Module 8: Part II: Gibbs Sampling with an Application to Missing Data

Rebecca C. Steorts

Tuesday, March 19, 2019

Agenda

- Gibbs sampling (Three-stage sampler)
- Gibbs sampling (multi-stage sampler)
- Gibbs sampling with an application to missing data (censoring)

Multi-stage Gibbs sampler

Assume three random variables, with joint pmf or pdf: $p(x, y, z)$..

Set x, y, and z to some values (x_o, y_o, z_o).

Sample $x|y, z$, then $y|x, z$, then $z|x, y$, then $x|y, z$, and so on.
More precisely,

0. Set (x_0, y_0, z_0) to some starting value.
1. Sample $x_1 \sim p(x|y_0, z_0)$.
   Sample $y_1 \sim p(y|x_1, z_0)$.
   Sample $z_1 \sim p(z|x_1, y_1)$.
2. Sample $x_2 \sim p(x|y_1, z_1)$.
   Sample $y_2 \sim p(y|x_2, z_1)$.
   Sample $z_2 \sim p(z|x_2, y_2)$.
   $\vdots$

Multi-stage Gibbs sampler

Assume d random variables, with joint pmf or pdf $p(v^1, \ldots, v^d)$.

At each iteration $(1, \ldots, M)$ of the algorithm, we sample from

$$
\begin{aligned}
v^1 &\mid v^2, v^3, \ldots, v^d \\
v^2 &\mid v^1, v^3, \ldots, v^d \\
&\vdots \\
v^d &\mid v^1, v^2, \ldots, v^{d-1}
\end{aligned}
$$

always using the most recent values of all the other variables.

The conditional distribution of a variable given all of the others is referred to as the *full conditional* in this context, and for brevity denoted $v^i \mid \cdots$.

Example: Censored data

In many real-world data sets, some of the data is either missing altogether or is partially obscured.

One way in which data can be partially obscured is by *censoring*, which means that we know a data point lies in some particular interval, but we do not observe it.

Medical data censoring

Suppose 6 patients participate in a cancer trial, however, patients 1, 2 and 4 leave the trial early.

Then we know when they leave the study, but we don't know information about them as the trial continues.

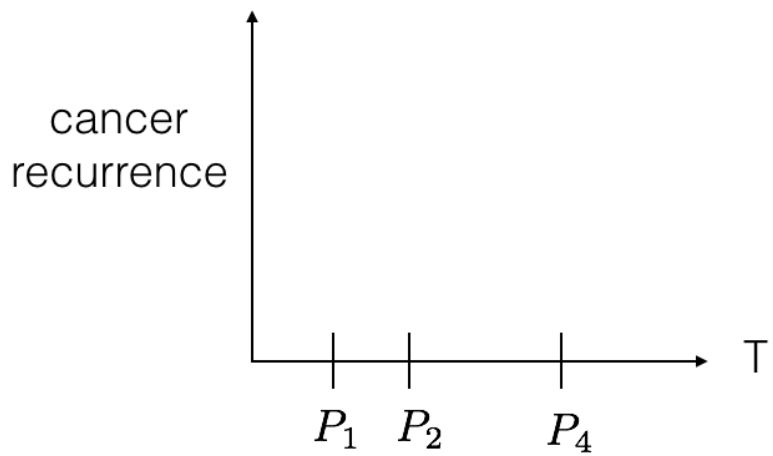


Figure 1: Example of censoring for medical data.

This is a certain type of missing data.

Heart Disease (Censoring) Example

- Researchers are studying the length of life (lifetime) following a particular medical intervention, such as a new surgical treatment for heart disease.
- The study consists of 12 patients.
- 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.

Model

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

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

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

where a, b, and r are known.

- 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.

Posterior inference

Goal: find $p(\theta, z_{1:n}|x_{1:n})$?

1. Straightforward approaches that are in closed form do not work (think about these on your own). Instead we turn to Gibbs!

2. To sample from $p(\theta, z_{1:n}|x_{1:n})$, we cycle through each of the full conditional distributions,

$$
\begin{aligned}
\theta \mid z_{1:n}, x_{1:n} \\
z_1 \mid \theta, z_{2:n}, x_{1:n} \\
z_2 \mid \theta, z_1, z_{3:n}, x_{1:n} \\
\vdots \\
z_n \mid \theta, z_{1:n-1}, x_{1:n}
\end{aligned}
$$

sampling from each in turn, always conditioning on the most recent values of the other variables.

