Module 7: Part II: Gibbs Sampling with an Application to Missing Data

Rebecca C. Steorts

Announcements

- For exam II, I will allow you either to take the Exam II (in class) or participate in Datathon.
- The choice is up to you. You must email myself and the TA team no later than this Friday if you plan on working on Datathon and let me know your teams (no larger than groups of two).
- You must work with a team from the class for Datathon, and complete a submission that will be graded by myself/TAs.
- Part of your analysis must be Bayesian in nature in order to receive a passing grade on the Datathon exam.
- If you have trouble finding a group, please let me know.
- If you do not follow these instructions for Datathon, you may not receive a passing grade on the exam.

Announcements

- How are final grades calculated in this class?

https://github.com/resteorts/modern-bayes/blob/master/syllabus/syllabus-sta360-fall21.pdf

- If you would like to know your raw score in the class, rank, and grade, please come see me on Thursday at the start of OH so I can quickly go through these or set up a quick time over zoom.

Agenda

- Three stage Gibbs sampler
- Gibbs sampling (multi-stage sampler)
- Missing data (censoring) application

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 $z|x, y$, 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 $v^i | \cdots$.

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.

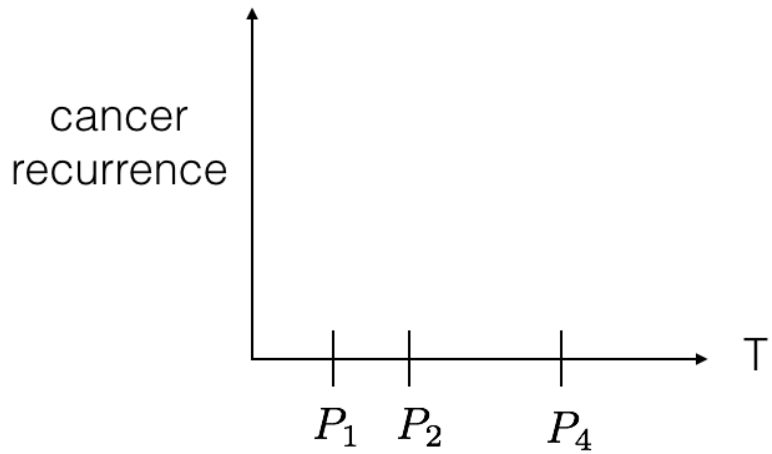


Figure 1: Example of censoring for medical data.

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.

Goal

The goal of this module is to introduce **censoring**, where we need to impute observations from the data.

We will do this through a new concept known a **latent variable**.

Then we will utilize concepts that we have learned throughout the semester which include:

- hierarchical modeling
- semi-conjugacy
- Gibbs sampling
- inverse CDF method
- MCMC diagnostics
- approximating posterior distributions and explaning our results
- performing sensitivity analyses

Background

A **latent variable** is the true version of the state of a random variable that is unknown and not directly observed.

Example: We do not know observed state of every patient (in years), so we can model this unsing a latent variable.

Latent Variable

Let Z_i denote a latent variable for each individual i in the study.

Case 1: If X_i is a censored value, then $X_i = c_i$. This means that

$$Z_i \geq c_i = Z_i > c_i,$$

if Z_i is continuous.

Case 2: If X_i is observed (without missingness) and assuming no noise, then

$$Z_i = X_i.$$

Here, $Z_i \leq c_i$. Why? We considered the other case already above.

Example Likelihood

Suppose that we have two patients, one fully observed at $X_1 = 6$ years and one censored that leaves the study early at $X_2 = 3+$ years.

1. Since the first patient $X_1 = 6$ is fully observed, $Z_1 = X_1 = 6$.
2. Since the second patient X_2 is censored, we know that $X_2 = 3+$. This means that $Z_2 > 3$.

Thinking ahead, we only need to impute censored values. (We won't ever impute fully observed data because we already know it!)

Likelihood

We can summarize the likelihood into the following:

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

where

1. X_i is observed (without missingness) and no noise so $Z_i = X_i$.
2. If X_i is a censored value, then $X_i = c_i$. Thus, $Z_i \geq c_i = Z_i > c_i$, if Z_i is continuous.

