# STA 601/360 Homework 7

*Yifei Wang*

*24 October, 2019*

**1. Hoff problem 7.2 Unit information prior: Letting $\Psi = \Sigma^{-1}$, show that a unit information prior for $(\theta, \Psi)$ is given by $\theta|\Psi \sim$ multivariateNormal$(\bar{y}, \Psi^{-1})$ and $\Psi \sim$ Wishart$(p+1, S^{-1})$, where $S = \sum(y_i - \bar{y})(y_i - \bar{y})^T/n$. This can be done by mimicking the procedure outlined in Exercise 5.6 as follows:**

**(a) Reparameterize the multivariate normal model in terms of the precision matrix $\Psi = \Sigma^{-1}$. Write out the resulting log likelihood, and find a probability density $p_U(\theta, \Psi) = p_U(\theta|\Psi)p_U(\Psi)$ such that $\log p(\theta, \Psi) = l(\theta, \Psi|Y)/n + c$, where $c$ does not depend on $\theta$ or $\Psi$. Hint: Write $(y_i - \theta)$ as $(y_i - \bar{y} + \bar{y} - \theta)$, and note that $\sum a_i^T B a_i$ can be written as $\mathbf{tr}(AB)$, where $A = a_i a_i^T$.**

We first reparameterize the multivariate normal model (p-dimensional) to be

$$p(\mathbf{Y}|\theta, \Psi) = (2\pi)^{-\frac{p}{2}} \det(\Psi)^{1/2} \exp(-\frac{1}{2}\sum(\mathbf{y}_i - \theta)^T \Psi(\mathbf{y}_i - \theta))$$

and the log-likelihood of it would be

$$\begin{aligned}
l(\theta, \Psi|\mathbf{Y}) &= \log p(\mathbf{Y}|\theta, \Psi) \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{1}{2}\sum(\mathbf{y}_i - \theta)^T\Psi(\mathbf{y}_i - \theta) \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{1}{2}\sum(\mathbf{y}_i - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \theta)^T\Psi(\mathbf{y}_i - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \theta) \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{1}{2}\sum\left[(\mathbf{y}_i - \bar{\mathbf{y}})^T\Psi(\mathbf{y}_i - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \theta)^T\Psi(\bar{\mathbf{y}} - \theta)\right] \\
&\qquad - \frac{1}{2}\sum\left[(\mathbf{y}_i - \bar{\mathbf{y}})^T\Psi(\bar{\mathbf{y}} - \theta) + (\bar{\mathbf{y}} - \theta)^T\Psi(\mathbf{y}_i - \bar{\mathbf{y}})\right] \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{1}{2}\sum\left[(\mathbf{y}_i - \bar{\mathbf{y}})^T\Psi(\mathbf{y}_i - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \theta)^T\Psi(\bar{\mathbf{y}} - \theta)\right] \\
&\qquad - \frac{1}{2}\left(\sum(\mathbf{y}_i - \bar{\mathbf{y}})^T\right)\Psi(\bar{\mathbf{y}} - \theta) - \frac{1}{2}(\bar{\mathbf{y}} - \theta)^T\Psi\sum(\mathbf{y}_i - \bar{\mathbf{y}}) \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{1}{2}\sum\left[(\mathbf{y}_i - \bar{\mathbf{y}})^T\Psi(\mathbf{y}_i - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \theta)^T\Psi(\bar{\mathbf{y}} - \theta)\right] \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{1}{2}\mathrm{tr}\left(\sum(\mathbf{y}_i - \bar{\mathbf{y}})^T\Psi(\mathbf{y}_i - \bar{\mathbf{y}})\right) - \frac{1}{2}\sum(\bar{\mathbf{y}} - \theta)^T\Psi(\bar{\mathbf{y}} - \theta) \\
&= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log(|\Psi|) - \frac{n}{2}\mathrm{tr}\left(S\Psi\right) - \frac{n}{2}(\bar{\mathbf{y}} - \theta)^T\Psi(\bar{\mathbf{y}} - \theta)
\end{aligned}$$

