STA360 Homework 7 (Ken Ye)

```
library(latex2exp)
library(ggplot2)
library(MASS)
library(ggrepel)
library(mvtnorm)
set.seed(0)
```

Question 3 (Exercise 7.3)

```
# read data
bluecrab <- as.matrix(read.table(url('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/bluecrab.dat')))
orangecrab <- as.matrix(read.table(url('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/orangecrab.dat')))
```

Part a

```
# function for getting posteior samples
sample.crab <- function(crab) {
  n <- nrow(crab)
  p <- ncol(crab)
  ybar <- colMeans(crab)

  # priors
  mu0 <- ybar
  lambda0 <- cov(crab)
  s0 <- cov(crab)
  nu0 <- 4

  sample.size <- 10000
  THETA <- matrix(nrow = sample.size, ncol = p)
  SIGMA <- array(dim = c(p, p, sample.size))

  # start value
  sigma <- s0

  # Gibbs sampling
  for (s in 1 : sample.size) {
    # update theta
    lambdan <- solve(solve(lambda0) + n * solve(sigma))
    mun <- lambdan %*% (solve(lambda0) %*% mu0 + n * solve(sigma) %*% ybar)
    theta <- mvrnorm(n = 1, mun, lambdan)

    # update sigma
    sn <- s0 + (t(crab) - c(theta)) %*% t(t(crab) - c(theta))
    sigma <- solve(rWishart(1, nu0 + n, solve(sn))[, , 1])

    # store sample
    THETA[s, ] <- theta
    SIGMA[, , s] <- sigma
  }

  # return samples
```

```
  list(theta = THETA, sigma = SIGMA)
}
```

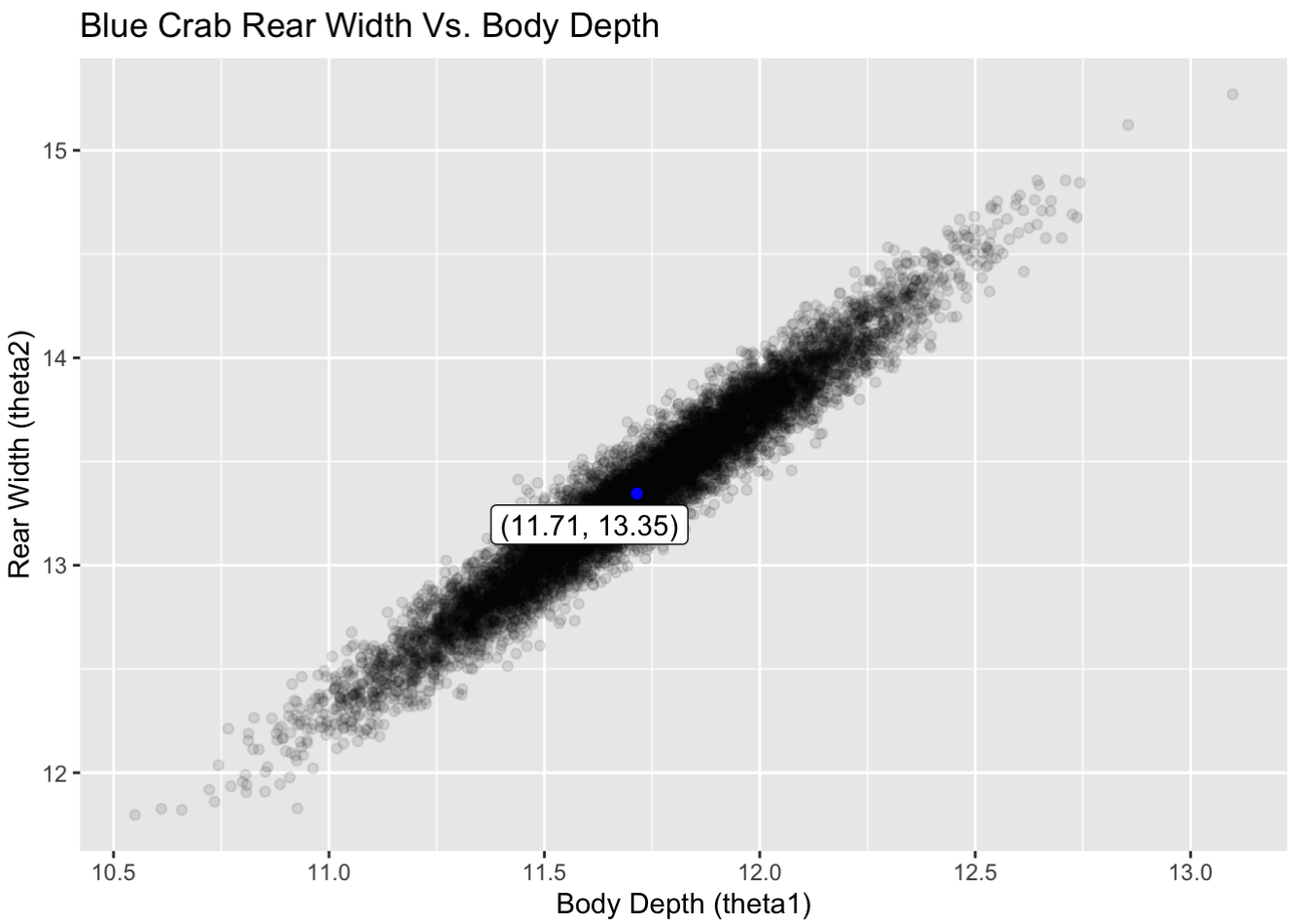
```
# obtain samples
blue.samples <- sample.crab(bluecrab)
orange.samples <- sample.crab(orangecrab)
```

