, cumsum)/(1:n.iter)
```

```
plot(1:n.iter, run.avg[,1], type = "l",
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Running Average Plot of ", theta)))
```

```
par(mfrow=c(2,3))
missing.index <- c(3,8,9,10,12)
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:n.iter, run.avg[,which(missing.index == ind)], type = "l",
       xlab = "Iteration", ylab = x.lab,
```

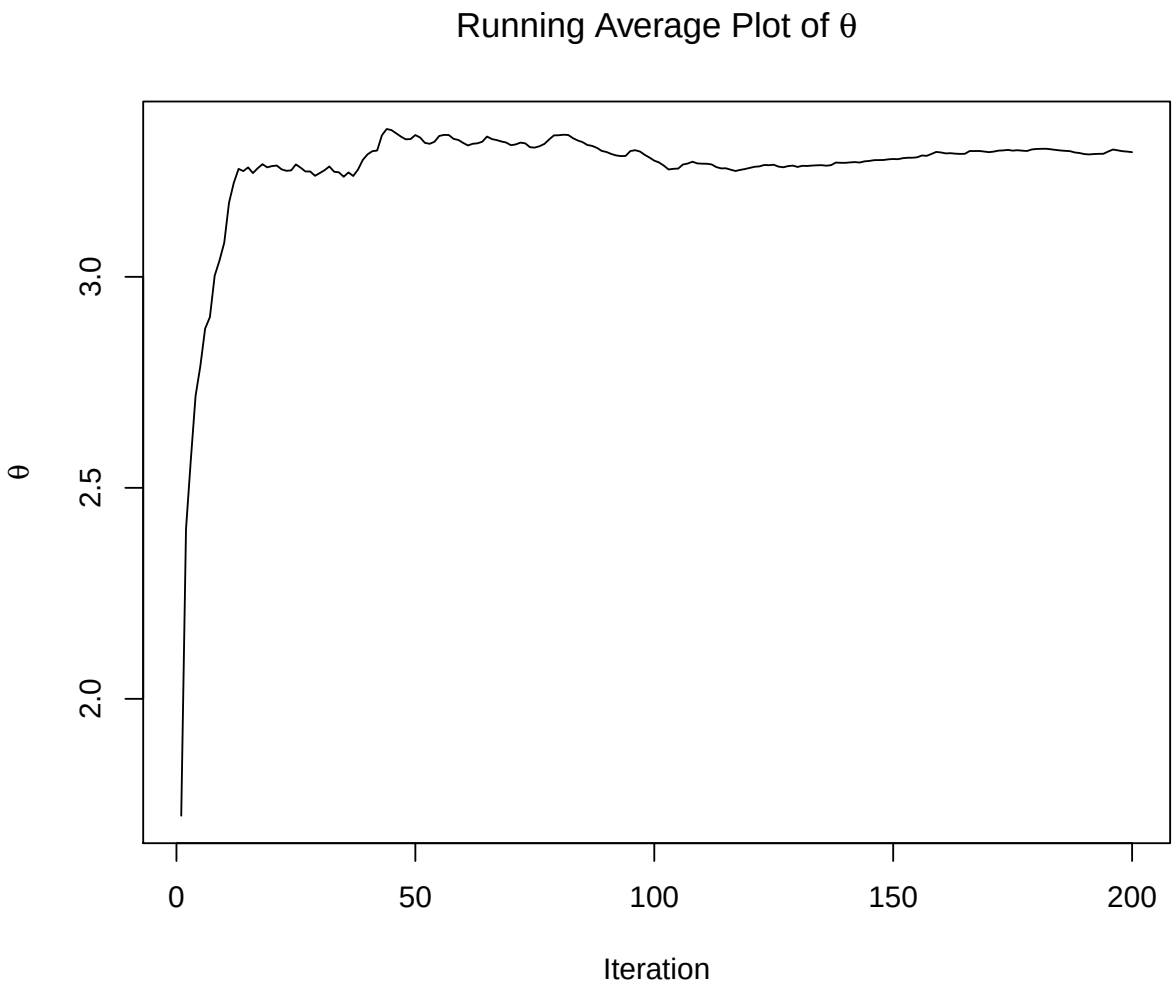
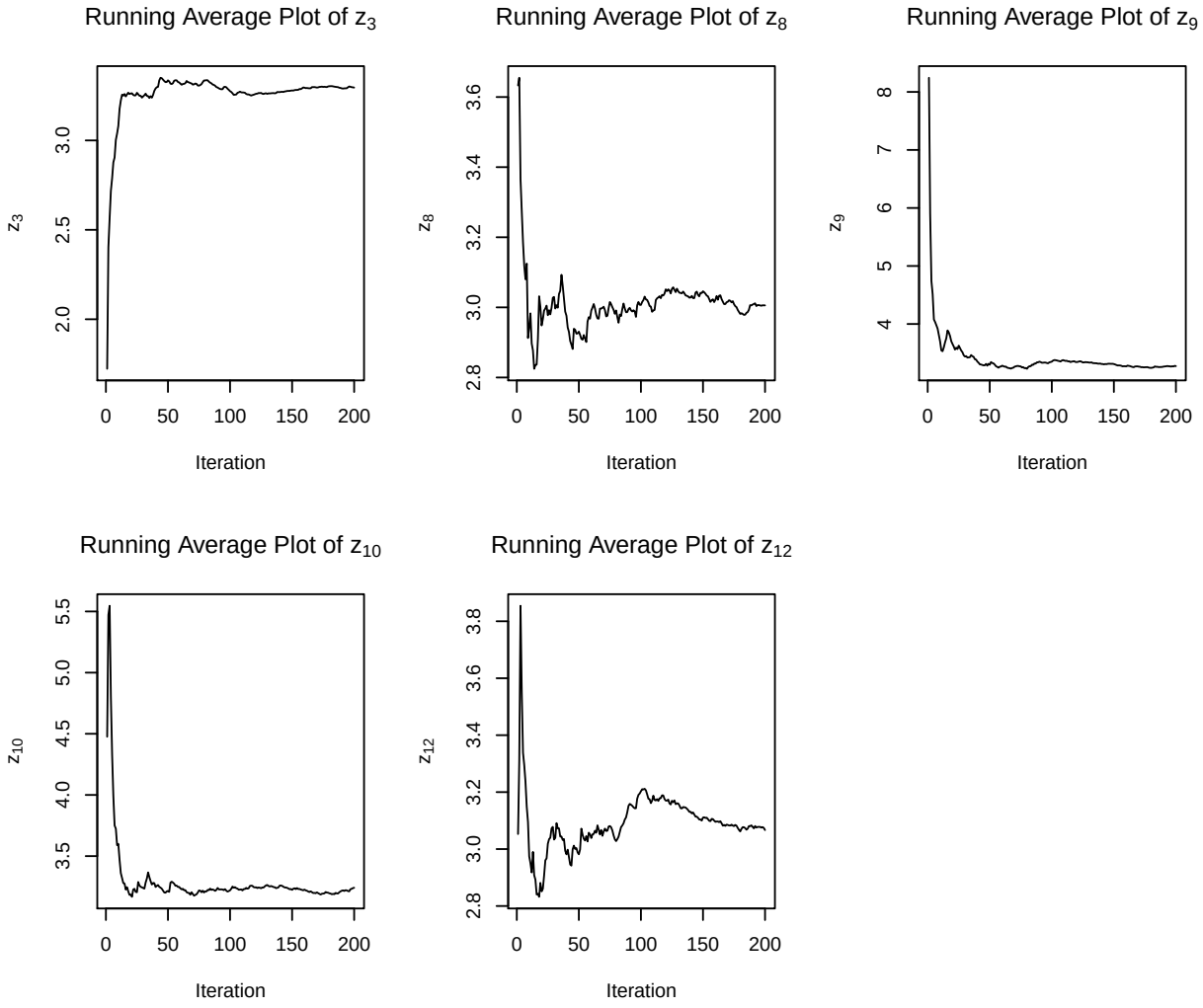


Figure 2: Running average plot of theta, 200 iterations