where $S = \frac{1}{n}\sum(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$ using property of matrix trace. Then assume we have our prior $\theta|\Psi \sim$ multivariateNormal$(\theta_0, \Sigma_0)$ and $\Psi \sim$ Wishart$(n_0, \Psi_0)$ Then we could write out

$$\begin{aligned}
\log(p_U(\theta, \Psi)) &= \log(p_U(\theta|\Psi)\, p_U(\Psi)) \\
&= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(\theta - \theta_0)^T\Sigma_0^{-1}(\theta - \theta_0) \\
&\quad - \frac{n_0 p}{2}\log 2 - -\frac{n_0}{2}\log|\Psi_0| - \log(\Gamma_p(\frac{n_0}{2})) + \frac{n_0 - p - 1}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}(\Psi_0^{-1}\Psi) \\
&= -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(\theta - \theta_0)^T\Sigma_0^{-1}(\theta - \theta_0) + \frac{n_0 - p - 1}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}(\Psi_0^{-1}\Psi) + c_0
\end{aligned}$$

where $c_0 = -\frac{p}{2}\log(2\pi) - \frac{n_0 p}{2}\log 2 - -\frac{n_0}{2}\log|\Psi_0| - \log(\Gamma_p(\frac{n_0}{2}))$ does not depend on $\theta$ and $\Psi$. Comparing this equation with the log-likelihood one, we can find that $\log(p_U(\theta,\Psi)) = \frac{1}{n}l(\theta,\Psi|\mathbf{Y}) + c$ when

$$\begin{aligned}\theta_0 &= \bar{\mathbf{y}}\\ \Sigma_0 &= \Psi^{-1}\\ n_0 &= p+1\\ \Psi_0 &= S^{-1}\end{aligned}$$

This shows that the unit information prior for $(\theta,\Psi)$ is given by

$$\begin{aligned}\theta|\Psi &\sim \text{multivariateNormal}(\bar{\mathbf{y}}, \Psi^{-1})\\ \Psi &\sim \text{Wishart}(p+1, S^{-1})\end{aligned}$$

**(b) Let $p_U(\Sigma)$ be the inverse-Wishart density induced by $p_U(\Phi)$. Obtain a density $p_U(\theta,\Sigma|y_1,\ldots,y_n) \propto p_U(\theta|\Sigma)p_U(\Sigma)p(y_1,\ldots,y_n|\theta,\Sigma)$. Can this be interpreted as a posterior distribution for $\theta$ and $\Sigma$?**

We now know about $p_U(\Psi)$ that $\Psi \sim \text{Wishart}(p+1, S^{-1})$, then we could induce $p_U(\Sigma)$ that $\Sigma \sim \text{inverseWishart}(p+1, S)$. We could then obtain the density

$$\begin{aligned}p_U(\theta,\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n) &\propto p_U(\theta|\Sigma)\,p(\Sigma)\,p_U(\mathbf{y}_1,\ldots,\mathbf{y}_n|\theta,\Sigma)\\ &\propto |\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta-\bar{\mathbf{y}})^T\Sigma^{-1}(\theta-\bar{\mathbf{y}})\right)\,|\Sigma|^{-(p+1)}\exp\left(-\frac{1}{2}\text{tr}(S\Sigma^{-1})\right)\\ &\quad |\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum(\mathbf{y}_i-\theta)^T\Sigma^{-1}(\mathbf{y}_i-\theta)\right)\\ &\propto |\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\theta^T(n+1)\Sigma^{-1}\theta + \theta^T\Sigma^{-1}(n+1)\bar{\mathbf{y}})\right)\times|\Sigma|^{-\frac{2p+n+2}{2}}\exp\left(-\frac{1}{2}\text{tr}((n+1)S\Sigma^{-1})\right)\end{aligned}$$

The final line is derived by expanding all the quadratic terms in the exponential and collapsing them with $\theta$. This shows that

$$\theta,\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n \sim \text{Normal}(\bar{\mathbf{y}}, \frac{\Sigma}{n+1}) \times \text{inverseWishart}(n+p+1, \frac{S^{-1}}{n+1})$$

Thus, this is definitely a joint posterior distribution of $\theta$ and $\Sigma$.

