

# Module 9: The Multivariate Normal Distribution

Rebecca C. Steorts

# Agenda

- ▶ Moving from univariate to multivariate distributions.
- ▶ The multivariate normal (MVN) distribution.
- ▶ Conjugate for the MVN distribution.
- ▶ The inverse Wishart distribution.
- ▶ Conjugate for the MVN distribution (but on the covariance matrix).
- ▶ Combining the MVN with inverse Wishart.
- ▶ See Chapter 7 (Hoff) for a review of the standard Normal density.

## Notation

Assume a matrix of covariates

$$\mathbf{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

- ▶ A column of  $\mathbf{x}$  represents a particular covariate we might be interested in, such as age of a person.
- ▶ Denote  $x_i$  as the  $i$ th **row vector** of the  $\mathbf{X}_{n \times p}$  matrix.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{ip} \\ \vdots \\ x_{ip} \end{pmatrix}$$

## Distribution of MVN

We assume that the population mean is  $\boldsymbol{\mu} = E(\mathbf{X})$  and  $\Sigma = \text{Var}(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$ , where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}.$$

## Notation

Suppose matrix  $A$  is invertible. The

$$\det(A) = \sum_{i=1}^{j=n} a_{ij} A_{ij}.$$

I recommend using the `det()` command in R.

Suppose now we have a square matrix  $H_{p \times p}$ .

$$\text{trace}(H) = \sum_i h_{ii},$$

where  $h_{ii}$  are the diagonal elements of  $H$ .

# Notation

- ▶ MVN is generalization of univariate normal.
- ▶ For the MVN, we write  $\mathbf{X} \sim \mathcal{MVN}(\boldsymbol{\mu}, \Sigma)$ .
- ▶ The  $(i, j)^{\text{th}}$  component of  $\Sigma$  is the covariance between  $X_i$  and  $X_j$  (so the diagonal of  $\Sigma$  gives the component variances).

Example:  $\text{Cov}(X_1, X_2)$  is just one element of the matrix  $\Sigma$ .

# Multivariate Normal

Just as the probability density of a scalar normal is

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right\}, \quad (1)$$

the probability density of the multivariate normal is

$$p(\vec{x}) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}. \quad (2)$$

Univariate normal is special case of the multivariate normal with a one-dimensional mean “vector” and a one-by-one variance “matrix.”

# Standard Multivariate Normal Distribution

Consider

$$Z_1, \dots, Z_n \stackrel{iid}{\sim} MVN(0, I)$$

$$f_z(z) = \prod_{i=1}^n \frac{1}{2\pi} e^{-z_i^2/2} \quad (3)$$

$$= (2\pi)^{-n} e^{z^T z/2} \quad (4)$$

- ▶  $E[Z] = 0$
- ▶  $\text{Var}[Z] = I$



## Conjugate to MVN

Suppose that

$$X_1 \dots X_n \stackrel{iid}{\sim} MVN(\theta, \Sigma).$$

Let

$$\pi(\boldsymbol{\theta}) \sim MVN(\boldsymbol{\mu}, \Omega).$$

What is the full conditional distribution of  $\boldsymbol{\theta} \mid \mathbf{X}, \Sigma$ ?

## Prior

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-p/2} \det \Omega^{-1/2} \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \right\} \quad (5)$$

$$\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \right\} \quad (6)$$

$$\propto \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \Omega^{-1} \boldsymbol{\mu} \right\} \quad (7)$$

$$\propto \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\mu} \right\} \quad (8)$$

$$= \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T A_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o \right\} \quad (9)$$

$\pi(\boldsymbol{\theta}) \sim MVN(\boldsymbol{\mu}, \Omega)$  implies that  $A_o = \Omega^{-1}$  and  $b_o = \Omega^{-1} \boldsymbol{\mu}$ .

## Likelihood

$$p(\mathbf{X} \mid \boldsymbol{\theta}, \Sigma) = \prod_{i=1}^n (2\pi)^{-p/2} \det \Sigma^{-n/2} \exp \left\{ -\frac{1}{2} (x_i - \boldsymbol{\theta})^T \Sigma^{-1} (x_i - \boldsymbol{\theta}) \right\} \quad (10)$$

$$\propto \exp -\frac{1}{2} \left\{ \sum_i x_i^T \Sigma^{-1} x_i - 2 \sum_i \boldsymbol{\theta}^T \Sigma^{-1} x_i + \sum_i \boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta} \right\} \quad (11)$$

$$\exp -\frac{1}{2} \left\{ -2 \boldsymbol{\theta}^T \Sigma^{-1} n \bar{x} + n \boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta} \right\} \quad (12)$$

$$\exp -\frac{1}{2} \left\{ -2 \boldsymbol{\theta}^T b_1 + \boldsymbol{\theta}^T A_1 \boldsymbol{\theta} \right\}, \quad (13)$$

where

$$b_1 = \Sigma^{-1} n \bar{x}, \quad A_1 = n \Sigma^{-1}$$

and

$$\bar{x} := \left( \frac{1}{n} \sum_i x_{i1}, \dots, \frac{1}{n} \sum_i x_{ip} \right)^T.$$