Likelihood

Recall the model is:

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

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

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

The pdf associated with this random variable is rather strange, as it consists of two point masses: one at Z_i and one at c_i. The formula is

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

.

Full conditionals

Recall the model is:

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

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

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

Let's start with $\theta | \cdots$

- Since $\theta \perp x_{1:n} \mid z_{1:n}$ (i.e., θ is conditionally independent of $x_{1:n}$ given $z_{1:n}$),

$$p(\theta | \cdots) = p(\theta | z_{1:n}, x_{1:n}) = p(\theta | z_{1:n}) \tag{4}$$

$$= \text{Gamma}\left(\theta \,\middle|\, a + nr,\ b + \textstyle\sum_{i=1}^{n} z_i\right) \tag{5}$$

  using the fact that the prior on θ is conjugate.

Full conditionals

Now we can easily find the full conditionals.

- Note 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

$$
\begin{aligned}
p(z_i|z_{-i}, x_{1:n}, \theta) &= p(z_i|x_i, \theta) && (6)\\
&\underset{z_i}{\propto} p(z_i, x_i, \theta) && (7)\\
&= p(\theta)p(z_i|\theta)p(x_i|z_i, \theta) && (8)\\
&\underset{z_i}{\propto} p(z_i|\theta)p(x_i|z_i, \theta) && (9)\\
&= p(z_i|\theta)p(x_i|z_i), && (10)
\end{aligned}
$$

where the last line holds by conditional independence.

Full conditionals (continued)

Recall that

$$p(z_i|z_{-i}, x_{1:n}, \theta) \propto p(z_i|\theta)p(x_i|z_i).$$

There are now two cases to consider.

1. 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.

2. 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.

Sampling from the truncated Gamma

We sample from the truncated gamma using a modified version of the inverse CDF method.

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

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

Therefore Y is a sample from the truncated Gamma.

Remark: when we implement the GS, we do not sample the observed values. We impute the censored values using the method just outlined.

Heart Disease (Censoring) Application

Let's recall the motivating application behind our model, which a study on 12 patients that have heart disease.

Recall that 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.

Application to censored data

```
# input the data
# z's are the observed values
z <- c(3.4,2.9,1.4,3.2,1.8,4.6,2.8)
# c's are the censored values
c <- c(1.2,1.7,2.0,1.4,0.6)
```

Application to censored data (continued)

```r
knitr::opts_chunk$set(cache=TRUE)
# 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){
  p0 <- pgamma(t, shape = a, rate = b)
  # Use the modification of the inverse CD method
  x <- runif(1, min = p0, max = 1)
  y <- qgamma(x, shape = a, rate = b)
  return(y)
}
```

Application to censored data (continued)

```
# Gibbs sampler
# z is the fully observed data
# c is the censored data
# n.iter is number of iterations
# init.theta and init.miss are initial values for sampler
# r,a, and b are fixed parameters
# burnin is number of iterations to use as burnin
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,])
}
```

Initialize Unknown Parameters

```
set.seed(5983)
# Z follows a Gamma(shape = r, scale= \theta)
# the scale is slightly more peaked
r <- 10
# theta follows a Gamma(1,1)
# Setting a,b both to one is a default prior on theta
a <- 1
b <- 1
n.iter <- 20
init.theta <- 1
init.missing <-
  rgamma(length(c), shape = r, rate = init.theta)
```

Run Gibbs sampler

```
res <- sampleGibbs(z, c, n.iter, init.theta, init.missing,
                   r, a, b)
```

Output of the Gibbs sampler

```
dim(res)
```

```
## [1] 20  6
```

```
head(res)
```

```
##           [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## [1,] 1.705932 8.067377 4.770327 3.845943 6.220137 9.527103
## [2,] 2.086170 5.059026 8.347300 4.648905 3.556692 2.680102
## [3,] 2.761540 1.383840 3.185447 3.943279 3.551234 4.483285
## [4,] 2.937102 4.766874 3.323249 4.078033 2.467942 4.177714
## [5,] 3.507482 2.467106 2.955495 3.333004 2.940252 3.399843
## [6,] 3.358133 4.017697 3.270878 2.803783 1.739449 1.922901
```