**2. Hoff problem 7.4 Marriage data: The file `agehw.dat` contains data on the ages of 100 married couples sampled from the U.S. population.**

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
age = read.table("agehw.dat", header = TRUE)
n = dim(age)[1]
p = dim(age)[2]
```

**(a) Before you look at the data, use your own knowledge to formulate a semiconjugate prior distribution for $\theta = (\theta_h, \theta_w)^T$ and $\Sigma$, where $\theta_h$, $\theta_w$ are mean husband and wife ages, and $\Sigma$ is the covariance matrix.**

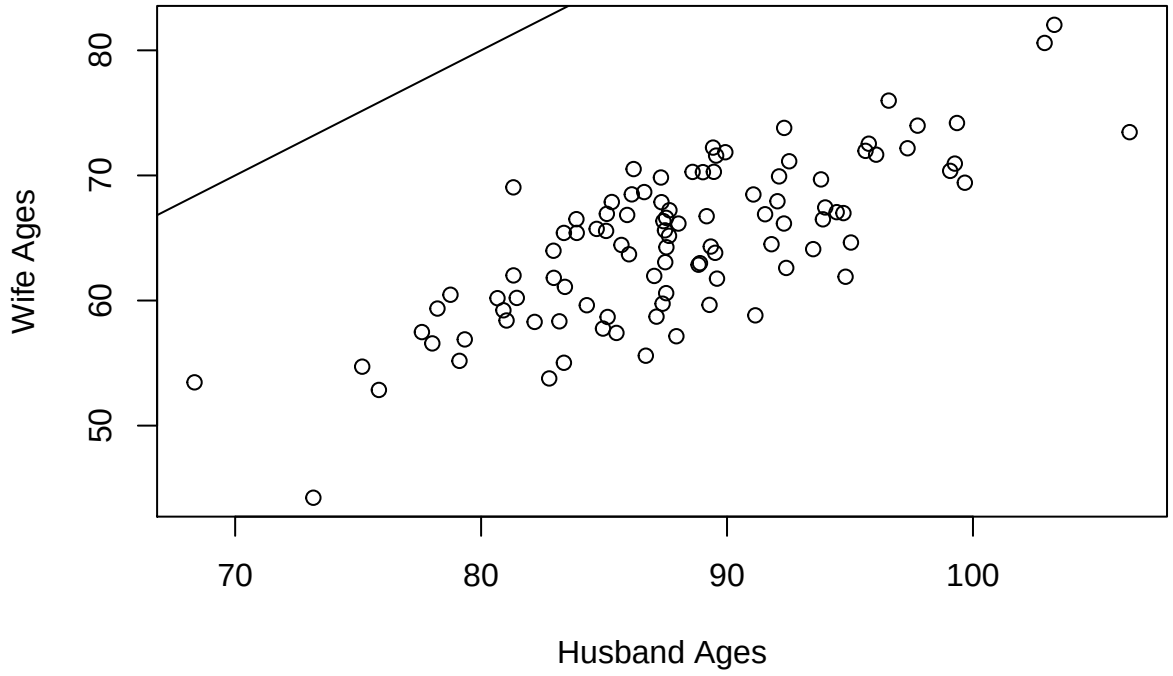
Considering the life expectacy and marriage ages, I think most of the ages will fall between $(25, 85)$. The average age of men might be higher than women, but I do not have enough evidence for this claims. Thus I would choose $\theta_0 = (55, 55)$ to be the prior mean of $\theta$. Also, most of the ages (95%) might not fall out of this range, so this gives the variance to be $((55-25)/2)^2 = 225$. I also believe that the ages of married couples are close to each other, with a correlation 0.8. This gives the covariance to be $0.8 * 225 = 180$. Thus the prior covariance matrix for $\theta$ is $\Lambda_0 = \begin{bmatrix} 225 & 180 \\ 180 & 225 \end{bmatrix}$.

For the variance $\Sigma$, the same logic of the range of ages still applies here. I will set $S_0 = \Sigma_0$, but make $\Sigma$ only loosely centered around $\Lambda_0$ by setting $\nu_0 = p + 2 = 4$.

**(b) Generate a prior predictive dataset of size $n = 100$, by sampling $(\theta, \Sigma)$ from your prior distribution and then simulating $Y_1, \ldots Y_n \sim$ i.i.d. multivariateNormal$(\theta, \Sigma)$. Generate several such datasets, make bivariate scatterplots for each dataset, and make sure they roughly represent your prior beliefs about what such a dataset would actually look like. If your prior predictive datasets do not conform to your beliefs, go back to part (a) and formulate a new prior. Report the prior that you eventually decide upon, and provide scatterplots for at least three prior predictive datasets.**