Part b

```
# create dataframes
bluecrab.df = data.frame(blue.samples$theta)
colnames(bluecrab.df) = c('theta1', 'theta2')
orangecrab.df = data.frame(orange.samples$theta)
colnames(orangecrab.df) = c('theta1', 'theta2')
```

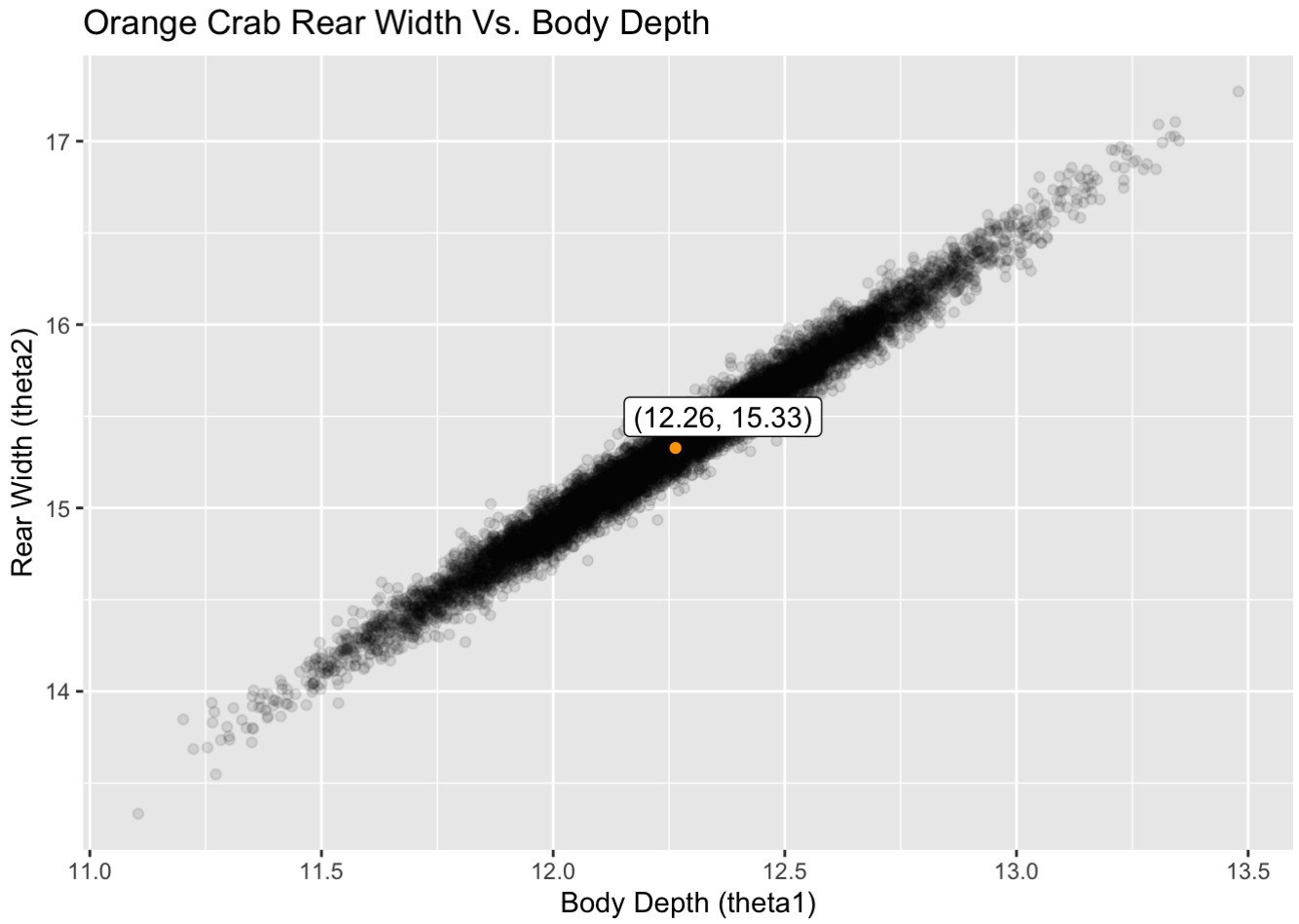
```
# blue crab distribution
bluecrab.means <- as.data.frame(t(as.matrix(colMeans(bluecrab.df[, c('theta1', 'theta2')]))))

ggplot(bluecrab.df, aes(x = theta1, y = theta2)) +
  geom_point(alpha = 0.1) +
  geom_point(data = bluecrab.means, color = 'blue') +
  geom_label_repel(data = bluecrab.means, aes(label = paste0("(", round(theta1, 2), ", ", round(theta2, 2),
")"))) +
  labs(title = 'Blue Crab Rear Width Vs. Body Depth',
       x = 'Body Depth (theta1)',
       y = 'Rear Width (theta2)')
```


Blue Crab Rear Width Vs. Body Depth
Rear Width (theta2)
Body Depth (theta1)
(11.71, 13.35)
15
14
13
12
10.5
11.0
11.5
12.0
12.5
13.0

```
# orange crab distribution
orangecrab.means <- as.data.frame(t(as.matrix(colMeans(orangecrab.df[, c('theta1', 'theta2')]))))

ggplot(orangecrab.df, aes(x = theta1, y = theta2)) +
  geom_point(alpha = 0.1) +
  geom_point(data = orangecrab.means, color = 'orange') +
  geom_label_repel(data = orangecrab.means, aes(label = paste0("(", round(theta1, 2), ", ", round(theta2, 2),
")"))) +
  labs(title = 'Orange Crab Rear Width Vs. Body Depth',
       x = 'Body Depth (theta1)',
       y = 'Rear Width (theta2)')
```



```
# compare
mean(orangecrab.df$theta1 > bluecrab.df$theta1)
```

```
## [1] 0.9056
```

```
mean(orangecrab.df$theta2 > bluecrab.df$theta2)
```

```
## [1] 0.9986
```

According to the distribution plots, it is clear that orange crabs tend to have both larger body depth (theta1) and rear width (theta2) than blue crabs. In fact, the mean (body depth, rear width) pair for orange crabs is (12.26, 15.33), whereas that of the blue crabs is (11.71, 13.35).