Output of the Gibbs sampler

```
tail(res)
```

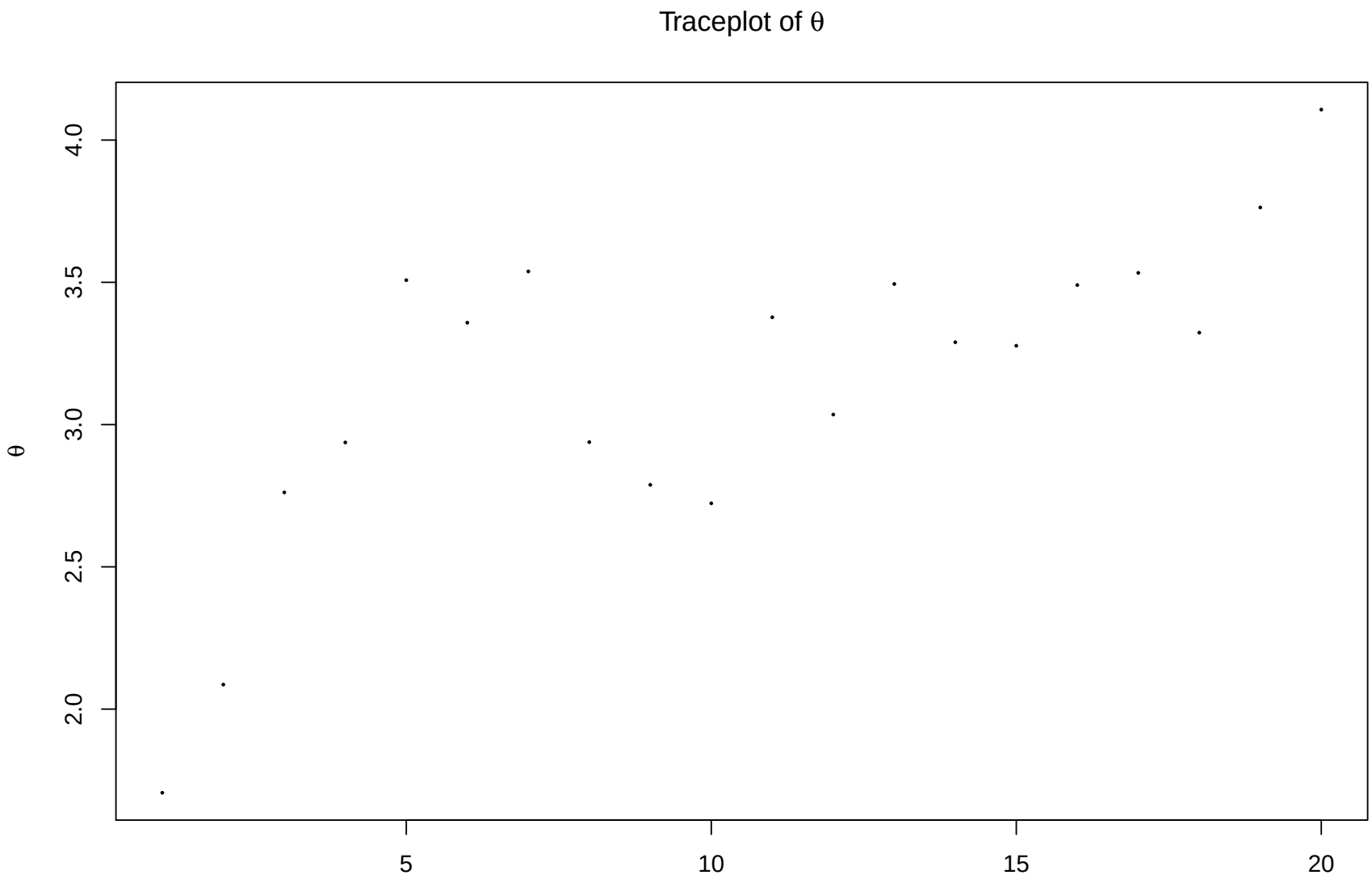
```
##           [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## [15,] 3.277044 2.271941 3.007430 2.229772 3.174802 3.235341
## [16,] 3.490379 2.991483 2.673884 3.348854 1.755765 2.466689
## [17,] 3.533226 2.726794 3.264838 3.026639 4.797629 1.402770
## [18,] 3.323041 3.120857 2.112456 2.608953 3.338799 2.615662
## [19,] 3.763003 2.826198 3.377921 2.776521 3.274876 2.538015
## [20,] 4.106968 2.385681 2.334286 3.791320 1.903190 1.848194
```