```
       main = bquote(paste("Running Average Plot of ", .(x.lab))))
}
plot.new()
```



From these running average plots, the plot of the running average for z_{12} may show that the average value may not have converged since it shows sharp increases in value. Also, it isn't clear if the average has settled to a particular value because of the downward trend seen at the end of the 200 iterations. Based on this plot, it does not seem that running the Gibbs sampler for 200 iterations is enough.

The below figures do not provide meaningful inference at this point since the sampler has not been run long enough.

```
# density plots
plot(density(res[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta)))
abline(v = mean(res[,1]), col = "red")
```

```
plot(density(res[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9])))
abline(v = mean(res[,4]), col = "red")
```

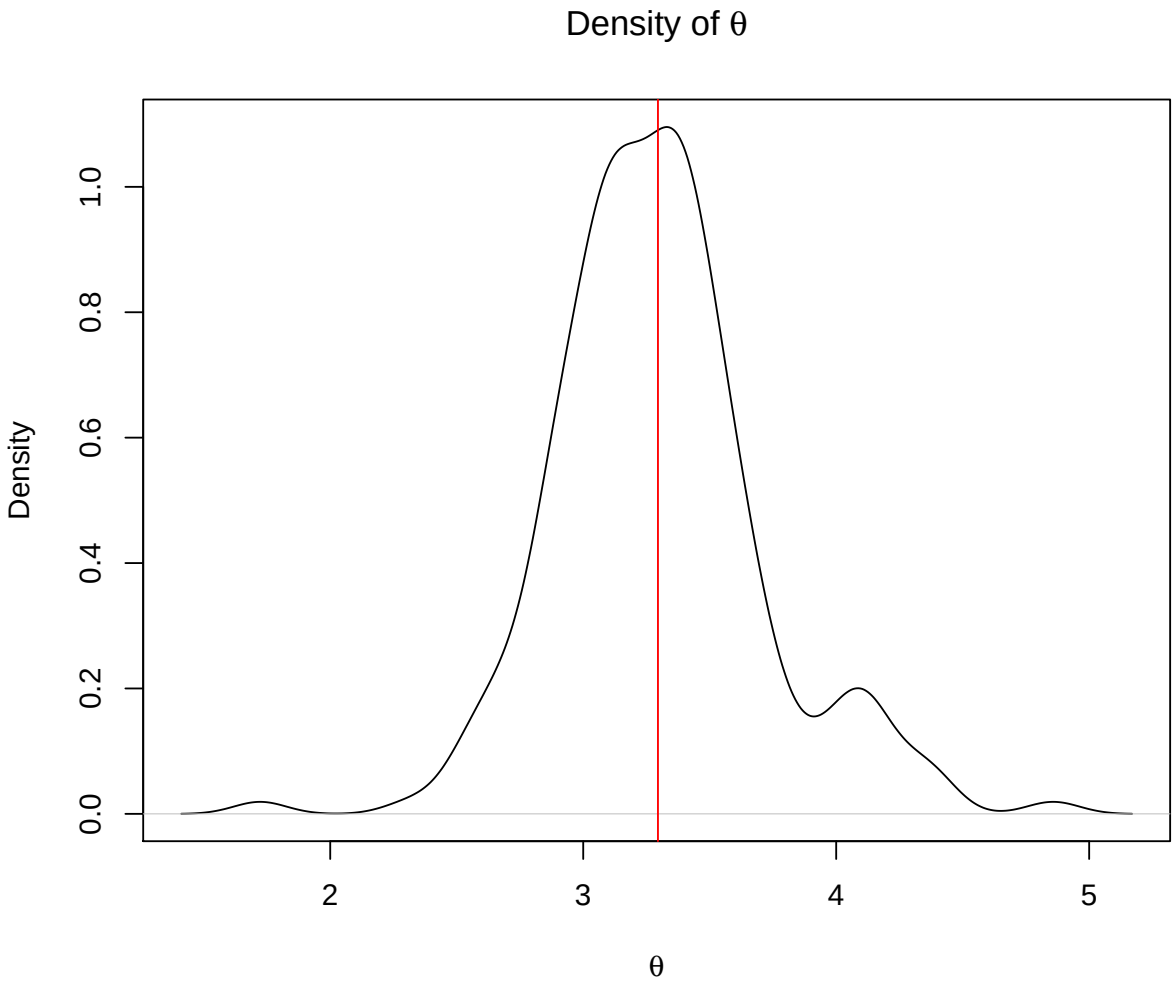


Figure 3: Estimated posterior density of theta, 200 iterations

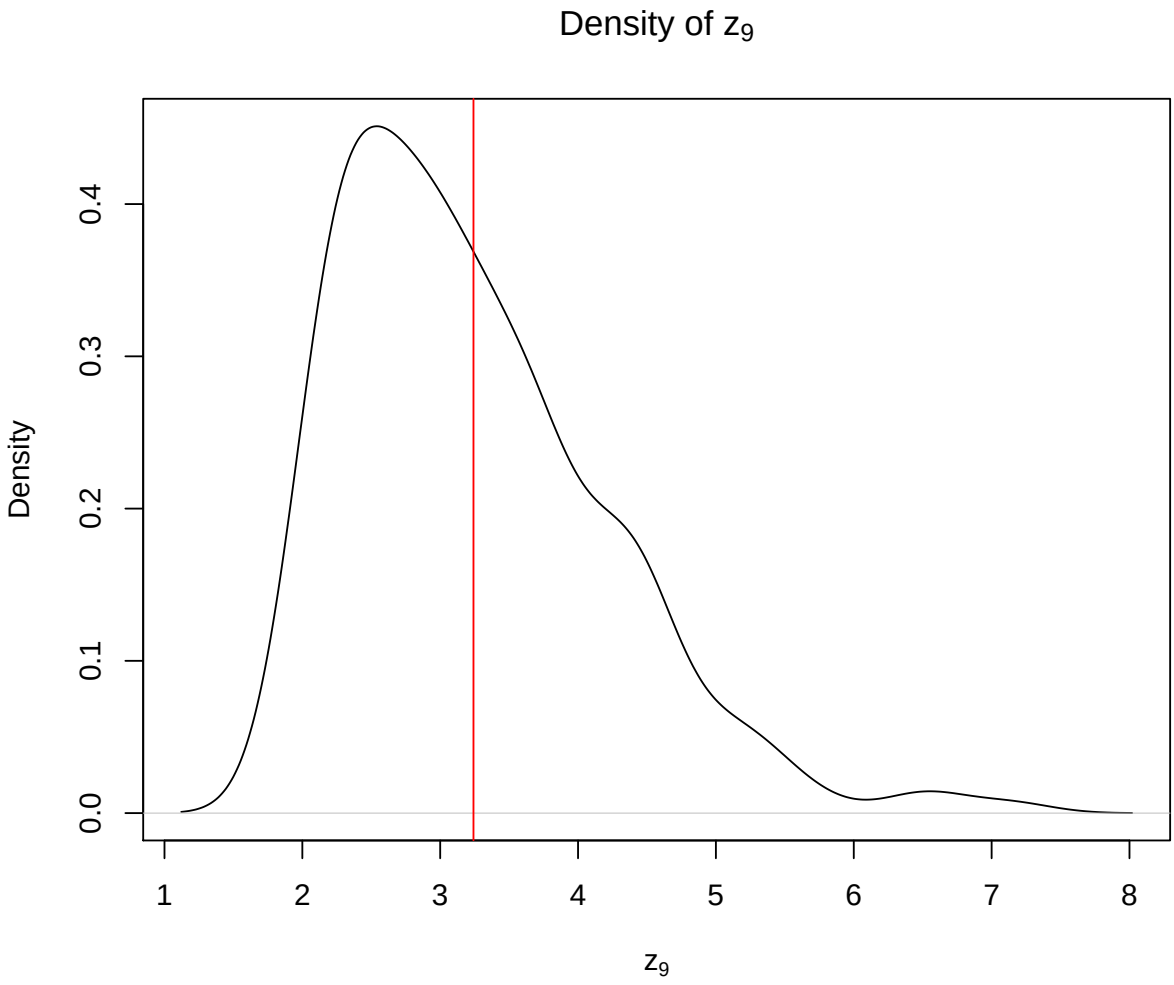


Figure 4: Estimated posterior density of z_9 (posterior mean in red), 200 iterations

2. (5 points) Now run the chain for 10,000 iterations and update your diagnostic plots (traceplots and running average plots). Report your findings for both traceplots and the running average plots for θ and the censored values. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.