Takeaway: The top line of the likelihood is when the observation was completely observed. The bottom line is when it was partially observed (censored), so we need to impute it.

Model

We can write a generative model to include the likelihood as

$$X_i = \begin{cases} Z_i & \text{if } Z_i \leq c_i \\ c_i & \text{if } Z_i > c_i \end{cases} \tag{2}$$

$$Z_1, \ldots, Z_n | \theta \overset{iid}{\sim} \text{Gamma}(r, \theta) \tag{3}$$

$$\theta \sim \text{Gamma}(a, b) \tag{4}$$

where a, b, and r are known.

- c_i is the censoring time for patient i.
- θ is the rate parameter for the lifetime distribution.
- Z_i is the lifetime for patient i, which is latent (unknown).

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:

$$X_i = \begin{cases} Z_i & \text{if } Z_i \leq c_i \\ c_i & \text{if } Z_i > c_i \end{cases} \tag{5}$$

$$Z_1, \ldots, Z_n | \theta \overset{iid}{\sim} \text{Gamma}(r, \theta) \tag{6}$$

$$\theta \sim \text{Gamma}(a, b) \tag{7}$$

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

The full conditionals are easy to calculate. 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 $z_{1:n}$),

$$p(\theta|\cdots) = p(\theta|z_{1:n}, x_{1:n}) = p(\theta|z_{1:n}) \tag{8}$$

$$= \text{Gamma}\left(\theta \,\middle|\, a + nr,\ b + \textstyle\sum_{i=1}^{n} z_i\right) \tag{9}$$

  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

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

Full conditionals (continued)

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

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

Sampling from the truncated Gamma

For the censored values of Z_i we know c_i.

Given θ, $Z_i|\theta \sim \text{Gamma}(r, \theta)$.

Let F be the CDF of $\text{Gamma}(Z_i \mid r, \theta)$ and truncate to (c_i, ∞).

Consider the following:

$$P(Z_i < z_i \mid c_i) = \frac{F(z_i) - F(c_i)}{1 - F(c_i)}.$$

Then $Z_i \mid c_i$ has a truncated Gamma distribution.

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

Application to censored data

As a part of homework 6, you will work on understanding how to put these details together. There is template file to help you with homework 6 that can be found at

https://github.com/resteorts/modern-bayes/blob/master/homeworks/homework-6/template-hw6.Rmd and https://github.com/resteorts/modern-bayes/blob/master/homeworks/homework-6/template-hw6.pdf.

Application to censored data

```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)

```r
# 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 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,])
}
```

Set parameter values