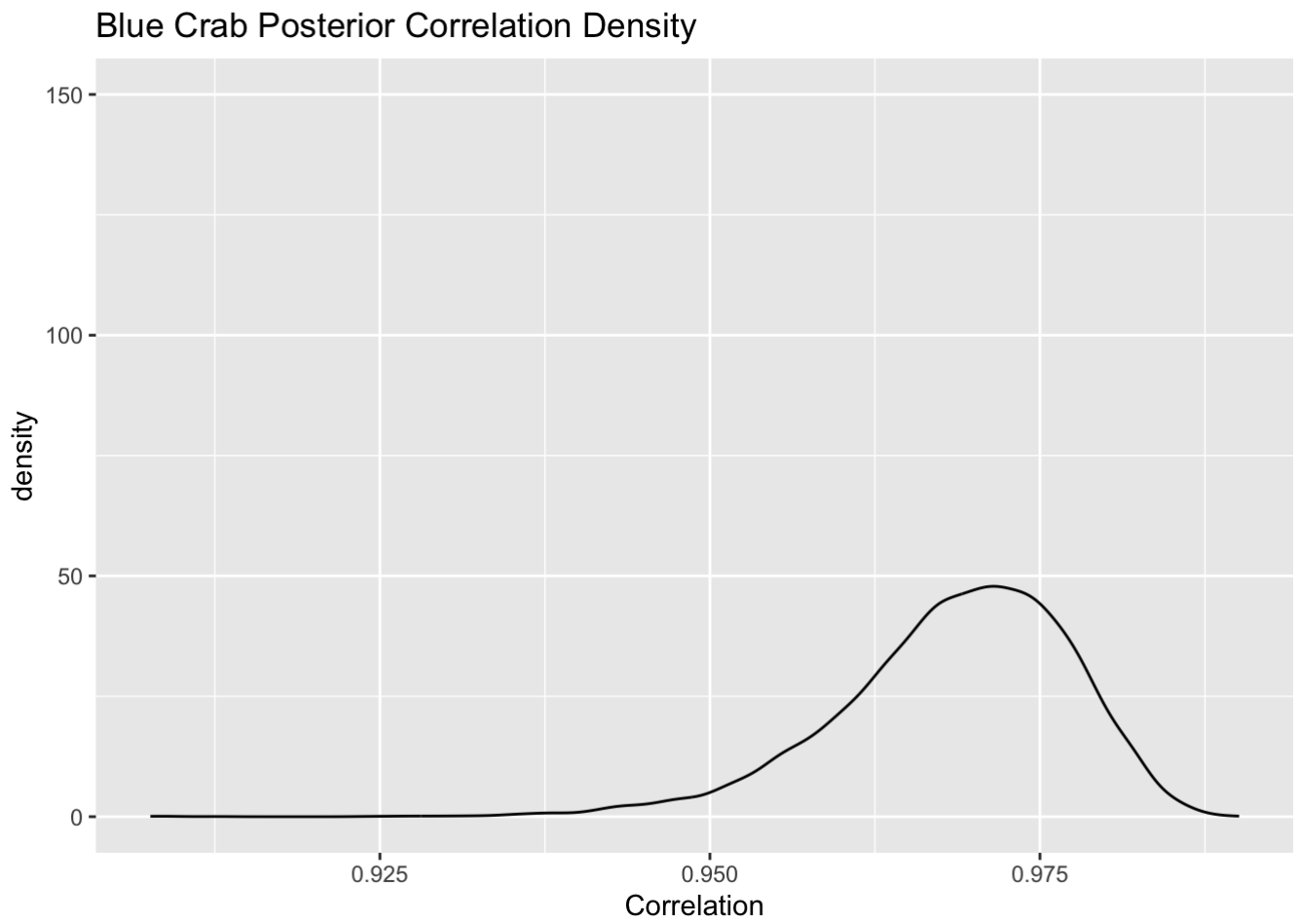
In addition, for all orange & blue crab pairs, 90.56% of the time the body depth of the orange crab is larger than that of the blue crab, and 99.86% of the time the rear width of the orange crab is larger than that of the blue crab.

