Homework 6 Solutions

```
set.seed(1)
library(mvtnorm)
```

```
samp.theta <- function(Y, Sigma, mu.0, lambda.0) {
  n <- nrow(Y)
  y.bar <- apply(Y, 2, mean)

  theta.var <- solve(solve(lambda.0) + n*solve(Sigma))
  theta.mean <- theta.var%*%(solve(lambda.0)%*%mu.0 + n*solve(Sigma)%*%y.bar)
  return(rmvnorm(1, theta.mean, theta.var))
}

samp.Sigma <- function(Y, theta, S.0, nu.0) {

  n <- nrow(Y)
  S.n <- S.0 + (t(Y) - c(theta))%*%t(t(Y) - c(theta))

  return(solve(rWishart(1, nu.0 + n, solve(S.n))[, , 1]))
}
```

Problem 1

Part a.

There isn't a single correct answer to this problem.

```
mu.0 <- c(40, 38)
nu.0 <- length(mu.0) + 2
S.0 <- lambda.0 <- rbind(c(100, 25), c(25, 100))
```

One possible answer is:

I set the prior mean to $\mu_0 = (40, 38)$ and set $\Lambda_0 = \begin{pmatrix} 100 & 25 \\ 25 & 100 \end{pmatrix}$, which assumes ages of married couples are correlated and that over 95% of ages are between 16 and 100. As in the test score example, we set $S_0 = \Lambda_0$ and center the prior for Σ weakly about S_0 by setting $\nu_0 = 2 + 2$.

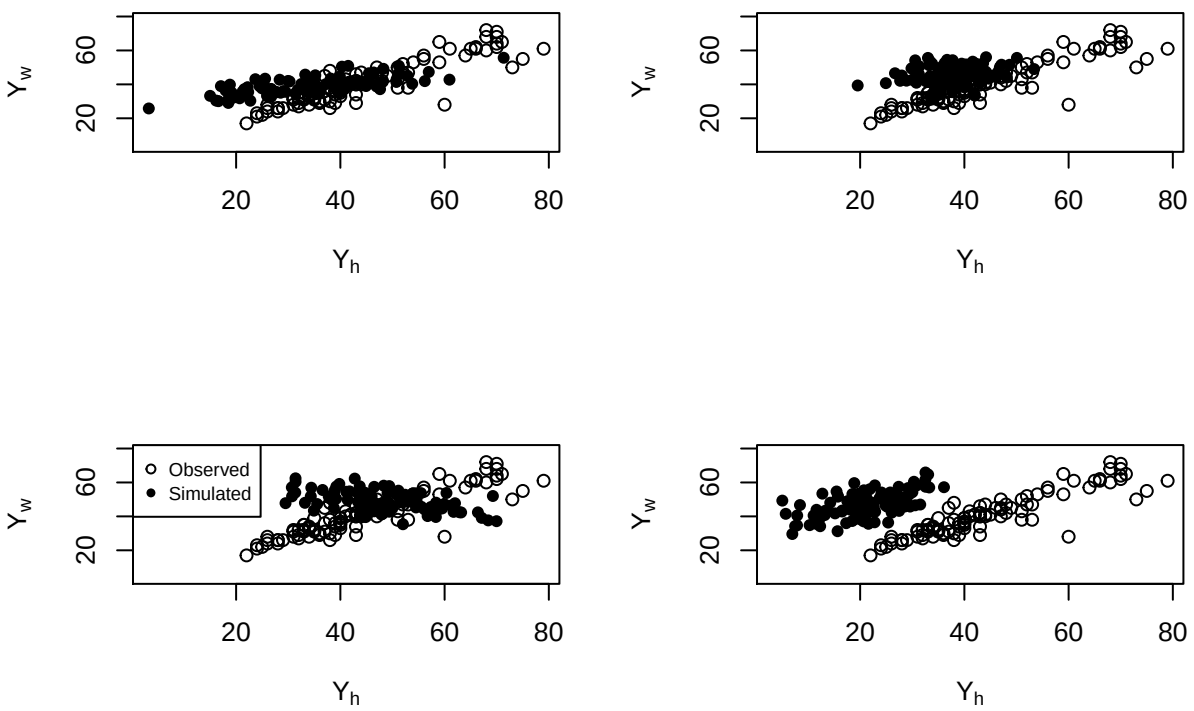
Part b.