```
# set number of iterations
new_iter <- 10000
# run sampler
res <- sampleGibbs(z, c, new_iter, init.theta, init.missing, r, a, b)
```

Below, we see traceplots for 10000 iterations of the Gibbs sampler. There are no noticeable "clumping" of data points that would suggest a problem with the sampler.

```
plot(1:new_iter, res[,1], pch = 16, cex = .35,
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Traceplot of ", theta)))
```

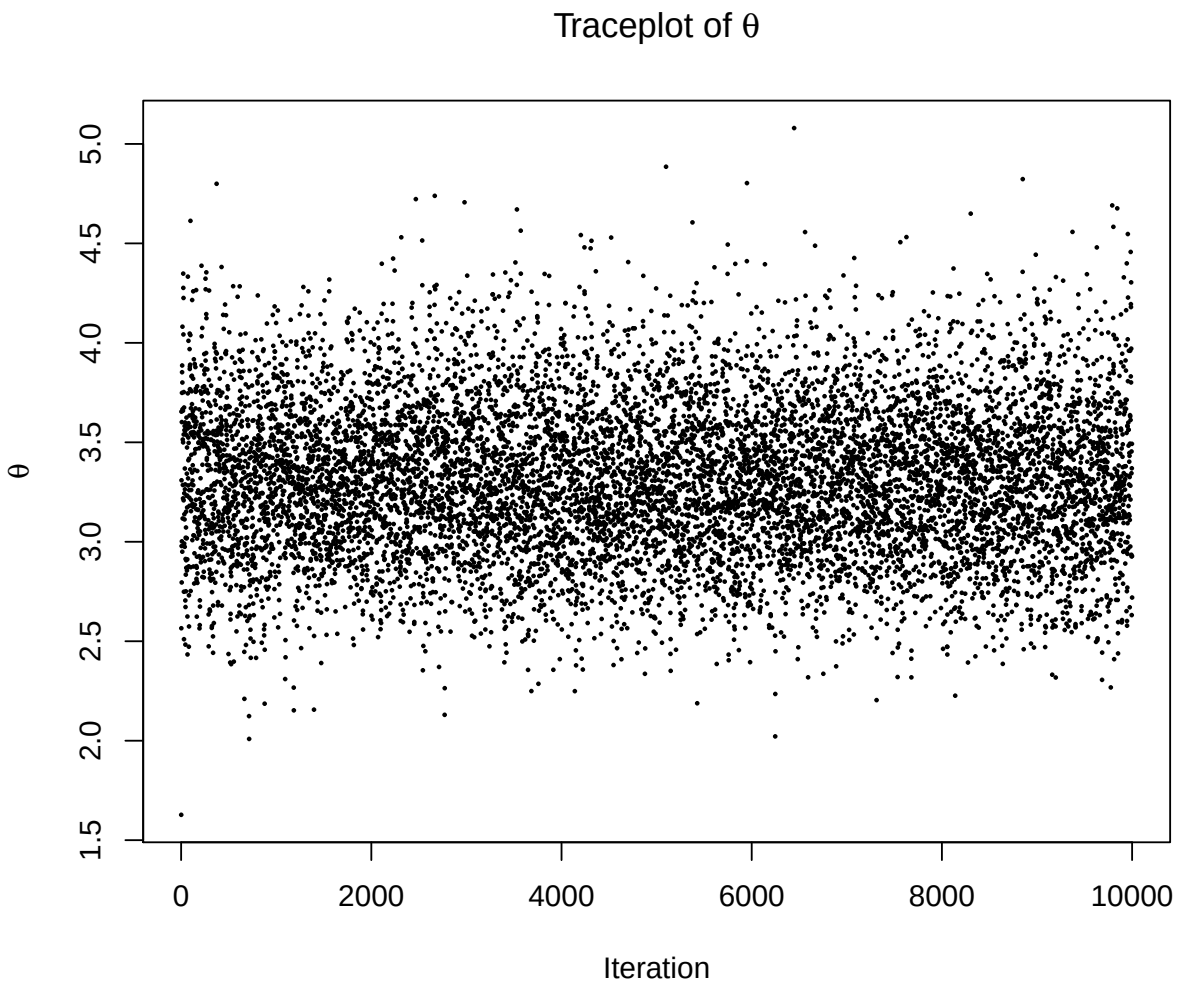


Figure 5: Traceplot of theta, 10000 iterations