Part c

```
# fuction for calculating correlation from covariance matrix
cal.corr <- function(cov){
  cov[1, 2] / (sqrt(cov[1, 1] * cov[2, 2]))
}
```

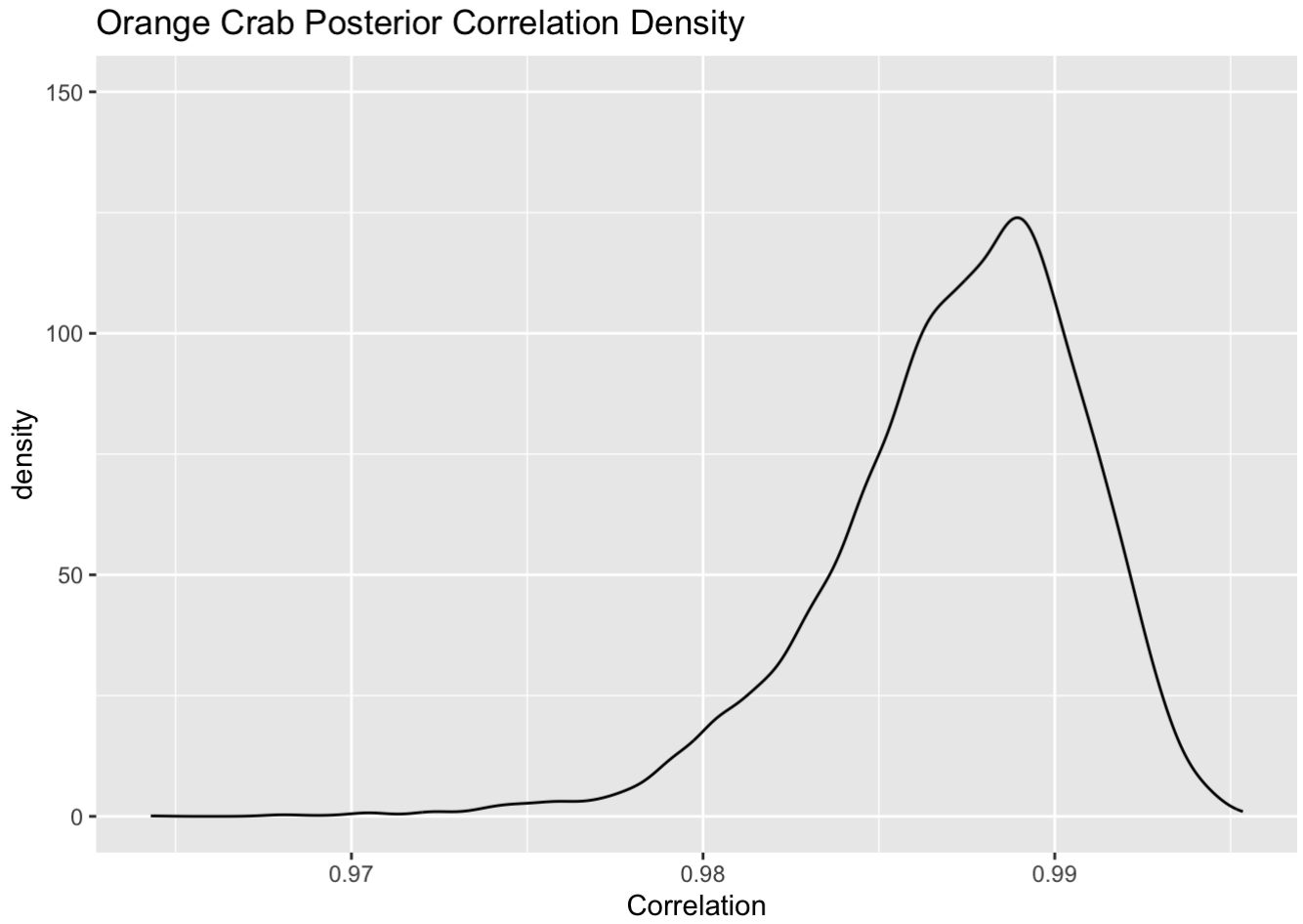
```
# blue crab posterior densities of the correlations
bluecrab.cor <- apply(blue.samples$sigma, MARGIN = 3, FUN = cal.corr)

bluecrab.cor.df <- data.frame(Correlation = bluecrab.cor)
ggplot(bluecrab.cor.df, aes(x = Correlation)) +
  geom_density() +
  labs(title = 'Blue Crab Posterior Correlation Density') +
  ylim(0, 150)
```




```
# orang crab posterior densities of the correlations
orangecrab.cor <- apply(orange.samples$sigma, MARGIN = 3, FUN = cal.corr)

orangecrab.cor.df <- data.frame(Correlation = orangecrab.cor)
ggplot(orangecrab.cor.df, aes(x = Correlation)) +
  geom_density() +
  labs(title = 'Orange Crab Posterior Correlation Density') +
  ylim(0, 150)
```




```
# Pr(phi_blue < phi_orange | y_blue, y_orange)
mean(bluecrab.cor < orangecrab.cor)
```

```
## [1] 0.9899
```