Solution depends on your choice in a.

```
Y <- read.table("http://www.stat.washington.edu/~pdhoff/Book/Data/hwdata/agehw.dat",
                    header = TRUE)
```

```
Y1s <- matrix(nrow = nrow(Y), ncol = 4)
Y2s <- matrix(nrow = nrow(Y), ncol = 4)
```

```
for (i in 1:4) {
   theta <- rmvnorm(1, mu.0, lambda.0)
   Sigma <- solve(rWishart(1, nu.0, solve(S.0))[, , 1])
   Y.sim <- rmvnorm(nrow(Y), theta, Sigma)
   Y1s[, i] <- Y.sim[, 1]
   Y2s[, i] <- Y.sim[, 2]
}
par(mfrow = c(2, 2))
range <- range(cbind(Y, Y1s, Y2s))
for (i in 1:4) {
  plot(Y[, 1], Y[, 2], ylim = range, xlim = range, xlab = expression(Y[h]), ylab = expression(Y[w]))
  points(Y1s[, i], Y2s[, i], pch = 16)
  if (i == 3) {legend("topleft", pch = c(1, 16),
                      legend = c("Observed", "Simulated"), cex = 0.75)}
}
```



This prior isn't perfect; ages appear to be less correlated under this prior than in the observed data. But, it's close enough to be plausible so I'll leave it be.

Part c.

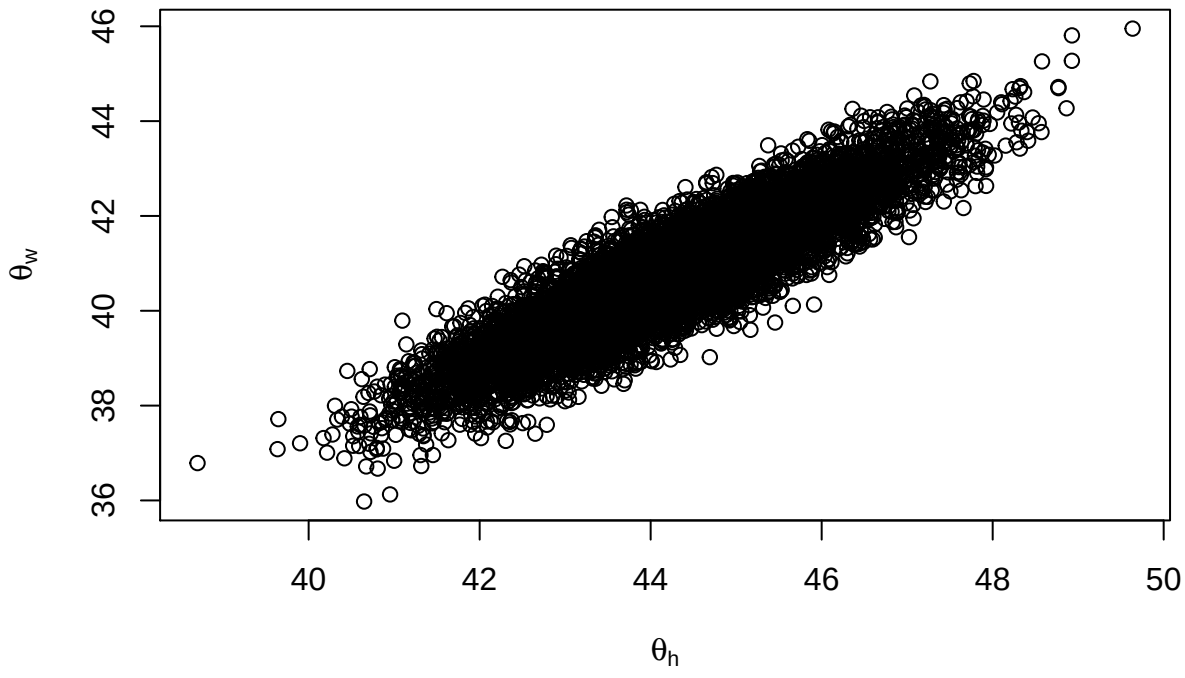
Solution depends on your choice in b.

```
theta <- apply(Y, 2, mean)
Sigma <- cov(Y)

S <- 10000
thetas.c <- matrix(nrow = S, ncol = length(theta))
Sigmas.c <- matrix(nrow = S, ncol = length(as.vector(Sigma)))
```