## Full conditional

$$p(\boldsymbol{\theta} \mid \mathbf{X}, \Sigma) \propto p(\mathbf{X} \mid \boldsymbol{\theta}, \Sigma) \times p(\boldsymbol{\theta}) \quad (14)$$

$$\propto \exp -\frac{1}{2} \left\{ -2\boldsymbol{\theta}^T b_1 + \boldsymbol{\theta}^T A_1 \boldsymbol{\theta} \right\} \quad (15)$$

$$\times \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T A_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o \right\} \quad (16)$$

$$\propto \exp \left\{ \boldsymbol{\theta}^T b_1 - \frac{1}{2} \boldsymbol{\theta}^T A_1 \boldsymbol{\theta} - \frac{1}{2} \boldsymbol{\theta}^T A_o \boldsymbol{\theta} + \boldsymbol{\theta}^T b_o \right\} \quad (17)$$

$$\propto \exp \left\{ \boldsymbol{\theta}^T (b_o + b_1) - \frac{1}{2} \boldsymbol{\theta}^T (A_o + A_1) \boldsymbol{\theta} \right\} \quad (18)$$

Then

$$A_n = A_o + A_1 = \Omega^{-1} + n\Sigma^{-1}$$

and

$$b_n = b_o + b_1 = \Omega^{-1} \mu + \Sigma^{-1} n\bar{x}$$

$$\boldsymbol{\theta} \mid \mathbf{X}, \Sigma \sim MVN(A_n^{-1} b_n, A_n^{-1}) = MVN(\mu_n, \Sigma_n)$$

# Interpretations

$$\theta \mid \mathbf{X}, \Sigma \sim MVN(A_n^{-1}b_n, A_n^{-1}) = MVN(\mu_n, \Sigma_n)$$

$$\mu_n = A_n^{-1}b_n = [\Omega^{-1} + n\Sigma^{-1}]^{-1}(b_o + b_1) \quad (19)$$

$$= [\Omega^{-1} + n\Sigma^{-1}]^{-1}(\Omega^{-1}\mu + \Sigma^{-1}n\bar{x}) \quad (20)$$

$$\Sigma_n = A_n^{-1} = [\Omega^{-1} + n\Sigma^{-1}]^{-1} \quad (21)$$

## inverse Wishart distribution

Suppose  $\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1})$  where  $\nu_o$  is a scalar and  $S_o^{-1}$  is a matrix.

Then

$$p(\Sigma) \propto \det(\Sigma)^{-(\nu_o+p+1)/2} \times \exp\{-\text{tr}(S_o \Sigma^{-1})/2\}$$

For the full distribution, see Hoff, Chapter 7 (p. 110).

## inverse Wishart distribution

- ▶ The inverse Wishart distribution is the multivariate version of the Gamma distribution.
- ▶ The full hierarchy we're interested in is

$$\mathbf{X} \mid \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}, \Sigma).$$

$$\boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Omega})$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

We first consider the conjugacy of the MVN and the inverse Wishart, i.e.

$$\mathbf{X} \mid \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}, \Sigma).$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

## Continued

What about  $p(\Sigma \mid \mathbf{X}, \boldsymbol{\theta}) \propto p(\Sigma) \times p(\mathbf{X} \mid \boldsymbol{\theta}, \Sigma)$ . Let's first look at

$$p(\mathbf{X} \mid \boldsymbol{\theta}, \Sigma) \tag{22}$$

$$\propto \det(\Sigma)^{-n/2} \exp\left\{-\sum_i (\mathbf{X}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\theta})/2\right\} \tag{23}$$

$$\propto \det(\Sigma)^{-n/2} \exp\left\{-\text{tr}\left(\sum_i (\mathbf{X}_i - \boldsymbol{\theta})(\mathbf{X}_i - \boldsymbol{\theta})^T \Sigma^{-1}/2\right)\right\} \tag{24}$$

$$\propto \det(\Sigma)^{-n/2} \exp\left\{-\text{tr}(S_{\boldsymbol{\theta}} \Sigma^{-1}/2)\right\} \tag{25}$$

where  $S_{\boldsymbol{\theta}} = \sum_i (\mathbf{X}_i - \boldsymbol{\theta})(\mathbf{X}_i - \boldsymbol{\theta})^T$ .