According to the plots, orange crabs (higher peak, and peak at around 0.99) appear to have a higher correlation between their body depth and rear width than the blue crabs (lower peak, and peak at around 0.97).

In addition, Pr(phi_blue < phi_orange | y_blue, y_orange) is 0.9899, suggesting that for all orange & blue crab pairs, 98.99% of the time orange crabs' correlation between their body depth and rear width is larger that of the blue crabs.

Question 4

```
# read data
Y <- readRDS("~/OneDrive - Duke University/2023 Spring/STA360/HW7/hw7train.rds")
test <- readRDS("~/OneDrive - Duke University/2023 Spring/STA360/HW7/hw7test.rds")
```

Part b

```
n <- dim(Y)[1]
p <- dim(Y)[2]

# priors
tao2 <- 1
mu0 <- rep(0,14)
sd0 <- (mu0/2)
L0 <- matrix(0, p, p)
diag(L0) <- 1
# L0 <- L0*outer(sd0, sd0)
nu0 <- p+2
S0 <- L0

# starting values
Sigma <- S0
Y.full <- Y
O <- 1*(!is.na(Y))
for (j in 1:p){
  Y.full[is.na(Y.full[,j]),j] <- mean(Y.full[,j], na.rm = TRUE)
}

# Gibbs sampler
THETA <- SIGMA <- Y.MISS <- NULL
for (s in 1:1000){
  # update theta
  ybar <- apply(Y.full, 2, mean)
  Ln <- solve(solve(L0) + n*solve(Sigma))
  mun <- Ln %*% (solve(L0) %*% mu0 + n*solve(Sigma) %*% ybar)
  theta <- rmvnorm(1, mun, Ln)

  # update Sigma
  Sn <- S0 + (t(Y.full) - c(theta)) %*% t(t(Y.full) - c(theta))
  Sigma <- solve(rWishart(1, nu0 + n, solve(Sn))[,,1])

  # update missing data
  for (i in 60:79){
    b <- (O[i,] == 0)
```