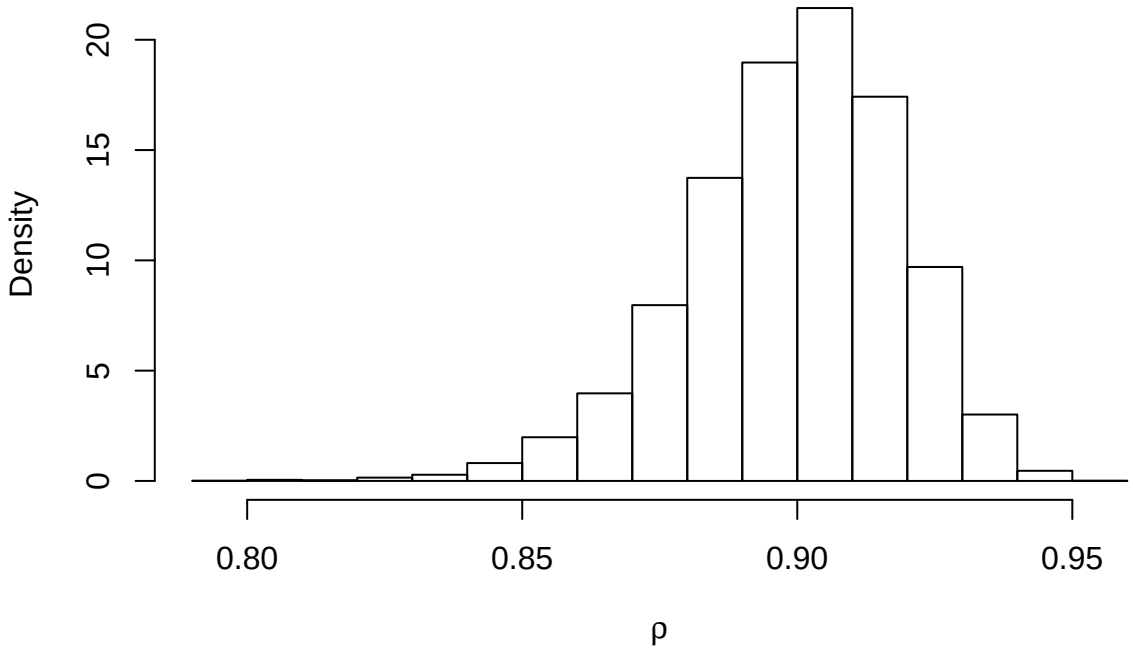
```
for (i in 1:S) {
 thetas.c[i, ] <- theta <- samp.theta(Y, Sigma, mu.0, lambda.0)
 Sigma <- samp.Sigma(Y, theta, S.0, nu.0)
 Sigmas.c[i, ] <- as.vector(Sigma)
}

post.ci.c <- rbind(quantile(thetas.c[, 1], c(0.025, 0.975)),
                   quantile(thetas.c[, 2], c(0.025, 0.975)),
                   quantile(Sigmas.c[, 2]/sqrt(Sigmas.c[, 1]*Sigmas.c[, 4]), c(0.025, 0.975)))
row.names(post.ci.c) <- c("$\\theta_h$", "$\\theta_w$", "$\\rho$")
kable(post.ci.c, digits=2)
```

	2.5%	97.5%
θ_h	41.67	46.99
θ_w	38.34	43.29
ρ	0.86	0.93

```
plot(thetas.c[, 1], thetas.c[, 2], xlab = expression(theta[h]), ylab = expression(theta[w]))
```




```
hist(Sigmas.c[, 2]/sqrt(Sigmas.c[, 1]*Sigmas.c[, 4]), xlab = expression(rho), main = "", freq = FALSE)
```