```
set.seed(4)
# priors
p = 2
theta_0 = rep(55, p)
lambda_0 = matrix(c(225, 180, 180, 225), nrow=2)
nu_0 = p + 2
S_0 = lambda_0

# prior predictive dataset
n = 100
m = 15
THETA.prior = array(dim = c(p, m))
SIGMA.prior = array(dim = c(p, p, m))
Y.prior = array(dim = c(n, p, m))

for (i in 1:m) {
  # generate
  THETA = mvrnorm(n = 1, mu = theta_0, Sigma = lambda_0)
  SIGMA = solve(rWishart(n = 1, df = nu_0, Sigma = solve(S_0))[,,1])
  Y = mvrnorm(n = n, mu = THETA, Sigma = SIGMA)

  # update
  THETA.prior[,i] = THETA
  SIGMA.prior[,,i] = SIGMA
  Y.prior[,,i] = Y
```

```r
  # visual check
  # plot(Y[,1], Y[,2], xlab = "Husband Ages", ylab = "Wife Ages")
}

# plotting
ids = sample(1:m, 3)
for (id in ids) {
  plot(Y.prior[,1,id], Y.prior[,2,id], xlab = "Husband Ages", ylab = "Wife Ages")
  abline(a = 0, b = 1)
}
```

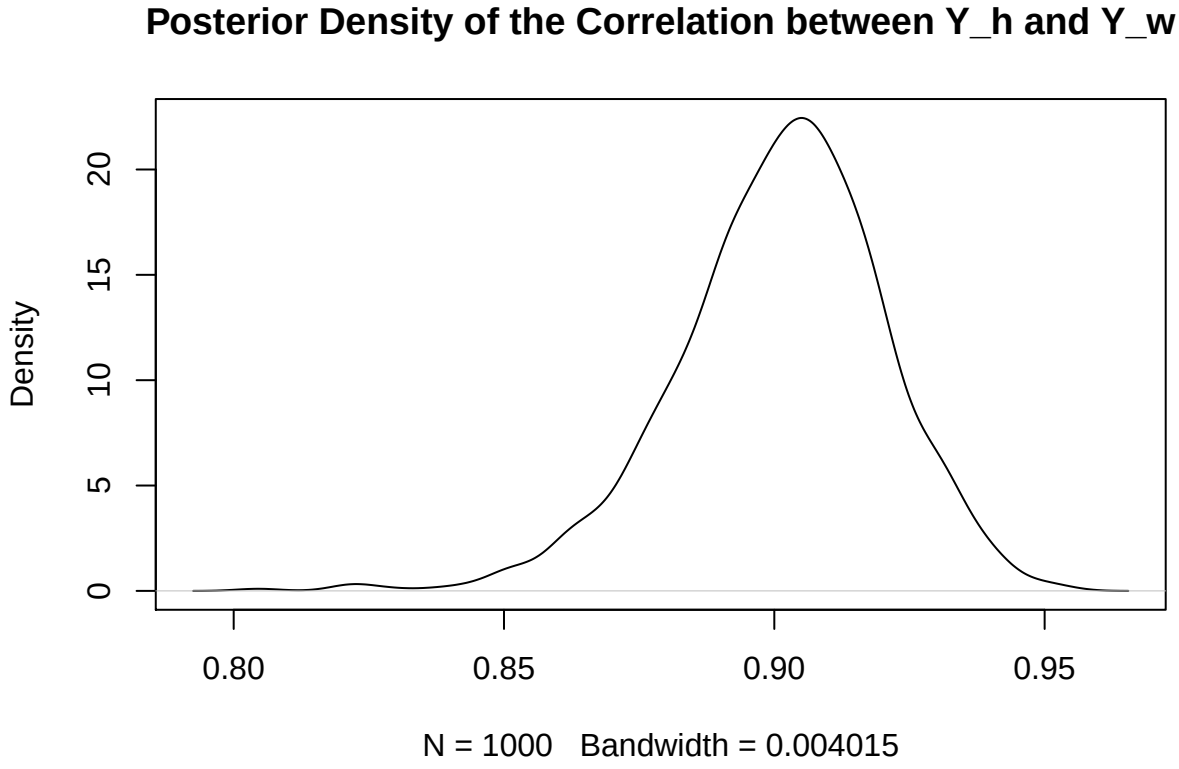

The generated prior predictive datasets do represent my prior beliefs on the real data.

**(c) Using your prior distribution and the 100 values in the dataset, obtain an MCMC approximation to $p(\theta, \Sigma | y_1, \ldots, y_{100})$. Plot the joint posterior distribution of $\theta_h$ and $\theta_w$, and also the marginal posterior density of the correlation between $Y_h$ and $Y_w$, the ages of a husband and wife. Obtain 95% posterior confidence intervals for $\theta_h$, $\theta_w$ and the correlation coefficient.**