Fact:

$$\sum_k b_k^T A b_k = \text{tr}(B^T B A),$$

where B is the matrix whose  $k$ th row is  $b_k$ .



## Continued

Now we can calculate  $p(\Sigma \mid \mathbf{X}, \theta)$

$$p(\Sigma \mid \mathbf{X}, \theta) \tag{26}$$

$$= p(\Sigma) \times p(\mathbf{X} \mid \theta, \Sigma) \tag{27}$$

$$\propto \det(\Sigma)^{-(\nu_o + p + 1)/2} \times \exp\{-\text{tr}(S_o \Sigma^{-1})/2\} \tag{28}$$

$$\times \det(\Sigma)^{-n/2} \exp\{-\text{tr}(S_\theta \Sigma^{-1})/2\} \tag{29}$$

$$\propto \det(\Sigma)^{-(\nu_o + n + p + 1)/2} \exp\{-\text{tr}((S_o + S_\theta) \Sigma^{-1})/2\} \tag{30}$$

This implies that

$$\Sigma \mid \mathbf{X}, \theta \sim \text{inverseWishart}(\nu_o + n, [S_o + S_\theta]^{-1} =: S_n)$$

## Continued

Suppose that we wish now to take

$$\boldsymbol{\theta} \mid \mathbf{X}, \Sigma \sim \text{MVN}(\mu_n, \Sigma_n)$$

(which we finished an example on earlier). Now let

$$\Sigma \mid \mathbf{X}, \boldsymbol{\theta} \sim \text{inverseWishart}(\nu_n, S_n^{-1})$$

There is no closed form expression for this posterior. Solution?

# Gibbs sampler

Suppose the Gibbs sampler is at iteration  $s$ .

1. Sample  $\boldsymbol{\theta}^{(s+1)}$  from it's full conditional:
  - a) Compute  $\mu_n$  and  $\Sigma_n$  from  $\mathbf{X}$  and  $\Sigma^{(s)}$
  - b) Sample  $\boldsymbol{\theta}^{(s+1)} \sim \text{MVN}(\mu_n, \Sigma_n)$
2. Sample  $\Sigma^{(s+1)}$  from its full conditional:
  - a) Compute  $S_n$  from  $\mathbf{X}$  and  $\theta^{(s+1)}$
  - b) Sample  $\Sigma^{(s+1)} \sim \text{inverseWishart}(\nu_n, S_n^{-1})$

# Working with Multivariate Normal Distribution

The R package, `mvtnorm`, contains functions for evaluating and simulating from a multivariate normal density.

```
library(mvtnorm)
```

```
## Warning: package 'mvtnorm' was built under R version 3.4
```

## Simulating Data

Simulate a single multivariate normal random vector using the `rmvnorm` function.

```
rmvnorm(n = 1, mean = rep(0, 2), sigma = diag(2))
```

```
##           [,1]      [,2]  
## [1,] -1.5098  1.344589
```

# Evaluation

Evaluate the multivariate normal density at a single value using the `dmvnorm` function.

```
dmvnorm(rep(0, 2), mean = rep(0, 2), sigma = diag(2))
```

```
## [1] 0.1591549
```

## Working with the Multivariate Normal

- ▶ Now let's simulate many multivariate normals.
- ▶ Each row is a different sample from this multivariate normal distribution.

```
rmvnorm(n = 3, mean = rep(0, 2), sigma = diag(2))
```

```
##           [,1]      [,2]  
## [1,] -1.0581411  1.0931620  
## [2,] -0.4217665 -0.6840540  
## [3,]  0.1649000  0.7576422
```

# Evaluation

We can evaluate the multivariate normal density at several values using the `dmvnorm` function.

```
dmvnorm(rbind(rep(0, 2), rep(1, 2), rep(2, 2)),  
        mean = rep(0, 2), sigma = diag(2))
```

```
## [1] 0.159154943 0.058549832 0.002915024
```



## Work with the Wishart density

- ▶ The R package, `stats`, contains functions for evaluating and simulating from a Wishart density.
- ▶ We can simulate a single Wishart distributed matrix using the `rWishart` function.

```
nu0 <- 2  
Sigma0 <- diag(2)  
rWishart(1, df = nu0, Sigma = Sigma0)[, , 1]
```

```
##           [,1]      [,2]  
## [1,] 0.2059065 0.520885  
## [2,] 0.5208850 1.322016
```

## inverse Wishart simulation

We can simulate a single inverse-Wishart distributed matrix using the `rWishart` function as well.

```
nu0 <- 2
Sigma0 <- diag(2)
solve(rWishart(1, df = nu0,
              Sigma = solve(Sigma0))[, , 1])
```

```
##           [,1]      [,2]
## [1,] 1.4586243 0.3195855
## [2,] 0.3195855 0.7997427
```

## An Application to Reading Comprehension

We will follow an example from Hoff (Section 7.4, p. 112).