Part d.

```
mu.0 <- rep(0, 2)
lambda.0 <- 10^5*diag(2)
S.0 <- 1000*diag(2)
nu.0 <- 3

theta <- apply(Y, 2, mean)
Sigma <- cov(Y)

S <- 10000
thetas.d <- matrix(nrow = S, ncol = length(theta))
Sigmas.d <- matrix(nrow = S, ncol = length(as.vector(Sigma)))

for (i in 1:S) {
 thetas.d[i, ] <- theta <- samp.theta(Y, Sigma, mu.0, lambda.0)
 Sigma <- samp.Sigma(Y, theta, S.0, nu.0)
 Sigmas.d[i, ] <- as.vector(Sigma)
}

post.ci.d <- rbind(quantile(thetas.d[, 1], c(0.025, 0.975)),
                   quantile(thetas.d[, 2], c(0.025, 0.975)),
                   quantile(Sigmas.d[, 2]/sqrt(Sigmas.d[, 1]*Sigmas.d[, 4]), c(0.025, 0.975)))
row.names(post.ci.d) <- c("$\\theta_h$", "$\\theta_w$", "$\\rho$")
kable(post.ci.d, digits=2)
```

	2.5%	97.5%
θ_h	41.67	47.17
θ_w	38.30	43.47
ρ	0.79	0.90

Part e.

Given that we have $n = 100$, the likelihood dominates the posterior and the results from both priors are very similar. The only exception is that the posterior 95% confidence interval for ρ is slightly larger under the diffuse prior used in Part d.. If we had less data, e.g., $n = 25$, we would expect larger 95% confidence intervals under the more diffuse prior used in Part d. than under the more informative prior used in Part c..