```
# MCMC gibbs sampling
gibbs_normal = function(S, mu_0, lambda_0, nu_0, S_0, Y) {
  Y_mean = colMeans(Y)
  n = dim(Y)[1]
  p = dim(Y)[2]
  THETA.mcmc = array(dim = c(p, S))
  SIGMA.mcmc = array(dim = c(p, p, S))
  SIGMA = cov(Y)
  for (s in 1:S) {
    # sample THETA
    lambda = solve(solve(lambda_0) + n*solve(SIGMA))
    mu = lambda %*% (solve(lambda_0) %*% mu_0 + n * solve(SIGMA) %*% Y_mean)
    THETA = mvrnorm(n = 1, mu = mu, Sigma = lambda)

    # sample SIGMA
    Sn = S_0 + (t(Y) - c(THETA)) %*% t(t(Y) - c(THETA))
    SIGMA = solve(rWishart(n = 1, df = nu_0 + n, Sigma = solve(Sn))[,,1])

    # update parameters
    THETA.mcmc[,s] = THETA
    SIGMA.mcmc[,,s] = SIGMA
  }
  return(list(theta = THETA.mcmc, sigma = SIGMA.mcmc))
}
```

```r
# Posterior Analysis
normal_post_analysis = function(normal.mcmc, plotting) {
  theta.mcmc = normal.mcmc$theta
  sigma.mcmc = normal.mcmc$sigma
  corr.mcmc = sigma.mcmc[1,2,] / sqrt(sigma.mcmc[1,1,]*sigma.mcmc[2,2,])

  # plotting
  if (plotting == TRUE) {
    plot(theta.mcmc[1,], theta.mcmc[2,], xlab = "Husband Ages", ylab = "Wife Ages",
         main = "Joint Posterior Distribution of theta")
    plot(density(corr.mcmc),
         main = "Posterior Density of the Correlation between Y_h and Y_w")
  }

  # Confidence Intervals
  print("Husband Ages - theta_h")
  print(quantile(theta.mcmc[1,], probs = c(0.025, 0.975)))
  print("Wife Ages - theta_w")
  print(quantile(theta.mcmc[2,], probs = c(0.025, 0.975)))
  print("Correlation")
  print(quantile(corr.mcmc, probs = c(0.025, 0.975)))
}
```

```r
set.seed(4)
# priors
mu_0 = rep(55, p)
lambda_0 = matrix(c(225, 180, 180, 225), nrow=2)
nu_0 = p + 2
S_0 = lambda_0

S = 1000
normal.mcmc = gibbs_normal(S = S, mu_0 = mu_0, lambda_0 = lambda_0,
                           nu_0 = nu_0, S_0 = S_0, Y = age)
normal_post_analysis(normal.mcmc = normal.mcmc, plotting = TRUE)
```