```
set.seed(5983)
# set parameter values and enter data
r <- 10
a <- 1
b <- 1
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 <- 100
init.theta <- 1
init.missing <-
  rgamma(length(c), shape = r, rate = init.theta)
```

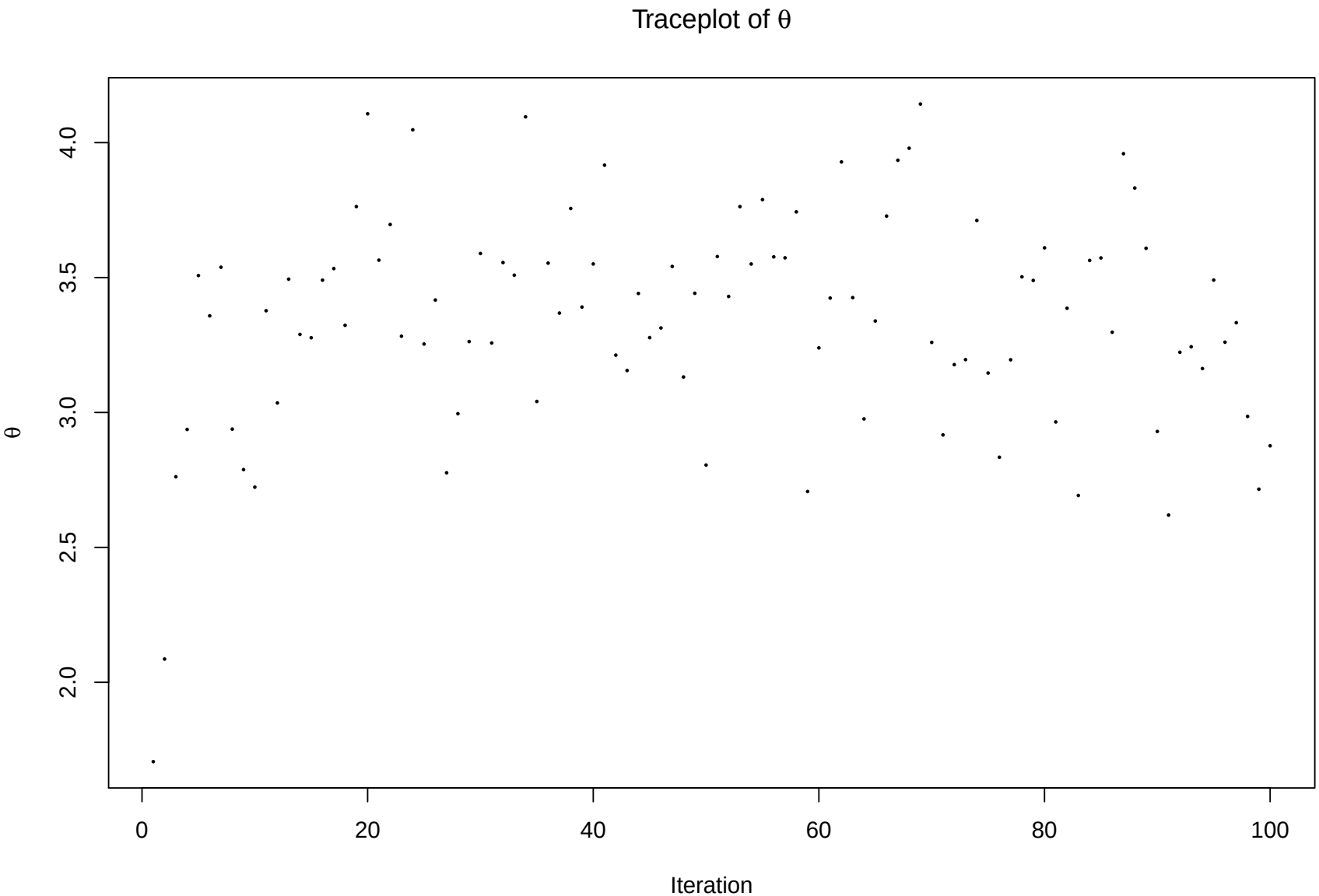
Run Gibbs sampler

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

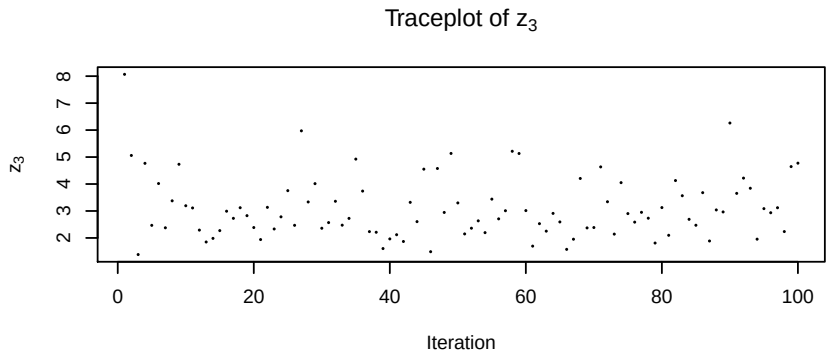