```
missing.index <- c(3,8,9,10,12)
par(mfrow=c(2,3))
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:new_iter, res[,which(missing.index == ind)], pch = 16, cex = .35,
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Traceplot of ", .(x.lab))))
}
plot.new()
```

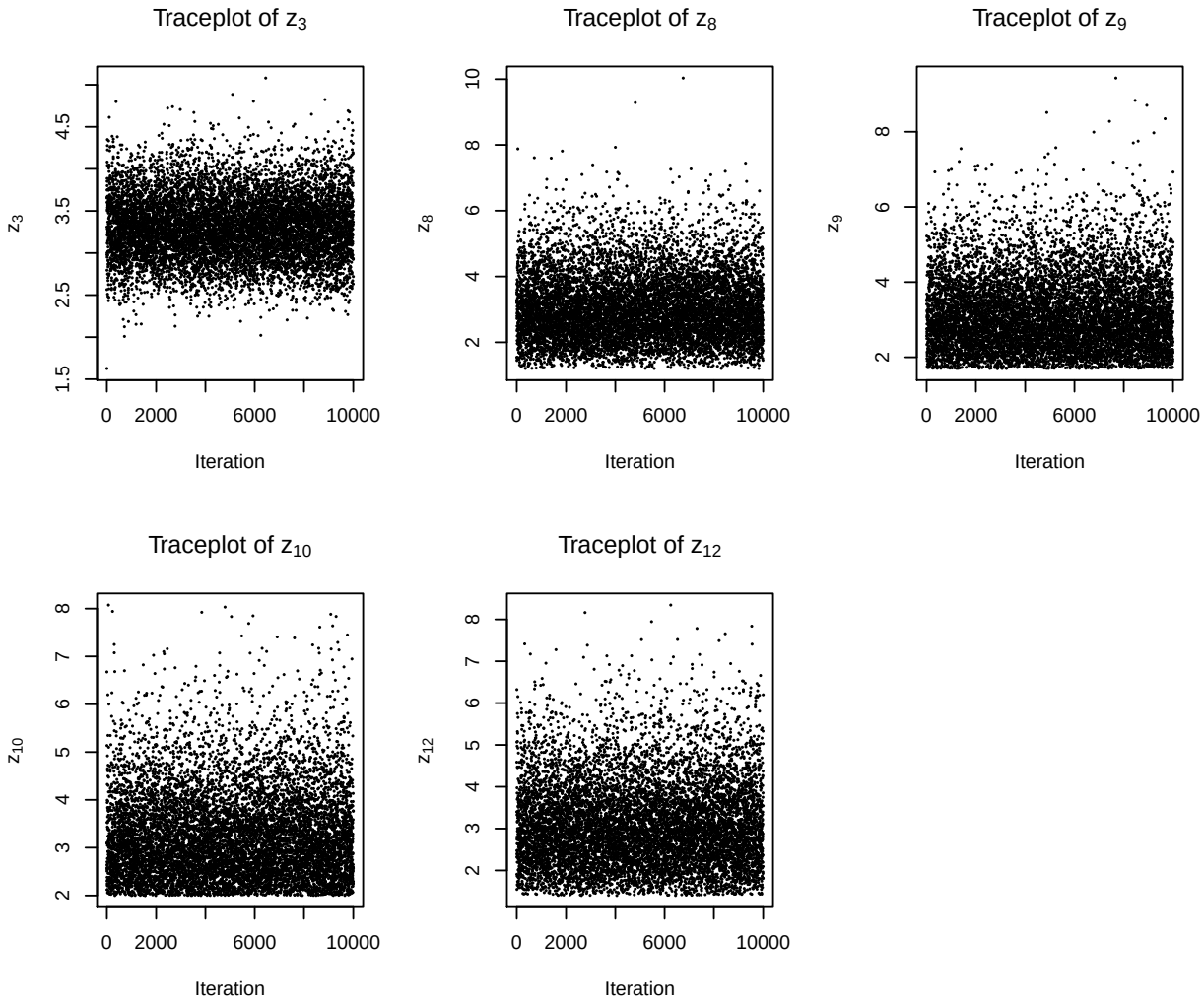


Figure 6: Traceplot of $z_3, z_8, z_9, z_{10}, z_{12}$., 10000 iterations

Below, we see running average plots for 10000 iterations of the Gibbs sampler, where from all of these it is clear that after 10000 iterations the sampler is having mixing issues, and should be run for long to check that "it has not failed to converge."