Output of the Gibbs sampler

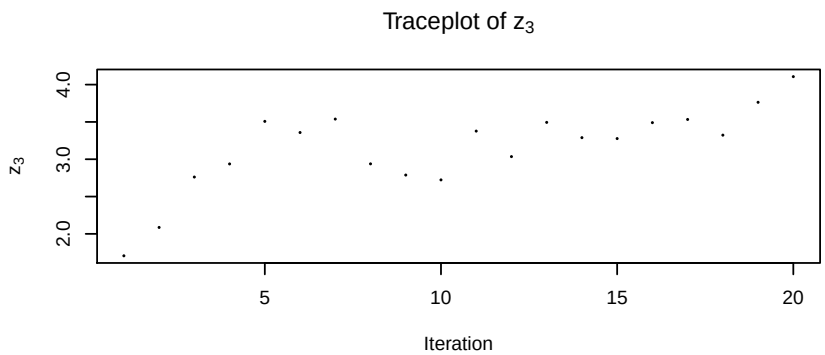
Let's first look at some diagnostics — trace plots and runnning average plots.

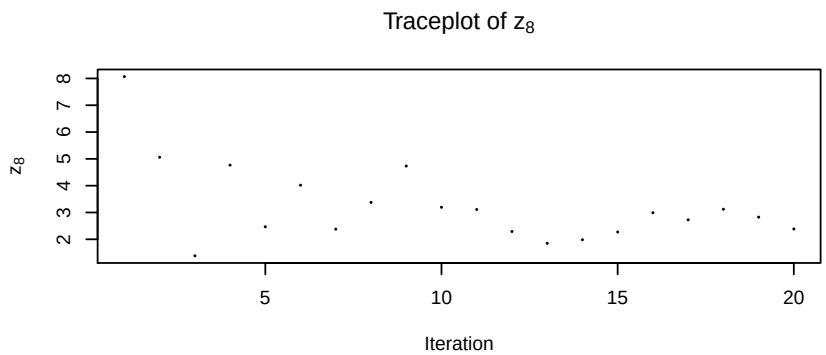
Traceplot of θ

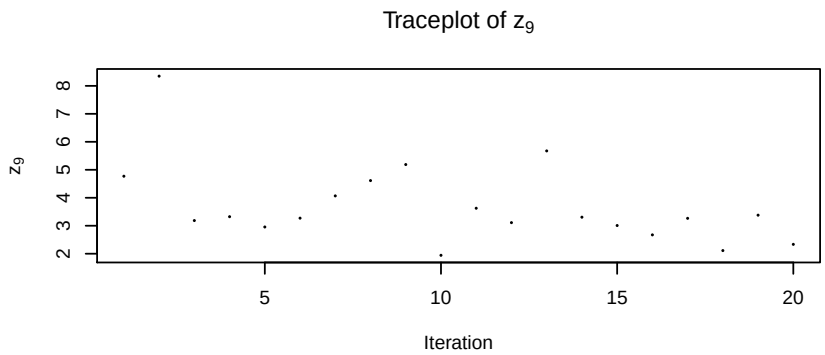
```
plot(1:n.iter, res[,1], pch = 16, cex = .35,
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Traceplot of ", theta)))
```