```
## [1] "Husband Ages - theta_h"
##     2.5%    97.5%
## 41.73089 47.09204
## [1] "Wife Ages - theta_w"
##     2.5%    97.5%
```

```
## 38.39290 43.33363
## [1] "Correlation"
##      2.5%     97.5%
## 0.8605982 0.9349632
```

**(d) Obtain 95% posterior confidence intervals for $\theta_h$, $\theta_w$ and the correlation coefficient using the following prior distributions:**

**i. Jeffreys' prior, described in Exercise 7.1;**

```
set.seed(4)
# Jeffreys' prior
Y_mean = colMeans(age)
n = dim(age)[1]
p = dim(age)[2]
SIGMA = cov(age)

S = 1000
THETA.mcmc = array(dim = c(p, S))
SIGMA.mcmc = array(dim = c(p, p, S))
for (s in 1:S) {
  # Update theta
  THETA = mvrnorm(n = 1, Y_mean, SIGMA / n)

  # Update sigma
  S_theta = (t(age) - c(Y_mean)) %*% t(t(age) - c(Y_mean))
  SIGMA = solve(rWishart(1, n + 1, solve(S_theta))[,,1])

  THETA.mcmc[,s] = THETA
  SIGMA.mcmc[,,s] = SIGMA
}

normal_post_analysis(normal.mcmc = list(theta = THETA.mcmc, sigma = SIGMA.mcmc), plotting = F)
```

```
## [1] "Husband Ages - theta_h"
##     2.5%    97.5%
## 41.58998 47.01728
## [1] "Wife Ages - theta_w"
##     2.5%    97.5%
## 38.25191 43.24955
## [1] "Correlation"
##      2.5%     97.5%
## 0.8617078 0.9361881
```

**ii. the unit information prior, described in Exercise 7.2;**

```
set.seed(4)
# data
n = dim(age)[1]
p = dim(age)[2]
Y_mean = colMeans(age)
Y_S = cov(age) * (n-1) / n
# unit information prior
```

```
mu_0 = colMeans(age)
nu_0 = p + 1
S_0 = Y_S
SIGMA = cov(age)

S = 1000
THETA.mcmc = array(dim = c(p, S))
SIGMA.mcmc = array(dim = c(p, p, S))
for (s in 1:S) {
  # sample SIGMA
  SIGMA = solve(rWishart(n = 1, df = nu_0 + n, Sigma = solve(S_0)/(n+p+1))[,,1])

  # sample THETA
  THETA = mvrnorm(n = 1, mu = Y_mean, Sigma = SIGMA/(n+1))

  # update parameters
  THETA.mcmc[,s] = THETA
  SIGMA.mcmc[,,s] = SIGMA
}

normal_post_analysis(normal.mcmc = list(theta = THETA.mcmc, sigma = SIGMA.mcmc), plotting = FALSE)
```

```
## [1] "Husband Ages - theta_h"
##     2.5%    97.5%
## 41.59525 47.01651
## [1] "Wife Ages - theta_w"
##     2.5%    97.5%
## 38.27656 43.25030
## [1] "Correlation"
##      2.5%     97.5%
## 0.8622253 0.9359293
```

**iii. a "diffuse prior" with $\mu_0 = 0$, $\Lambda_0 = 10^5 \times \mathbf{I}$, $S_0 = 1000 \times \mathbf{I}$ and $\nu_0 = 3$.**

```
set.seed(4)
# diffuse prior
mu_0 = rep(0, 2)
lambda_0 = 10^5 * diag(c(1,1))
nu_0 = 3
S_0 = 1000 * diag(c(1,1))

S = 1000
normal.mcmc = gibbs_normal(S = S, mu_0 = mu_0, lambda_0 = lambda_0,
                           nu_0 = nu_0, S_0 = S_0, Y = age)
normal_post_analysis(normal.mcmc = normal.mcmc, plotting = FALSE)
```

```
## [1] "Husband Ages - theta_h"
##     2.5%    97.5%
## 41.56936 47.06819
## [1] "Wife Ages - theta_w"
##     2.5%    97.5%
## 38.21129 43.40227
## [1] "Correlation"
##      2.5%     97.5%
```