```
# get running averages
run.avg <- apply(res, 2, cumsum)/(1:new_iter)
```

```r
plot(1:new_iter, run.avg[,1], type = "l",
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Running Average Plot of ", theta)))
```

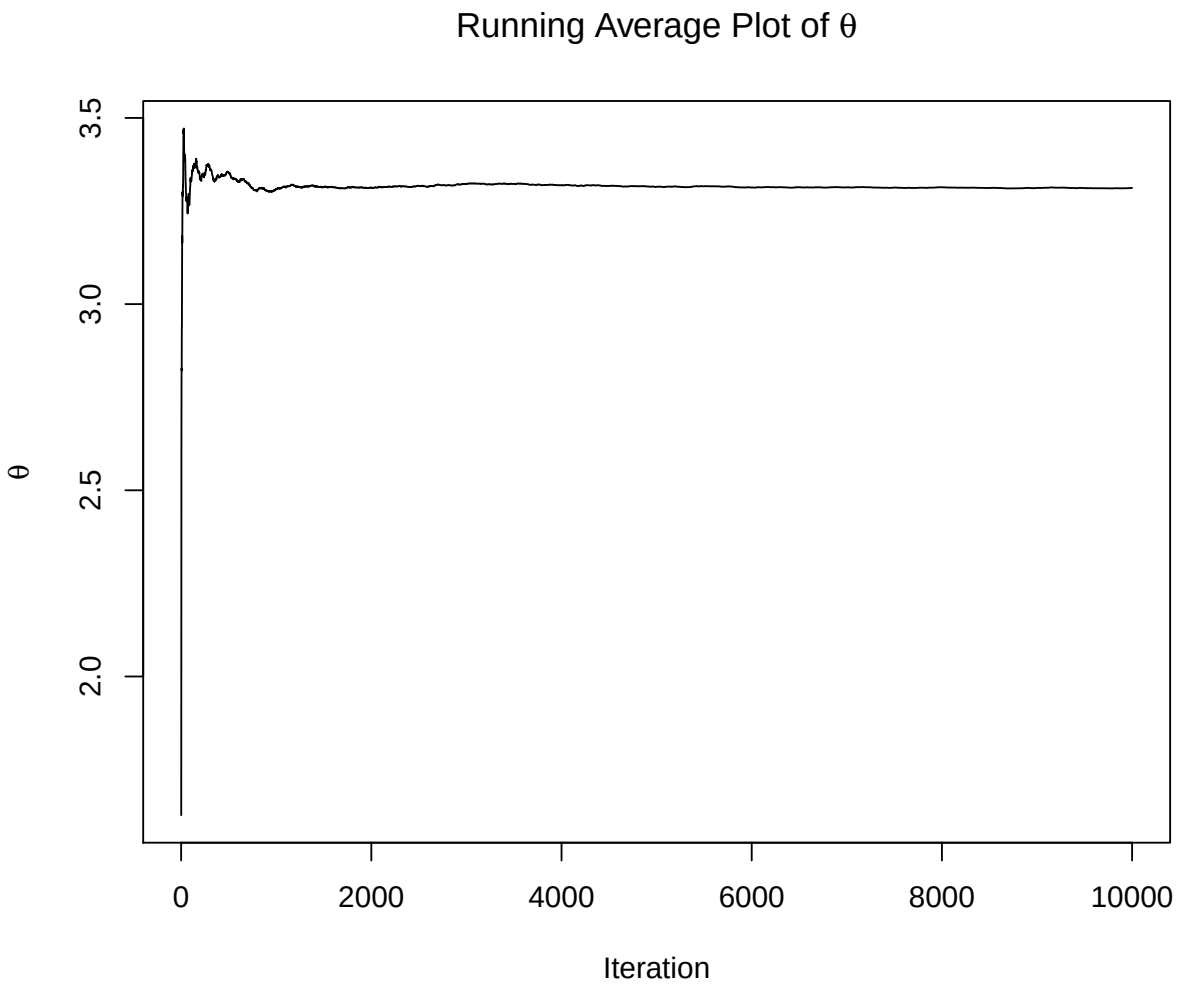
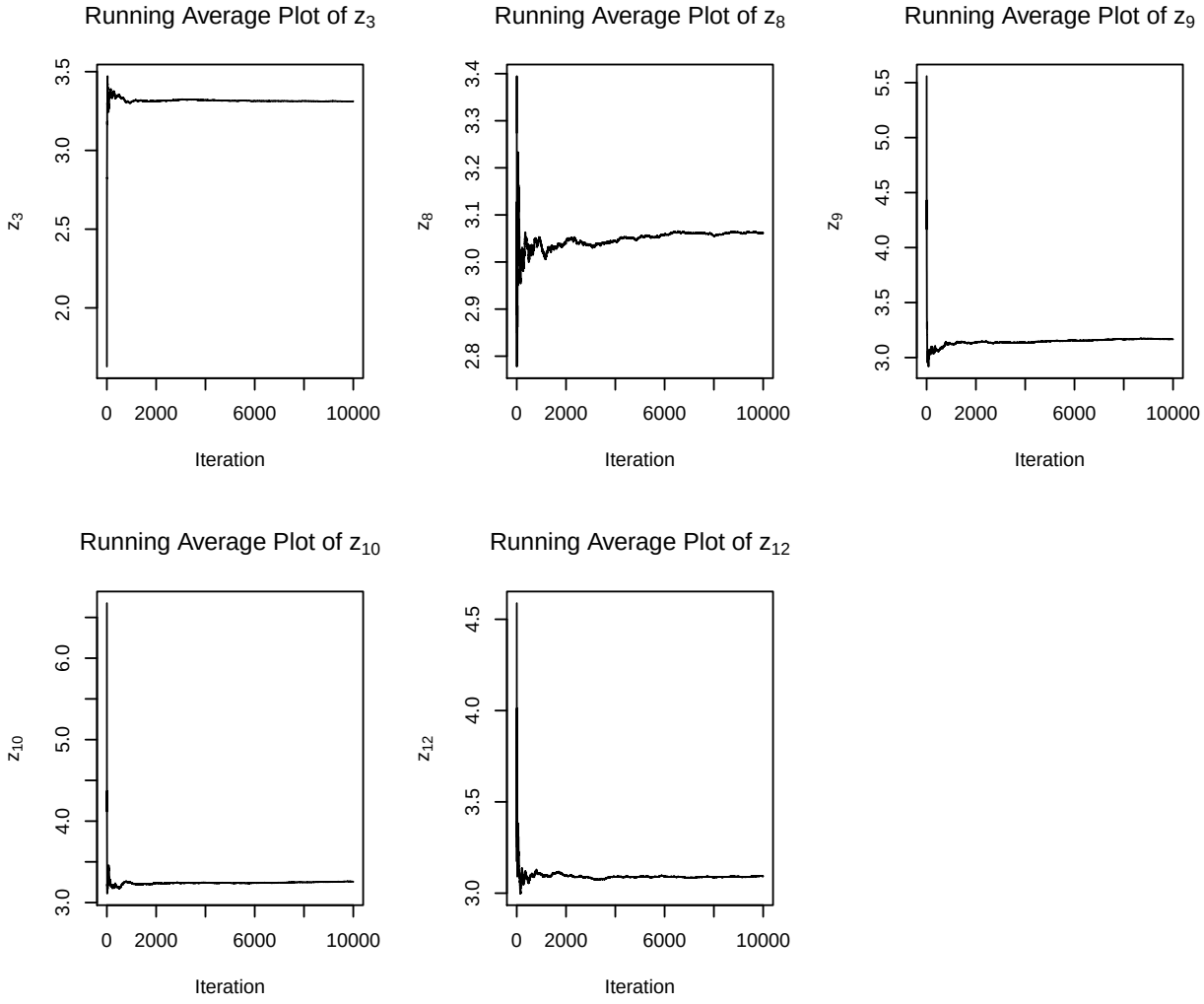


Figure 7: Running average plot of theta, 10000 iterations

```r
par(mfrow=c(2,3))
missing.index <- c(3,8,9,10,12)
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:new_iter, run.avg[,which(missing.index == ind)], type = "l",
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Running Average Plot of ", .(x.lab))))
}
plot.new()
```



In all of the running average plots, the average values seem to have converged to a single value by 10000 iterations Since the traceplots didn't show any problems and the running average plots didn't either, we can conclude that the Gibbs sampler was run for enough iterations.

3. (5 points) Give plots of the estimated density of $\theta \mid \cdots$ and $z_9 \mid \cdots$. Be sure to give brief explanations of your results and findings. (Present plots for 10,000 iterations).

Based on the density of θ shown below, we can see that our parameter of interest, the rate parameter for the lifetime distribution, most likely has a value of about 3.3, which is the mean of the distribution and is close to the maximum density value. For a range of possible values for θ, we can use the range (2.5, 4.3) since the majority of the density for θ lies in that range.

Based on the density of z_9, the censored value for the lifetime of patient #9, we can see that the value of z_9 most likely lies in the range (2,5), where the majority of the density of patient #9 is. Since the distribution is skewed to the right, z_9 most likely has a value of about 2.7, where the maximum density value is.