Let's first look at some diagnostics — trace plots and runnning average plots.

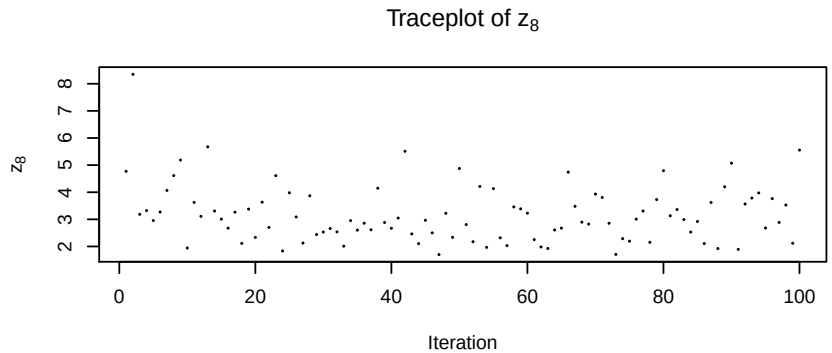
Traceplot of θ

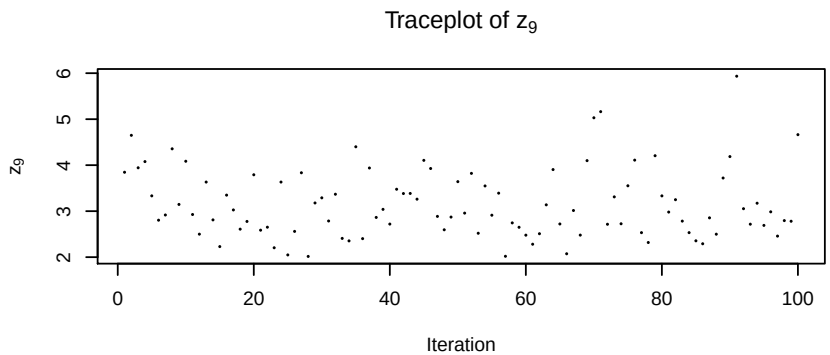
```
plot(1:n.iter, res[,1], pch = 16, cex = .35,
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Traceplot of ", theta)))
```

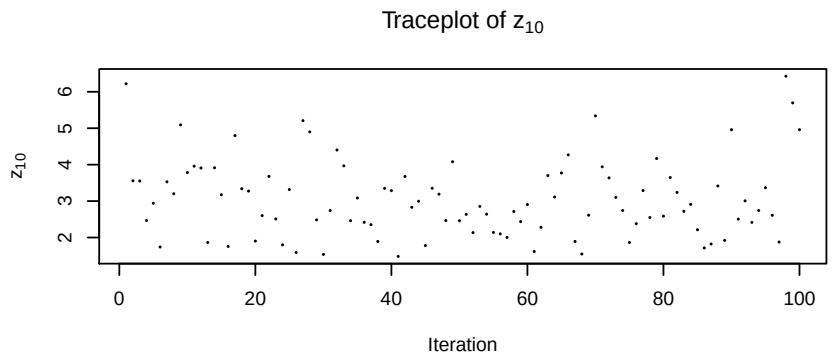