```
## 0.7914405 0.9028120
```

**(e) Compare the confidence intervals from (d) to those obtained in (c). Discuss whether or not you think that your prior information is helpful in estimating $\theta$ and $\Sigma$, or if you think one of the alternatives in d) is preferable. What about if the sample size were much smaller, say $n = 25$?**

I think my prior is helpful in estimating $\theta$ and $\Sigma$, but the choice of different prior does not matter much in this case. The sample size $n = 100$ is pretty large compare to the amount of information in any of those priors. The prior sample size $\nu_0 \leq 4$ for all of the priors above, which is way less informative than the data itself. Thus the results of any of those priors are quite similar.

If we have a smaller sample size, this may be different. The priors would be more informative compared to the data. This might have a greater effect on the posterior. I randomly choose 25 samples from the original data and test on my prior and the diffuse prior.

```
# data
age25 = age[sample(1:100, 25),]
n = dim(age25)[1]
p = dim(age25)[2]
```

```
set.seed(4)
# my priors
mu_0 = rep(55, p)
lambda_0 = matrix(c(225, 180, 180, 225), nrow=2)
nu_0 = p + 2
S_0 = lambda_0

S = 1000
normal.mcmc = gibbs_normal(S = S, mu_0 = mu_0, lambda_0 = lambda_0,
                           nu_0 = nu_0, S_0 = S_0, Y = age25)
normal_post_analysis(normal.mcmc = normal.mcmc, plotting = FALSE)
```

```
## [1] "Husband Ages - theta_h"
##     2.5%    97.5%
## 42.49606 53.82591
## [1] "Wife Ages - theta_w"
##     2.5%    97.5%
## 38.42282 48.95635
## [1] "Correlation"
##      2.5%     97.5%
## 0.8215099 0.9596418
```

```
set.seed(4)
# diffuse prior
mu_0 = rep(0, 2)
lambda_0 = 10^5 * diag(c(1,1))
nu_0 = 3
S_0 = 1000 * diag(c(1,1))

S = 1000
normal.mcmc = gibbs_normal(S = S, mu_0 = mu_0, lambda_0 = lambda_0,
                           nu_0 = nu_0, S_0 = S_0, Y = age25)
normal_post_analysis(normal.mcmc = normal.mcmc, plotting = FALSE)
```