```
# density plots
plot(density(res[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta)))
abline(v = mean(res[,1]), col = "red")
```

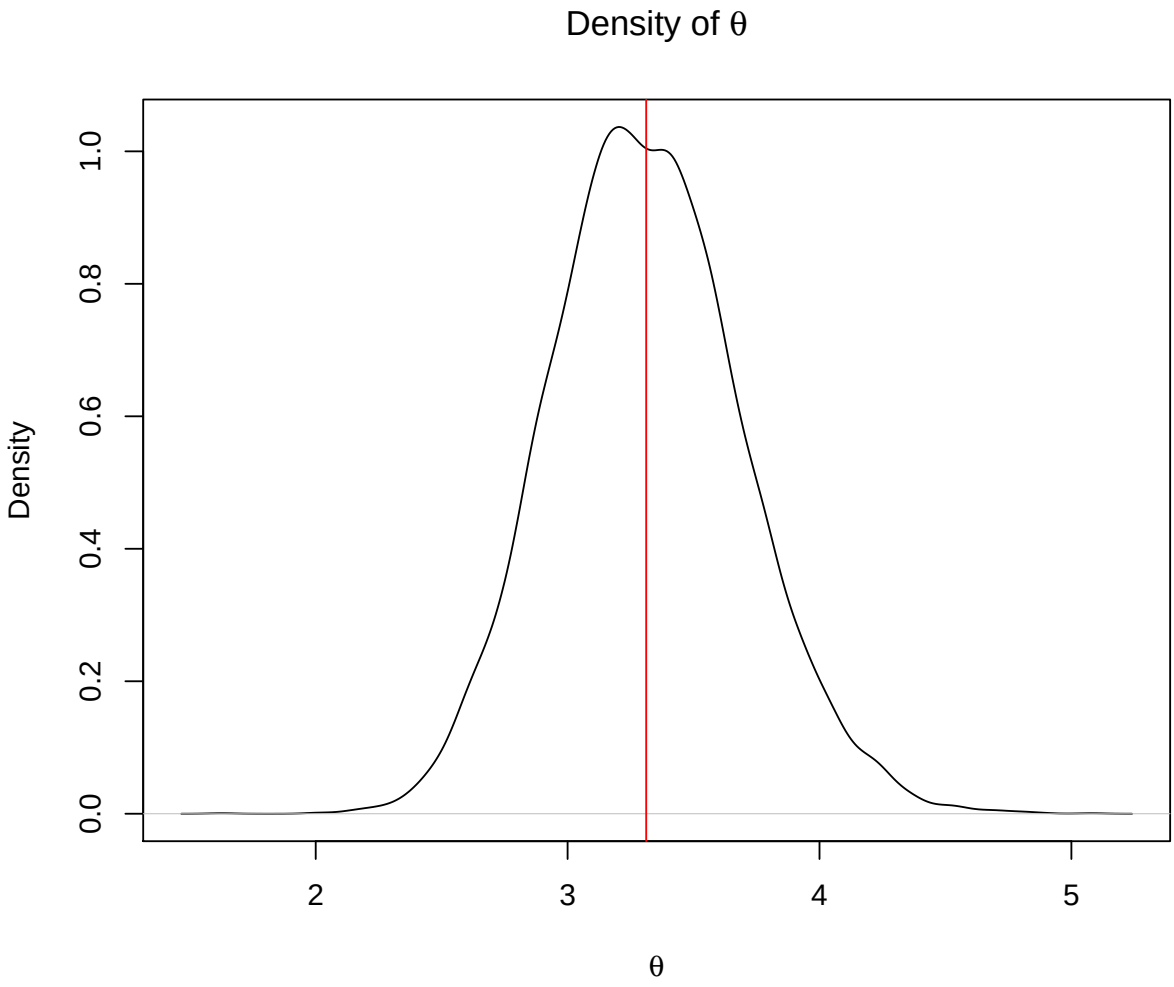


Figure 8: Estimated posterior density of theta, 10000 iterations