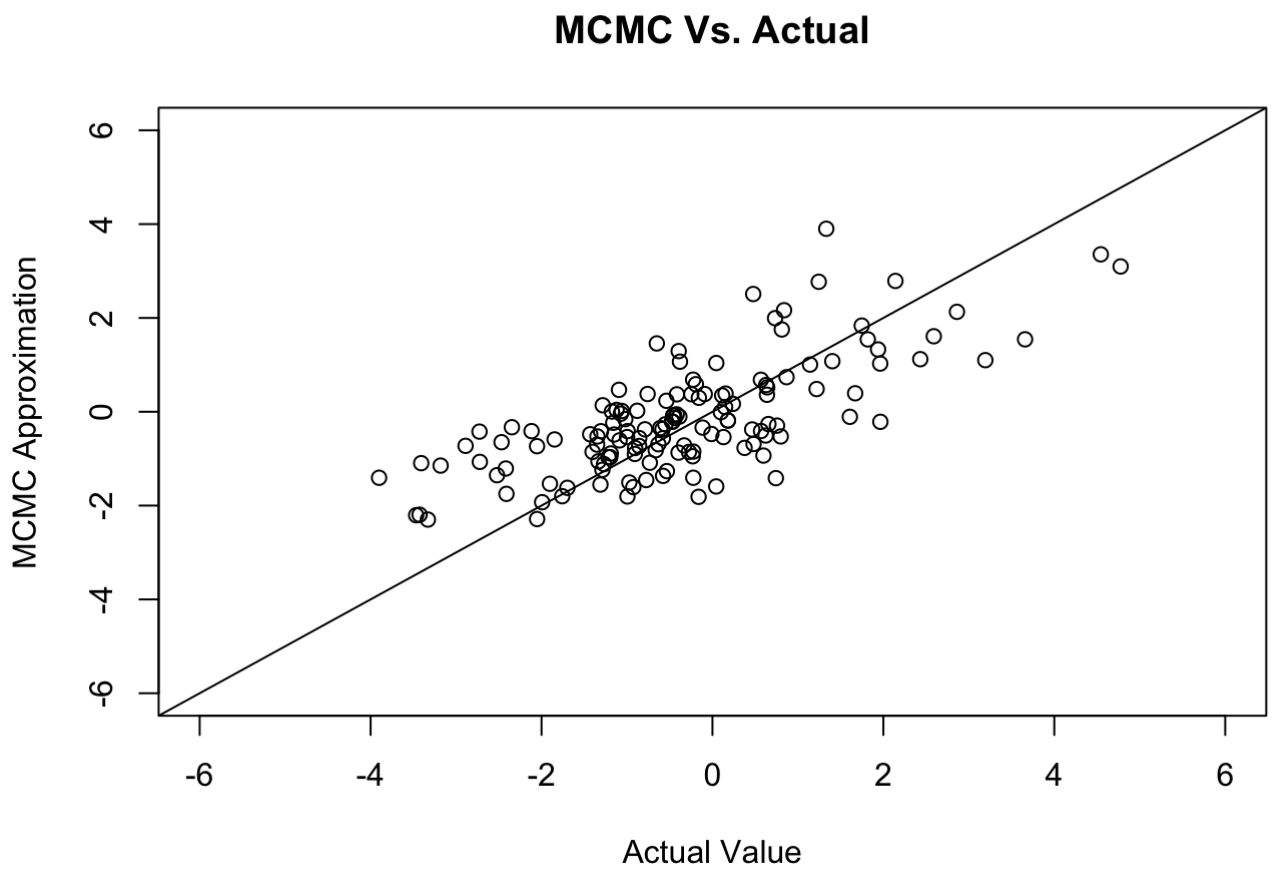
```
    a <- (O[i,] == 1)
    iSa <- solve(Sigma[a, a])
    beta.j <- Sigma[b, a] %*% iSa
    Sigma.j <- Sigma[b, b] - Sigma[b, a] %*% iSa %*% Sigma[a, b]
    theta.j <- theta[b] + beta.j %*% t(t(Y.full[i, a]) - theta[a])
    Y.full[i, b] <- rmvnorm(1, theta.j, Sigma.j)
  }

  # save results
  THETA <- rbind(THETA, theta)
  SIGMA <- rbind(SIGMA, c(Sigma))
  Y.MISS <- rbind(Y.MISS, Y.full[O == 0])
}
```

Part c

```
# construct m by p2 matrix
YB.hat <- apply(Y.MISS, 2, mean)
YB.hat <- matrix(YB.hat, nrow = 20)
```

```
# compare prediction to actual
plot(test[1,1], YB.hat[1,1],
     xlab = "Actual Value",
     ylab = "MCMC Approximation",
     xlim = c(-6,6),
     ylim = c(-6,6))
title("MCMC Vs. Actual")
abline(0,1)
for (i in 1:20){
  for (j in 1:7){
    points(test[i,j], YB.hat[i,j])
  }
}
```



According to this scatter plot, our MCMC approximations are quite close to the actual value, which can be told from the fact that all points are close to the line x = y, meaning the values we imputed are decent.

```
# MCMC approximation error sum
sum((YB.hat - test)^2)
```

```
## [1] 156.8921
```

```
# posterior theta estimate error sum
THETA.hat <- apply(THETA, 2, mean)
sum((THETA.hat - test)^2)
```

```
## [1] 327.1208
```

The prediction error using MCMC approximation is 156.8649, which is smaller than the prediction error (327.117) using posterior mean estimate of theta. This makes sense because with the MCMC approximation method we are making more “individualized” predictions for each missing value, as opposed to using the population statistics.