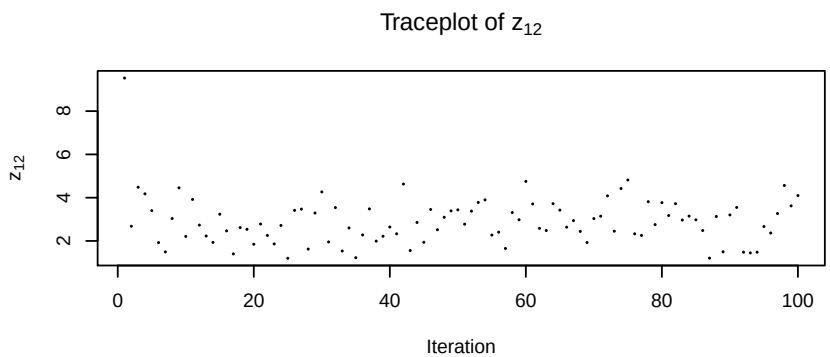
Traceplot of censored observations



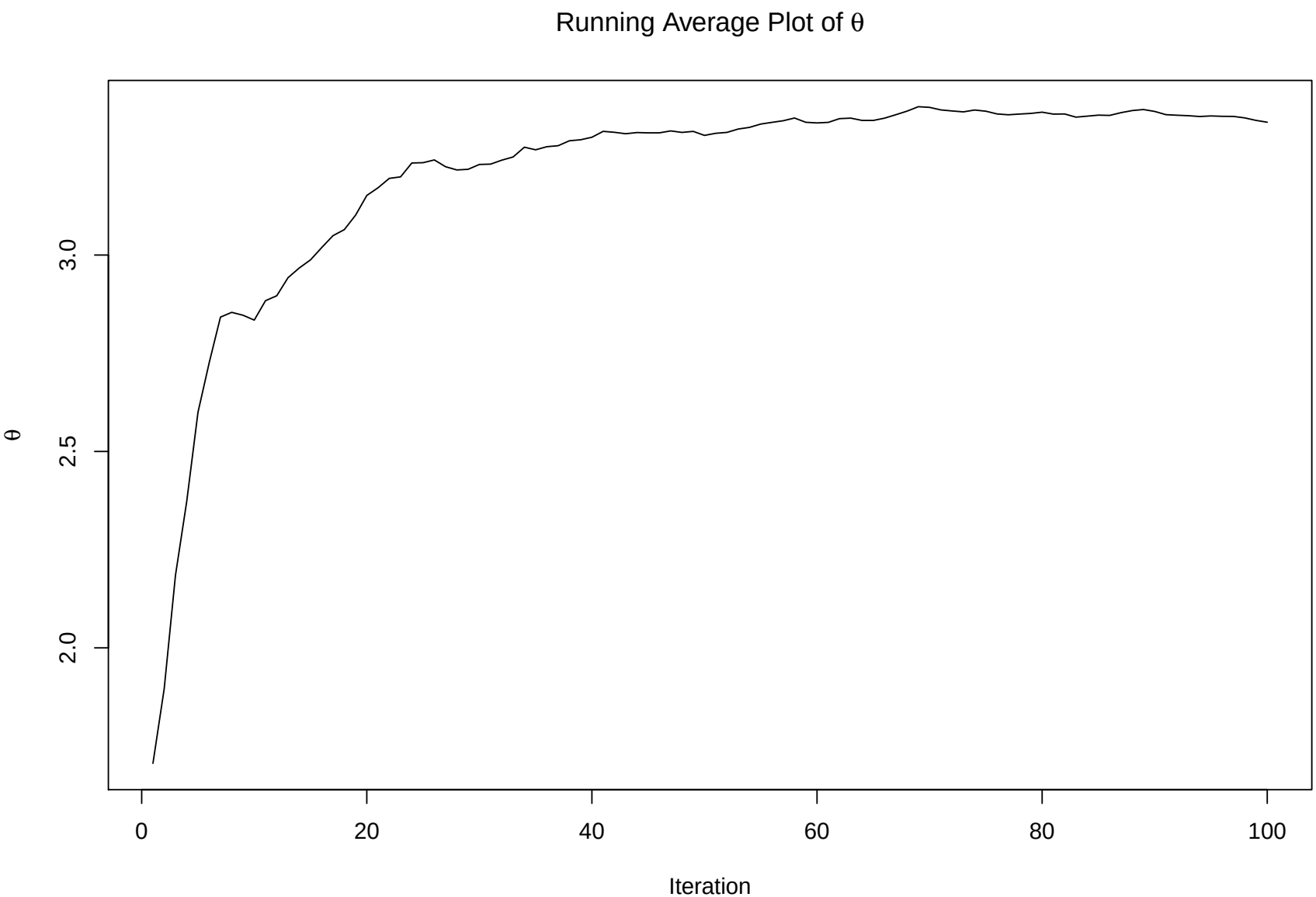




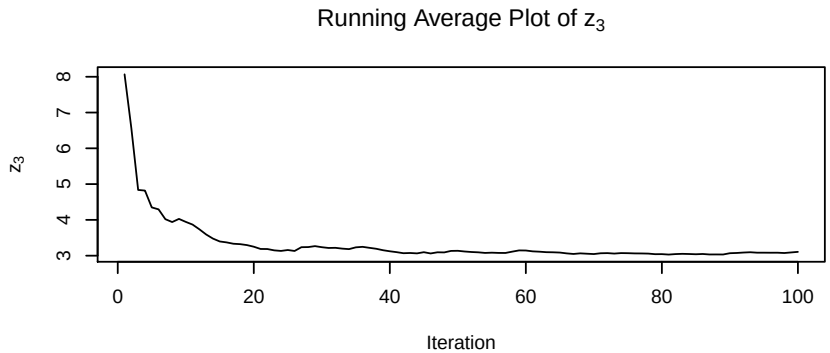


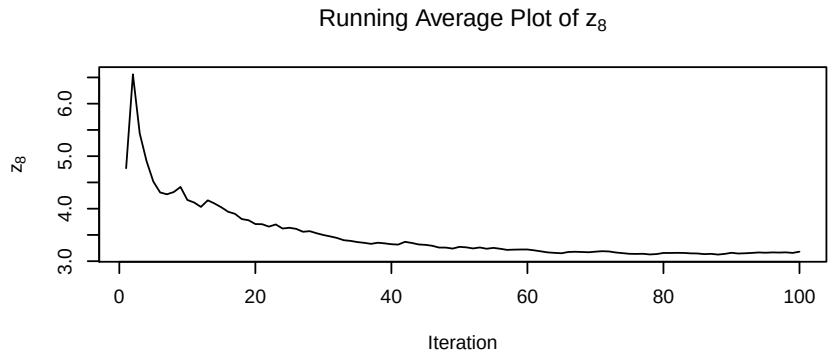


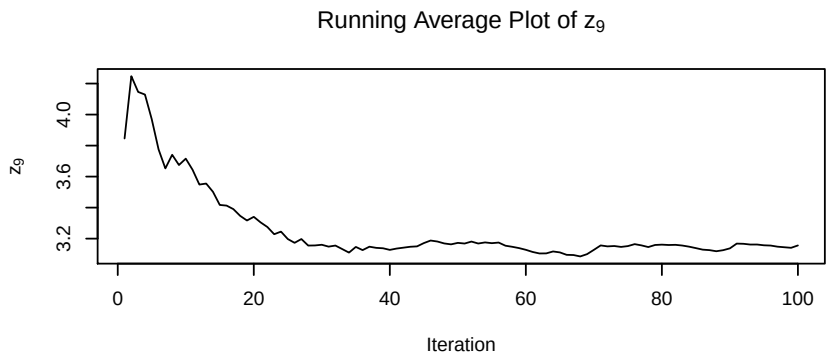
Running average plots

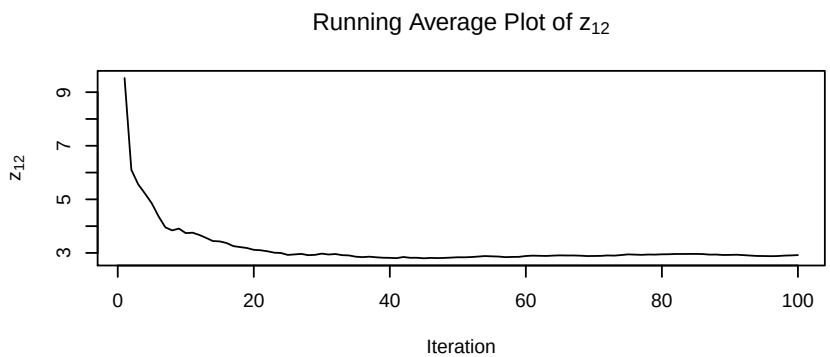


Running average plots