```r
plot(density(res[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9])))
abline(v = mean(res[,4]), col = "red")
```

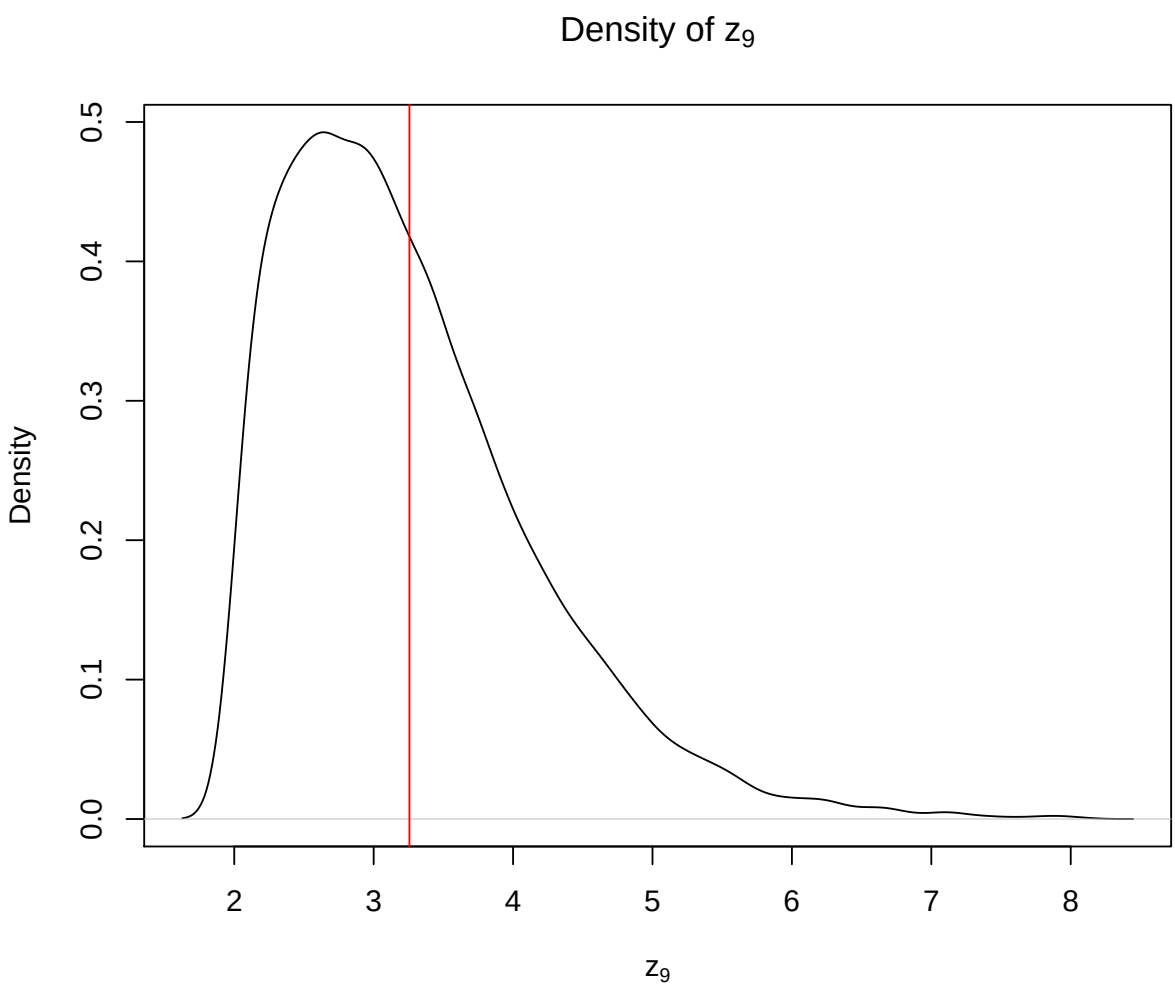
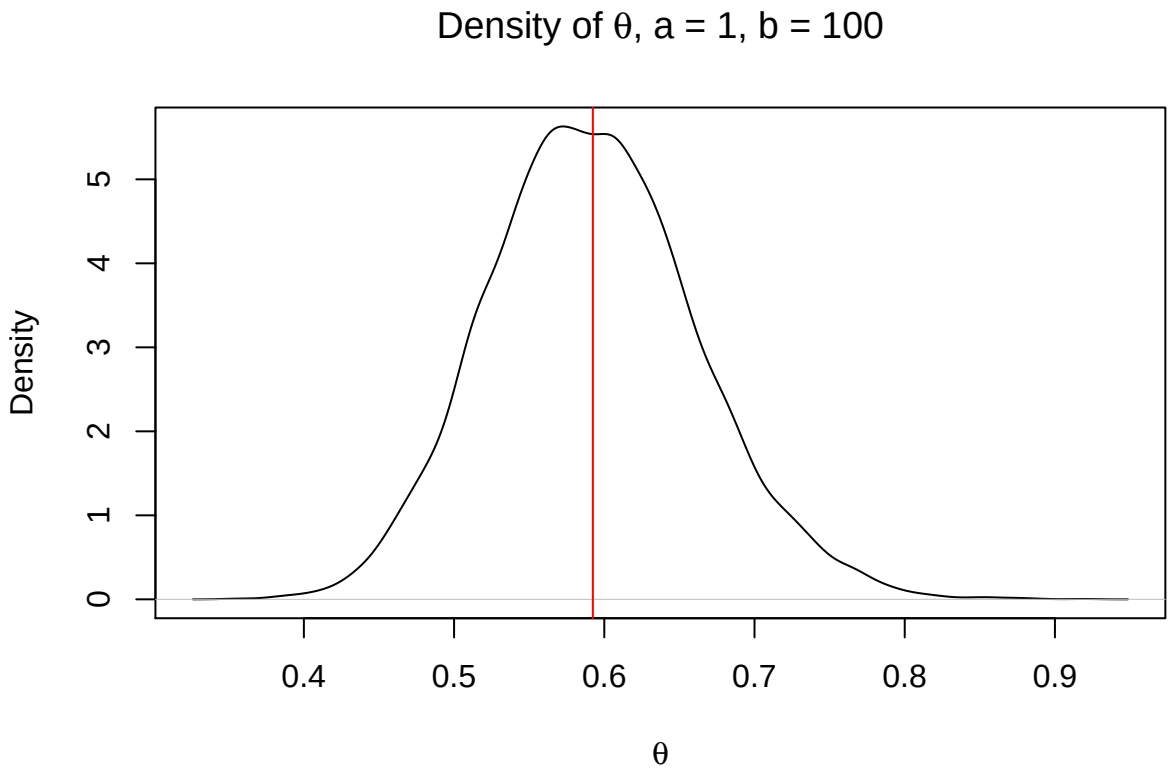


Figure 9: Estimated posterior density of z_9 (posterior mean in red), 10000 iterations

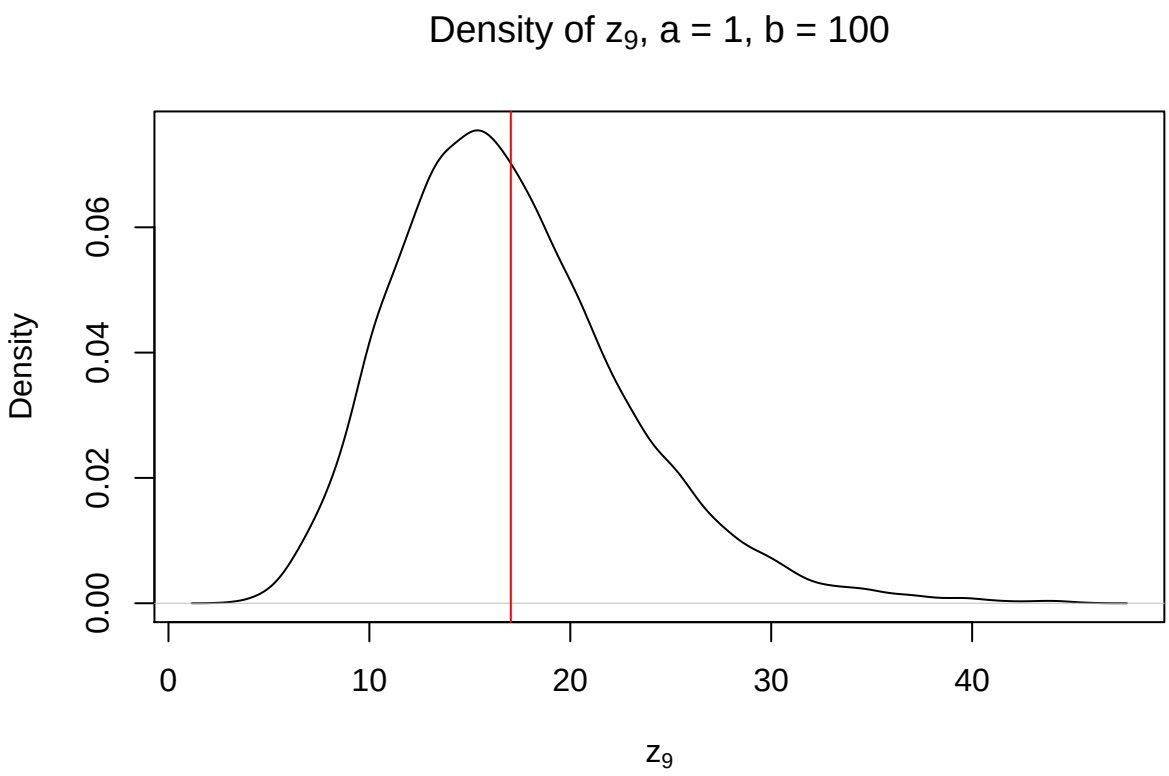
4. (5 points) Finally, let's suppose that $r = 10, a = 1, b = 100$. Do the posterior densities in part (c) change for $\theta \mid \cdots$ and $z_9 \mid \cdots$? Do the associated posterior densities change when $r = 10, a = 100, b = 1$? Please provide plots and an explanation to back up your answer. (Use 10,000 iterations for the Gibbs sampler).