```
## [1] "Husband Ages - theta_h"
```

```
##     2.5%    97.5%
## 41.48655 53.90917
## [1] "Wife Ages - theta_w"
##    2.5%   97.5%
## 37.2803 49.1968
## [1] "Correlation"
##      2.5%     97.5%
## 0.5225701 0.8850822
```

In this scenario, the effect of prior on correlation is greater when $n = 100$, especially when using the diffuse prior. Since the diffuse prior assumes that the age of husband has 0 correlation to the age of wife, that is also to say they are independent, the posterior correlation is being dragged towards 0 with a greater effect when $n = 25$.

### 3. Math problem: Conditional Gaussian Density

Verify that if

$$Y = (Y_a.Y_b) \sim N(\theta = (\theta_a, \theta_b), \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix})$$

is a multivariate normal then the conditional distribution is given by

$$Y_b|Y_a \sim N(\theta_b + \Sigma_{ba}\Sigma_{aa}^{-1}(y_a - \theta_a),\ \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})$$

To do this you need to use the matrix inverse identity given in class and compute the conditional distribution from firstprinciples. (Use the Block Matrix inversion formula: https://en.wikipedia.org/wiki/Block_matrix#Block_matrix_inversion)

Suppose we have

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_a \\ \theta_b \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix}, \quad \mathbf{\Lambda} = \mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix}$$

and according to the block matrix inversion formula we have

$$\begin{aligned}
\Lambda_{aa} &= \Sigma_{aa}^{-1} + \Sigma_{aa}^{-1}\Sigma_{ab}(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1}\Sigma_{ba}\Sigma_{aa}^{-1} \\
\Lambda_{ab} &= -\Sigma_{aa}^{-1}\Sigma_{ab}(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1} \\
\Lambda_{ba} &= -(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1}\Sigma_{ba}\Sigma_{aa}^{-1} \\
\Lambda_{bb} &= (\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1}
\end{aligned}$$

We know that

$$N(\mathbf{Y} \mid \theta, \mathbf{\Sigma}) = (2\pi)^{-p/2} \det(\mathbf{\Sigma})^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{Y} - \theta)^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \theta)\right\}$$

and by only taking into account of the exponential term we have

$$\begin{aligned}
&(\mathbf{Y} - \theta)^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \theta) \\
&= (\mathbf{Y} - \theta)^T \mathbf{\Lambda}(\mathbf{Y} - \theta) \\
&= (\mathbf{Y}_a - \theta_a)^T \mathbf{\Lambda}_{aa}(\mathbf{Y}_a - \theta_a) - (\mathbf{Y}_a - \theta_a)^T \mathbf{\Lambda}_{ab}(\mathbf{Y}_b - \theta_b) \\
&\quad - (\mathbf{Y}_b - \theta_b)^T \mathbf{\Lambda}_{ba}(\mathbf{Y}_a - \theta_a) - (\mathbf{Y}_b - \theta_b)^T \mathbf{\Lambda}_{bb}(\mathbf{Y}_b - \theta_b) \\
&= (\mathbf{Y}_a - \theta_a)^T \mathbf{\Lambda}_{aa}(\mathbf{Y}_a - \theta_a) - 2(\mathbf{Y}_a - \theta_a)^T \mathbf{\Lambda}_{ab}(\mathbf{Y}_b - \theta_b) - (\mathbf{Y}_b - \theta_b)^T \mathbf{\Lambda}_{bb}(\mathbf{Y}_b - \theta_b)
\end{aligned}$$

We note that there is a quadratic and a linear term of $\mathbf{Y}_b$ inside the exponential term. This shows that the conditional Normal distribution is also a Normal distribution $Y_{b|a} \sim N(\theta_{b|a}, \Sigma_{b|a})$. What we left is to identify the expectation $\theta_{b|a}$ and the covariance $\Sigma_{b|a}$.

The covariance is easy to identify, which is the inverse of the coefficient of the quadratic term of $\mathbf{Y}_b$. Thus we have

$$\Sigma_{b|a} = \Lambda_{bb}^{-1} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$$

As for the expectation, let's assume we have another random variable $\mathbf{z} = \mathbf{y}_b + A\mathbf{y}_a$ where $A = -\Sigma_{ba}\Sigma_{aa}^{-1}$. We first prove that $\mathbf{z}$ is independent of $\mathbf{y}_a$ by

$$\begin{aligned}
\operatorname{cov}(\mathbf{z}, \mathbf{y}_a) &= \operatorname{cov}(\mathbf{y}_b, \mathbf{y}_a) + \operatorname{cov}(A\mathbf{y}_a, \mathbf{y}_a) \\
&= \Sigma_{ba} + A\Sigma_{aa} \\
&= \Sigma_{ba} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{aa} \\
&= 0
\end{aligned}$$

Therefore $\mathbf{z}$ and $\mathbf{y}_a$ are uncorrelated and since they are jointly normal, we know they are independent. Now clearly we have $\mathbb{E}(\mathbf{z}) = \theta_b + A\theta_a$, therfore if follows that

$$\begin{aligned}
\theta_{b|a} &= \mathbb{E}(\mathbf{y}_b|\mathbf{y}_a) \\
&= \mathbb{E}(\mathbf{z} - A\mathbf{y}_a|\mathbf{y}_a) \\
&= \mathbb{E}(\mathbf{z}|\mathbf{y}_a) - \mathbb{E}(A\mathbf{y}_a|\mathbf{y}_a) \\
&= \mathbb{E}(\mathbf{z}) - A\mathbf{y}_a \\
&= \theta_b + A\theta_a - A\mathbf{y}_a \\
&= \theta_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\mathbf{y}_a - \theta_a)
\end{aligned}$$

In summary, we have

$$Y_b|Y_a \sim N(\theta_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\mathbf{y}_a - \theta_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})$$