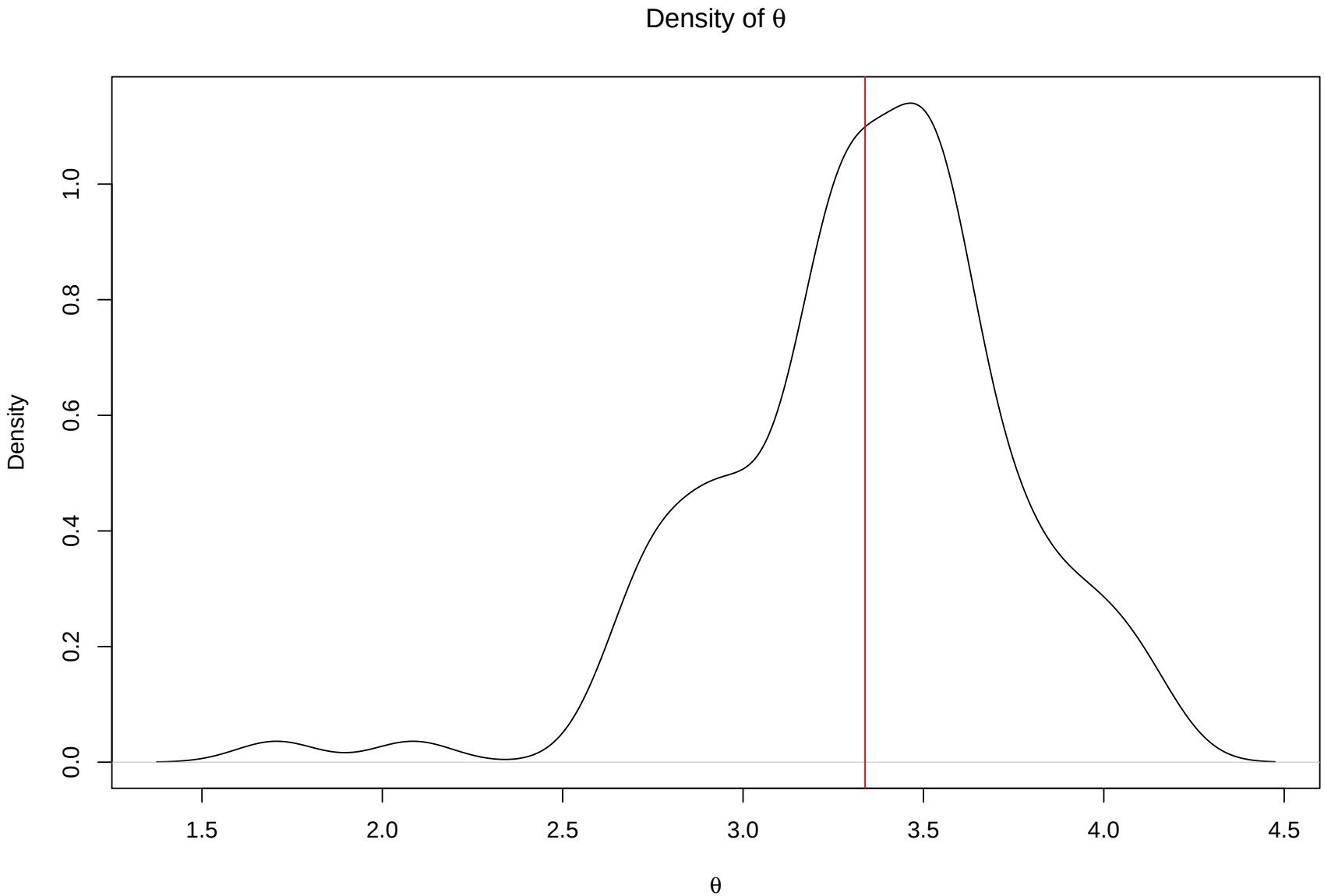




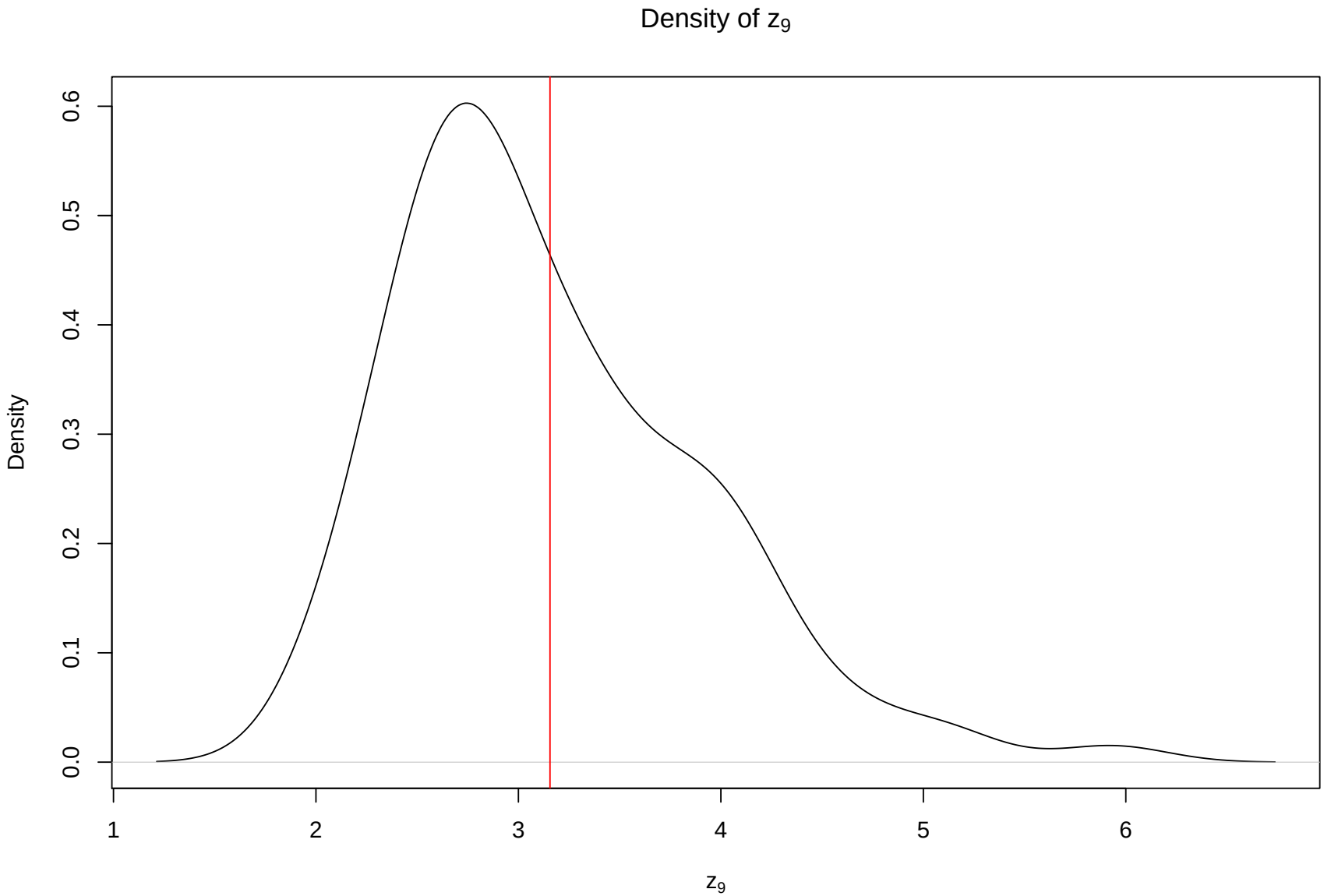




The estimated density of θ



The estimated density of z_9



Homework 6

Using the data and functions given to you in this module, investigate the following questions. The homework question is summarized for you below and more fully on homework 6.

1. 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.

2. Now run the chain for 10,000 iterations and update your diagnostic plots (trace plots and running average plots). Report your findings for both trace plots 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 $r = 10, a = 100, b = 1$?

Resources

See
https://www.johndcook.com/CompendiumOfConjugatePriors.pdf
for derviations of conjugate families of distributions.

Detailed Takeways

- Three stage Gibbs sampler
- Multistage Gibbs sampler
- Case study for censored data (heart disease)
- Background: Latent variable
- Utilizing a latent variable model
- Conditional distributions
- Truncated Gamma
- Interactive Application in Class (Homework 6)
- Follow up after Homework 6 – What did you learn?