```r
# Set seed for reproducibility
set.seed(123)
# set parameter values
r <- 10
a <- 1
b <- 100
# run sampler
```

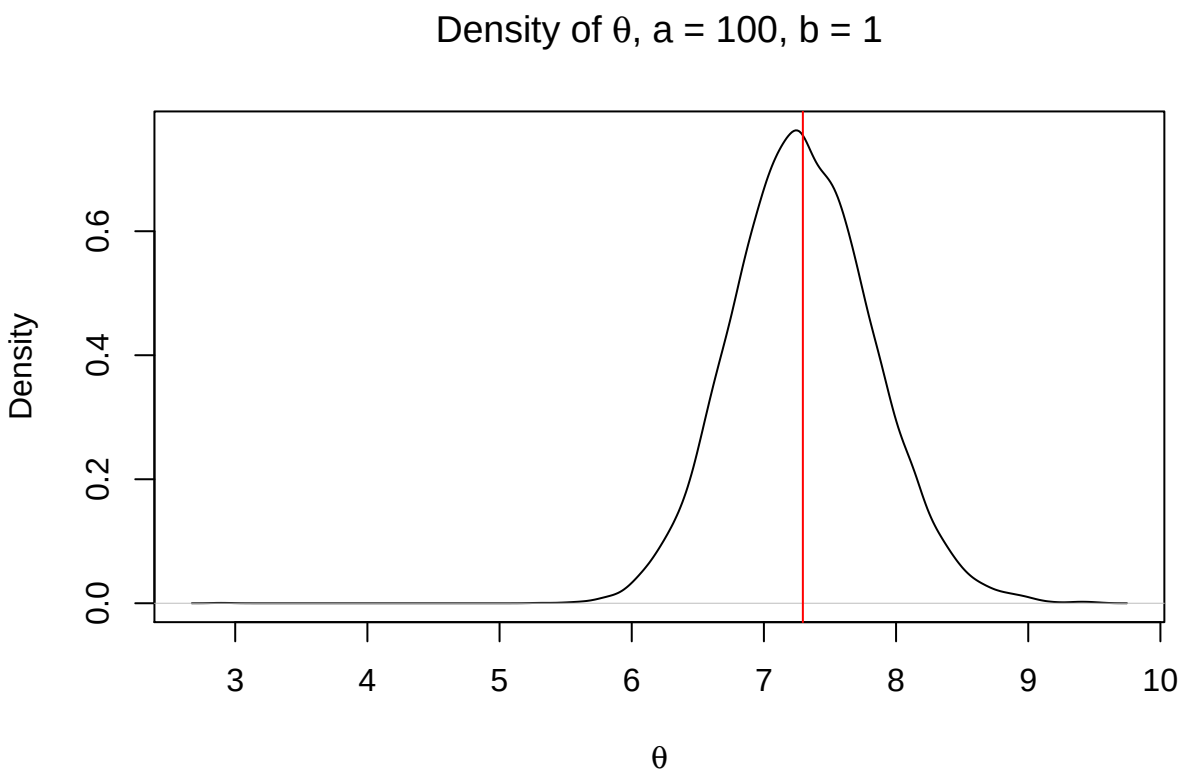
```r
res <- sampleGibbs(z, c, new_iter, init.theta, init.missing, r, a, b)
# density plots
plot(density(res[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta, ", a = 1, b = 100")))
abline(v = mean(res[,1]), col = "red")
```



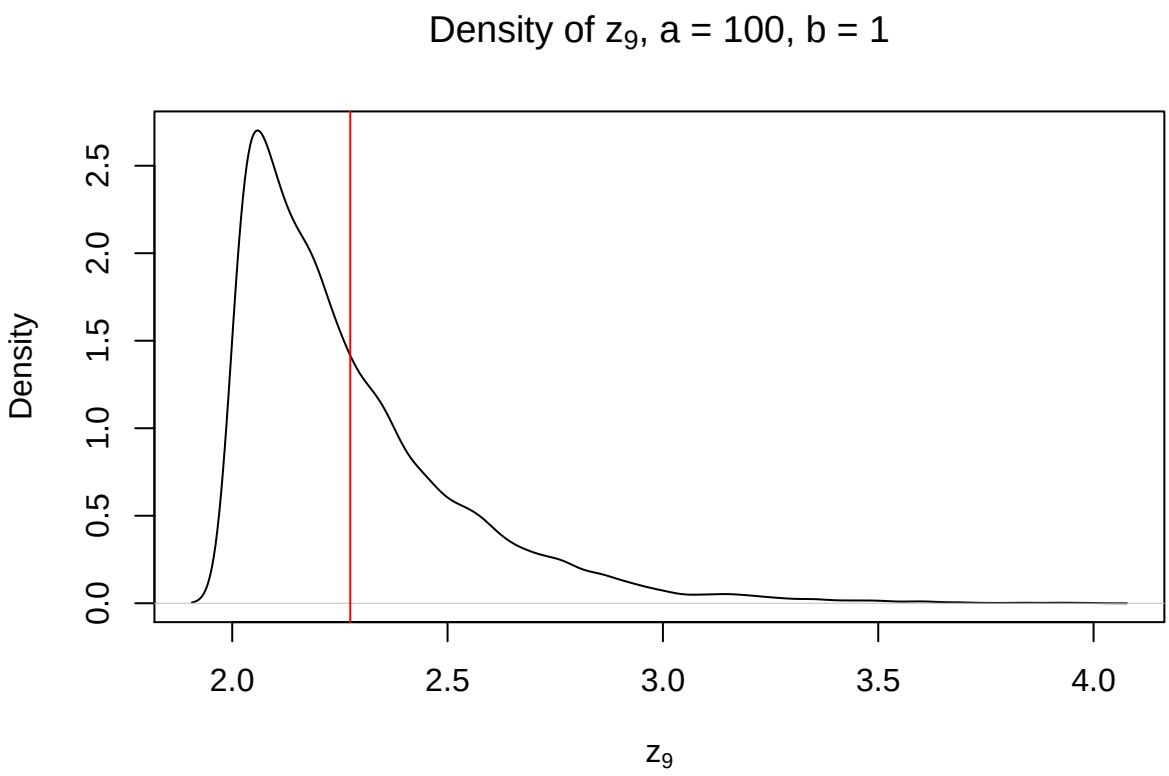

```r
plot(density(res[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9],  ", a = 1, b = 100")))
abline(v = mean(res[,4]), col = "red")
```




```
# Set seed for reproducibility
set.seed(123)
# set parameter values
r <- 10
a <- 100
b <- 1
# run sampler
res <- sampleGibbs(z, c, new_iter, init.theta, init.missing, r, a, b)
# density plots
plot(density(res[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta, ", a = 100, b = 1")))
abline(v = mean(res[,1]), col = "red")
```




```
plot(density(res[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9], ", a = 100, b = 1")))
abline(v = mean(res[,4]), col = "red")
```



The posterior densities for $\theta \mid \cdots$ and $z_9 \mid \cdots$ do change from part (c), showing that these densities are sensitive to large changes in the prior parameters for a and b. Both the mean and variance of these densities change drastically. When $r = 1, a = 1, b = 100$, the density of θ maintains the same shape, but most of the distribution now lies in the range (0.4, 0.8) instead of (2.5, 4.3) while the mean shifts from 3.3 to 0.6. When $r = 1, a = 100, b = 1$, the density of θ, again, maintains the same shape, but most of the distribution now lies in the range (6, 9). When $r = 1, a = 1, b = 100$, the density of z_9 becomes less skewed to the right than in part (c), and most of the distribution now lies in the range (5, 30) instead of (2,5). The mean also shifts greatly, changing from about 3.3 to about 17. When $r = 1, a = 100, b = 1$, the density of z_9 becomes more skewed to the right, and most of the distribution now lies in the range (2, 3).