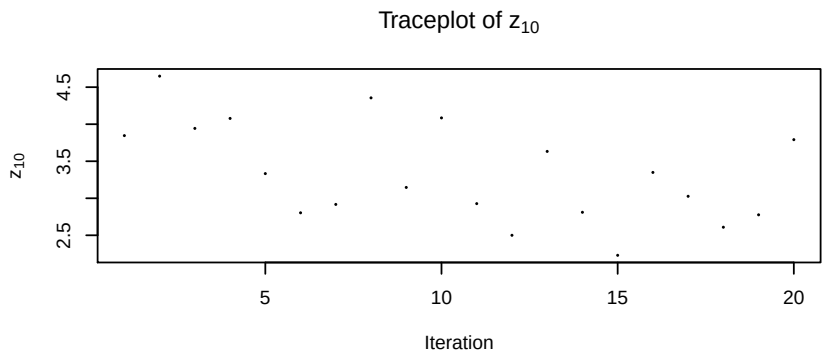


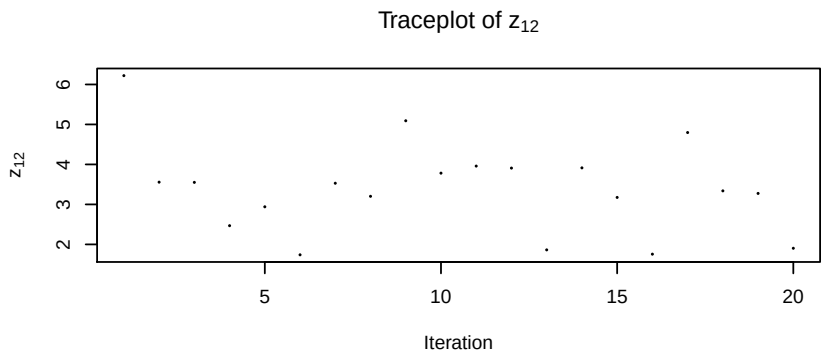
Traceplot of censored observations



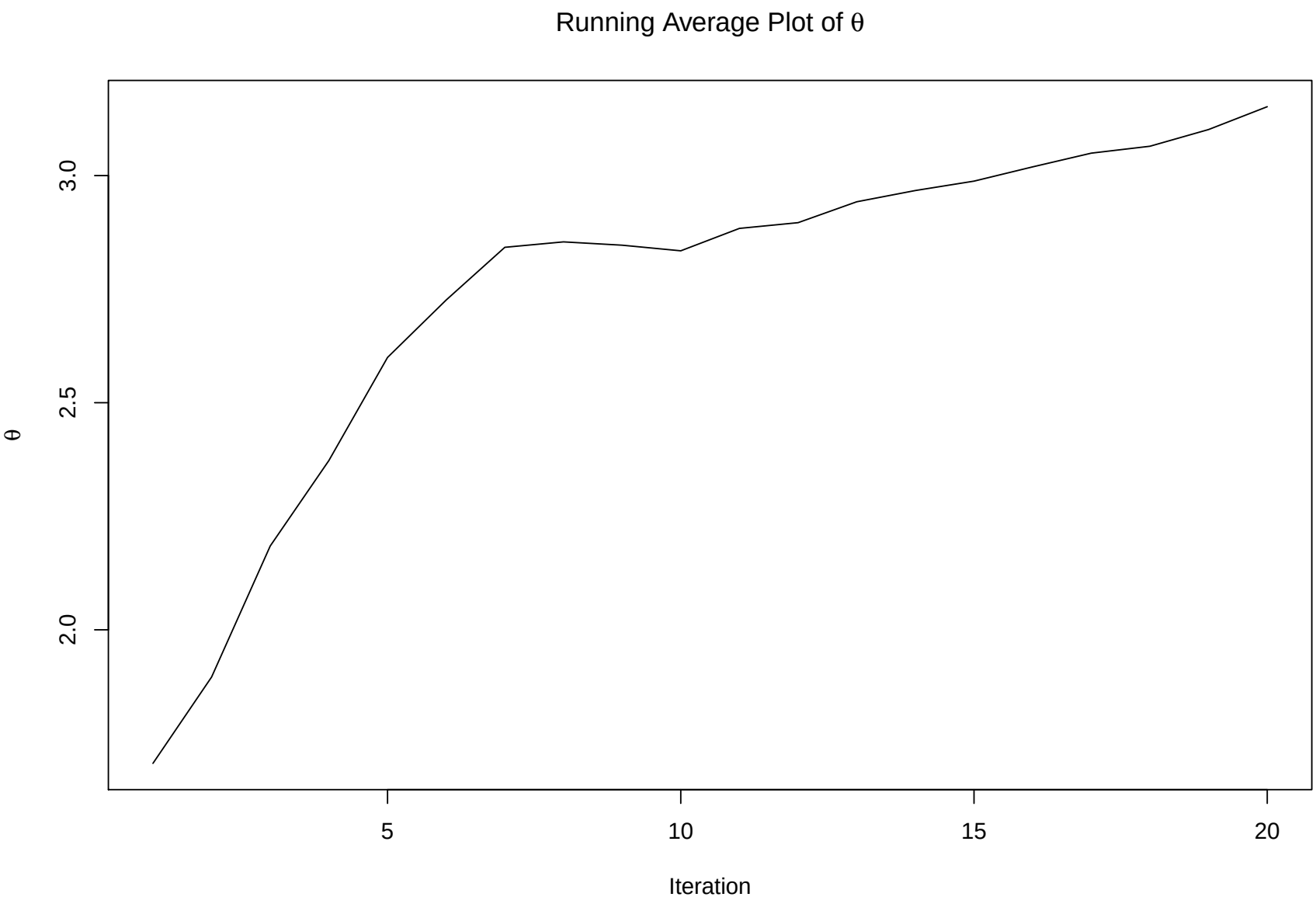




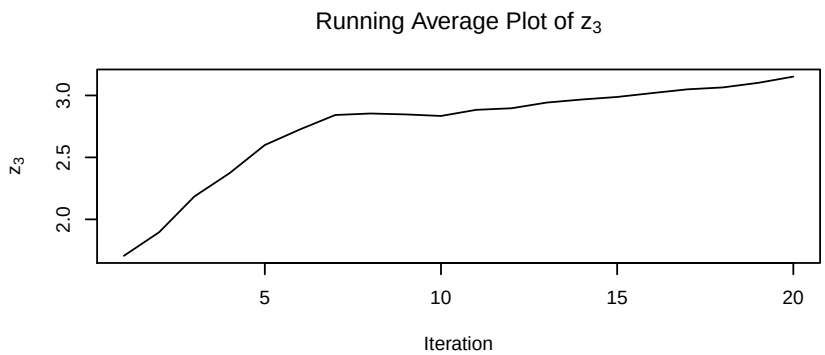


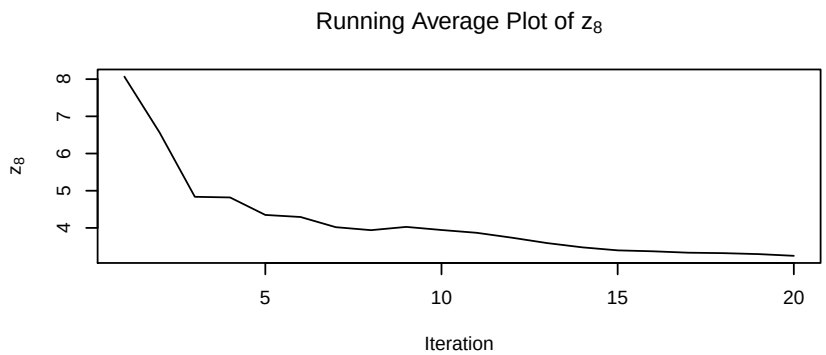


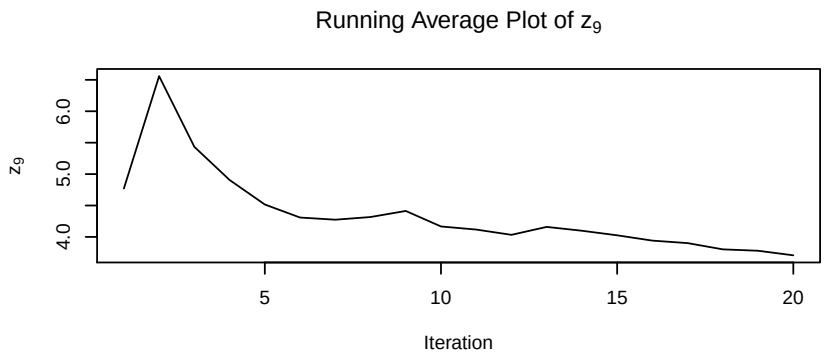