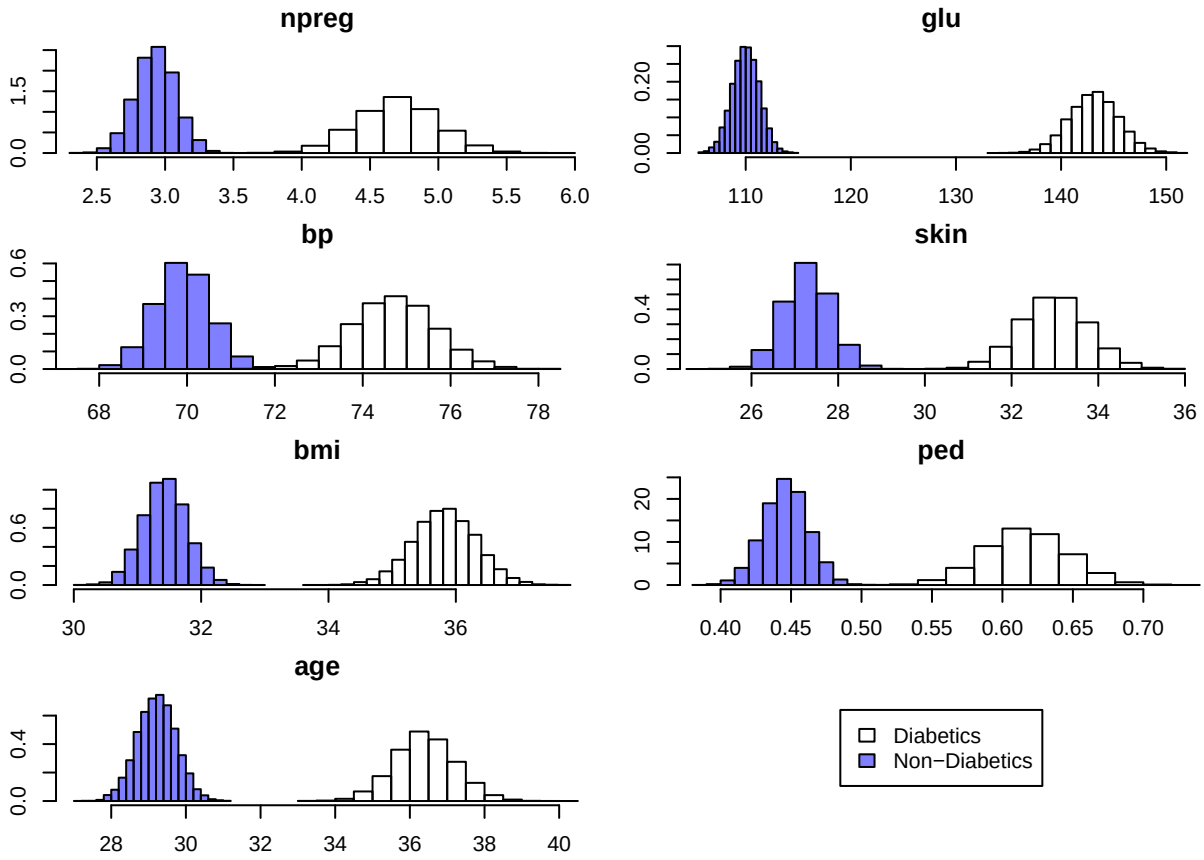
Problem 2

Part a.

```
data <- read.table("http://www.stat.washington.edu/~pdhoff/Book/Data/hwdata/azdiabetes.dat",
                   header = TRUE)
Y.n <- data[data[, "diabetes"] == "No", !names(data) == "diabetes"]
Y.d <- data[data[, "diabetes"] == "Yes", !names(data) == "diabetes"]

mu.0.n <- apply(Y.n, 2, mean)
mu.0.d <- apply(Y.d, 2, mean)
lambda.0.n <- S.0.n <- cov(Y.n)
lambda.0.d <- S.0.d <- cov(Y.d)
nu.0.d <- nu.0.n <- 9

theta.n <- apply(Y.n, 2, mean)
theta.d <- apply(Y.d, 2, mean)
Sigma.n <- cov(Y.n)
Sigma.d <- cov(Y.d)

S <- 10000
thetas.d <- matrix(nrow = S, ncol = length(theta.d))
thetas.n <- matrix(nrow = S, ncol = length(theta.n))

Sigmas.d <- matrix(nrow = S, ncol = length(as.vector(Sigma.d)))
Sigmas.n <- matrix(nrow = S, ncol = length(as.vector(Sigma.n)))
for (i in 1:S) {

    thetas.d[i, ] <- theta.d <- samp.theta(Y.d, Sigma.d, mu.0.d, lambda.0.d)
    Sigma.d <- samp.Sigma(Y.d, theta.d, S.0.d, nu.0.d)
    Sigmas.d[i, ] <- as.vector(Sigma.d)

    thetas.n[i, ] <- theta.n <- samp.theta(Y.n, Sigma.n, mu.0.n, lambda.0.n)
    Sigma.n <- samp.Sigma(Y.n, theta.n, S.0.n, nu.0.n)
    Sigmas.n[i, ] <- as.vector(Sigma.n)
}
```

```
par(mfrow = c(4, 2))
par(mar = rep(2, 4))
for (j in 1:7) {
  range <- range(cbind(thetas.d[, j], thetas.n[, j]))
  hist(thetas.n[, j], xlim = range, xlab = "", col = rgb(0, 0, 1, 0.5),
       main = names(Y.d)[j], freq = FALSE)
  hist(thetas.d[, j], xlim = range, add = TRUE,
```

```
       freq = FALSE)
}
plot(range, range, type = "n", axes = FALSE)
legend("center", fill = c("white", rgb(0, 0, 1, 0.5)),
       legend = c("Diabetics", "Non-Diabetics"))
```