Running average plots

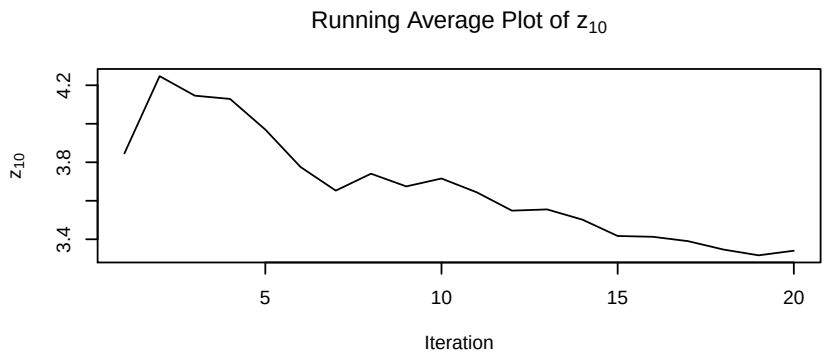


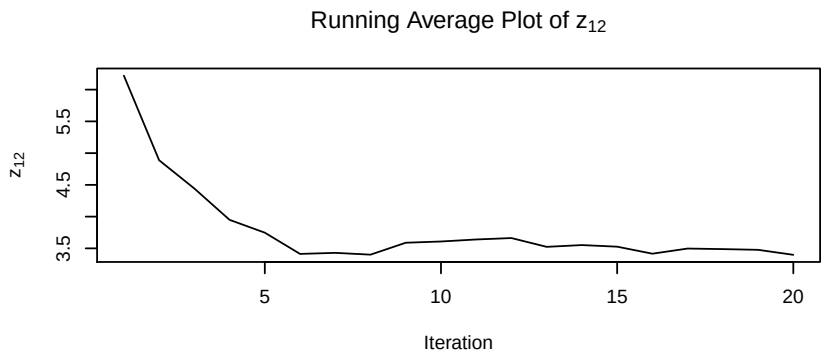
Running average plots



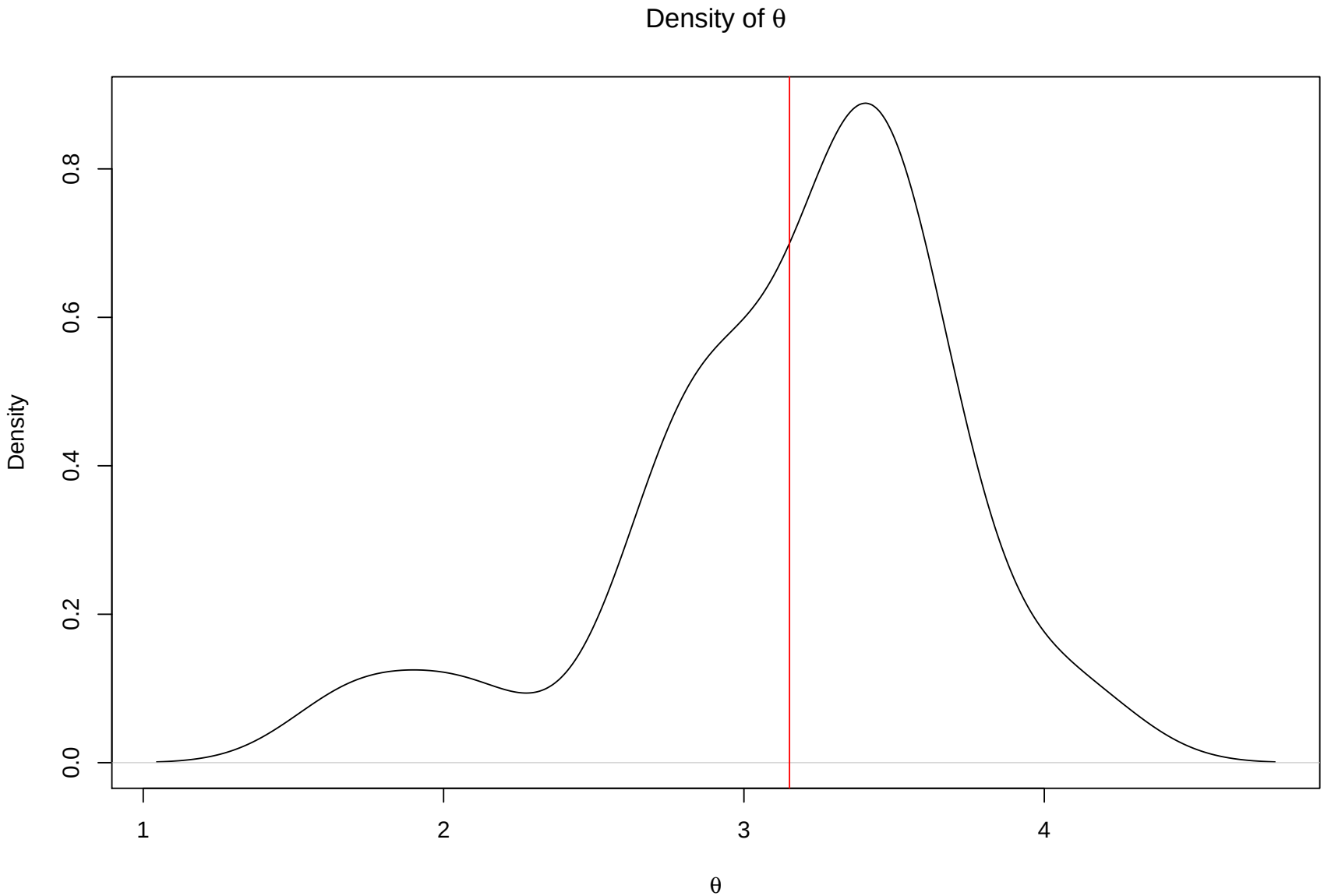




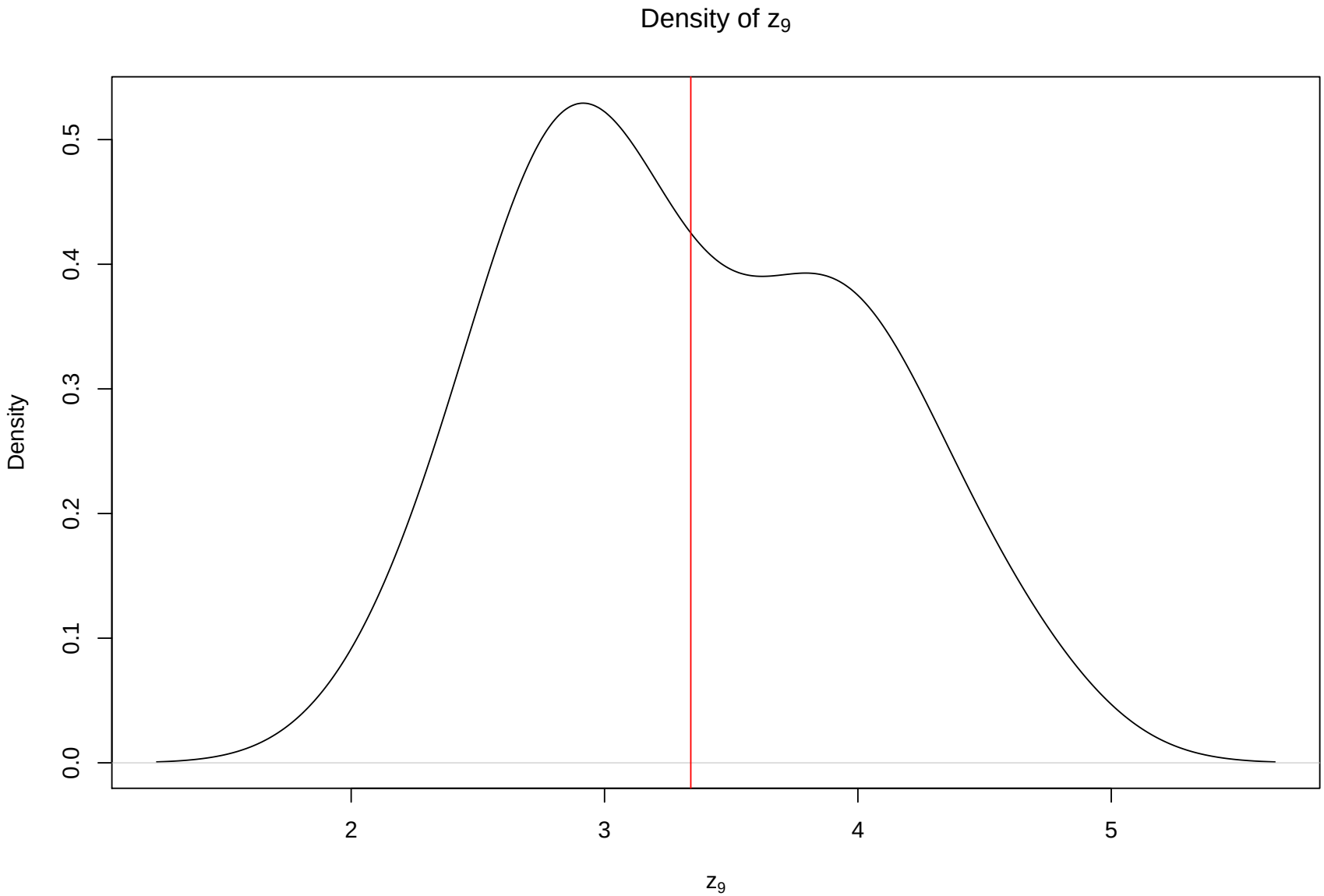




The estimated density of θ



The estimated density of z_9



Homework 6

Using the data and functions given to you (posted to Sakai with Homework 6), investigate the following questions:

1. Write code to produce traceplots and running average plots for the censored values for 40 iterations. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.
2. 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.
3. 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).
4. Finally, let's suppose that $r = 10, a = 1, b = 100$. Does your posterior change? What about when $r = 10, a = 100, b = 1$? (Use 10,000 iterations for the Gibbs sampler).