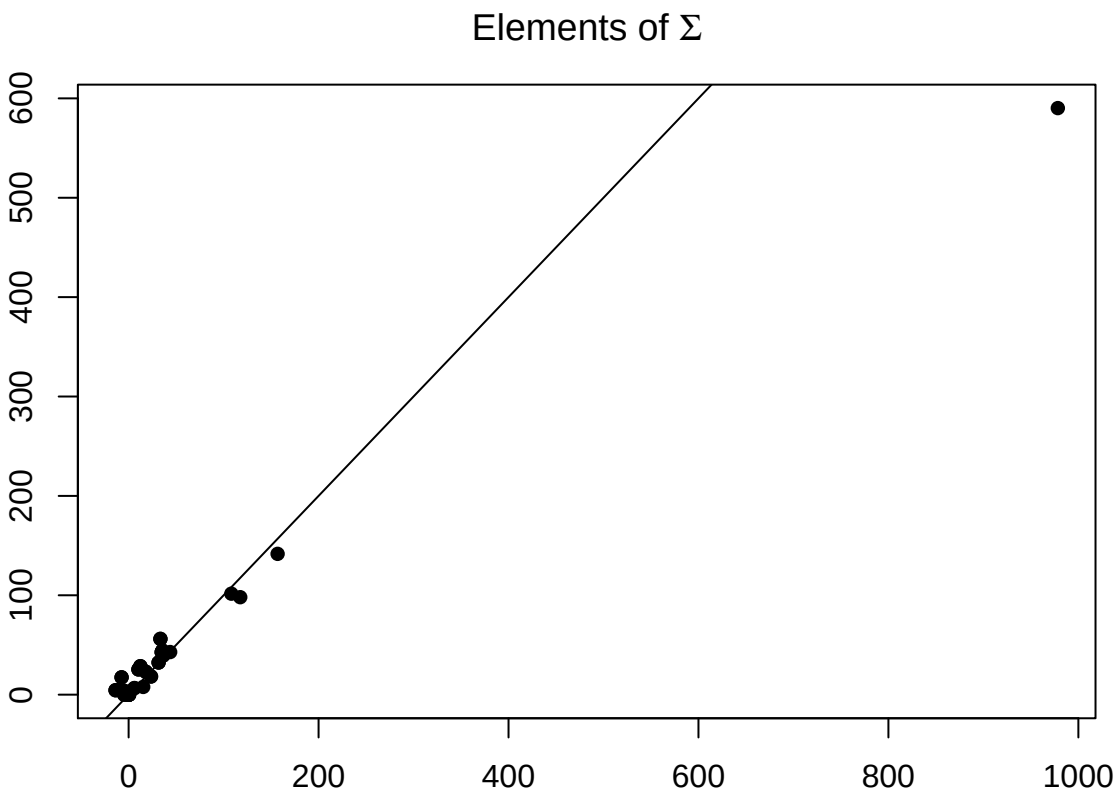
The figure shows that the means of all of the variables appear to be higher in the diabetic group, this is consistent with our comparisons of $\theta_{d,i}$ and $\theta_{n,i}$ in the table.

```
post.prob <- numeric(length(theta.d))
for (i in 1:length(post.prob)) {
  post.prob[i] <- mean(thetas.d[, i] > thetas.n[, i])
}
names(post.prob) <- paste("$\\text{Pr}\\left(\\theta_{d,", 1:7, "} > \\theta_{n,", 1:7, "}\\right)$",
                          sep = "")
kable(post.prob, digits=2)
```

$\Pr\left(\theta_{d,1} > \theta_{n,1}\right)$	1
$\Pr\left(\theta_{d,2} > \theta_{n,2}\right)$	1
$\Pr\left(\theta_{d,3} > \theta_{n,3}\right)$	1
$\Pr\left(\theta_{d,4} > \theta_{n,4}\right)$	1
$\Pr\left(\theta_{d,5} > \theta_{n,5}\right)$	1
$\Pr\left(\theta_{d,6} > \theta_{n,6}\right)$	1
$\Pr\left(\theta_{d,7} > \theta_{n,7}\right)$	1

Part b.

```
par(mfrow = c(1, 1))
par(mar = rep(3, 4))
plot(apply(Sigmas.d, 2, mean),
     apply(Sigmas.n, 2, mean),
     xlab = "Diabetics", ylab = "Non-Diabetics",
     main = expression(paste("Elements of ", Sigma, sep = "")),
     pch = 16)
abline(a = 0, b = 1)
```



The single point off the 45° line corresponds to the estimated variance of glucose levels; glucose levels are more variable among diabetic subjects than diabetic subjects. This is consistent with our expectations, given that diabetes is a disease related to glucose level regulation.

```
# Figure out which point is the outlier
  which(abs(apply(Sigmas.d, 2, mean) - apply(Sigmas.n, 2, mean)) == max(abs(apply(Sigmas.d, 2, mean) - a
```

```
## [1] 9
```