A sample of 22 children are given reading comprehension tests before and after receiving a particular instructional method.

Each student  $i$  will then have two scores,  $Y_{i,1}$  and  $Y_{i,2}$  denoting the pre- and post-instructional scores respectively.

Denote each student's pair of scores  $\mathbf{Y}_i$

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on first test} \\ \text{score on second test} \end{pmatrix}$$

## Model set up

$$\mathbf{Y}_i \mid \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}_i, \Sigma).$$

$$\boldsymbol{\theta}_i \sim \text{MVN}(\boldsymbol{\mu}_0, \Lambda_0)$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

Let  $\theta_i = (\theta_1, \theta_2)$ .

## Prior settings

$$\mathbf{Y}_i \mid \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}_i, \Sigma).$$

$$\boldsymbol{\theta}_i \sim \text{MVN}(\boldsymbol{\mu}_0, \Lambda_0)$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

The exam was designed to give average scores of around 50 out of 100, so  $\boldsymbol{\mu}_0 = (50, 50)^T$  would be a good choice for our prior mean.

## Prior settings

$$\mathbf{Y}_i \mid \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}_i, \Sigma).$$

$$\boldsymbol{\theta}_i \sim \text{MVN}(\boldsymbol{\mu}_0, \Lambda_0)$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

Since the true mean cannot be below 0 or above 100, we will use a prior variance that puts little probability outside of this range.

We'll take the prior variances on  $\theta_1$  and  $\theta_2$  to be

$$\lambda_{0,1}^2 = \lambda_{0,2}^2 = (50/2)^2 = 625$$

so that the prior probability that  $P(\theta_j \notin [0, 100]) = 0.05$ .

The two exams are measuring similar things, so we will take the prior correlation of 0.5 or rather  $\lambda_{1,2} = 625/2 = 312.5$

## Prior settings (continued)

$$\mathbf{Y}_i \mid \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}_i, \Sigma).$$

$$\boldsymbol{\theta}_i \sim \text{MVN}(\boldsymbol{\mu}_0, \Lambda_0)$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

What about the prior settings for  $\Sigma$ ?

We take  $S_o$  to be about the same as  $\Lambda_o$ .

We will center  $\Sigma$  around  $S_o$  by setting  $\nu_0 = p + 2 = 4$ .

## Load in data

```
# read in data
Y <- structure(c(59, 43, 34, 32, 42, 38, 55, 67, 64,
                 45, 49, 72, 34, 70, 34, 50, 41, 52,
                 60, 34, 28, 35, 77, 39, 46, 26, 38,
                 43, 68, 86, 77, 60, 50, 59, 38, 48,
                 55, 58, 54, 60, 75, 47, 48, 33),
               .Dim = c(22L, 2L), .Dimnames = list(NULL,
               c("pretest", "posttest")))
# number of observations
```

Quick calculations

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

```
## [1] 22
```

```
(ybar <- apply(Y,2,mean))
```



## Application to reading comprehension

```
# set hyper-parameters  
mu0 <- c(50,50)  
L0 <- matrix(c(625,312.5,312.5,625),nrow=2)  
nu0 <- 4  
S0 <- L0
```

## Gibbs sampler

```
## Warning: package 'MCMCpack' was built under R version 3.0.2

## Loading required package: coda

## Loading required package: MASS

## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)

## ## Copyright (C) 2003-2018 Andrew D. Martin, Kevin M. Quinn

## ##
## ## Support provided by the U.S. National Science Foundation

## ## (Grants SES-0350646 and SES-0350613)
## ##
```

# Gibbs sampler

```
THETA <- SIGMA <- NULL
set.seed(1)
for (s in 1:5000) {

  ## update theta
  Ln <- solve(solve(L0) + n*solve(Sigma))
  mun <- Ln %*% (solve(L0) %*% mu0 + n*solve(Sigma) %*% yba)
  theta <- rmvnorm(1, mun, Ln)

  ## update Sigma
  Sn <- S0 + (t(Y) - c(theta)) %*% t(t(Y)-c(theta))

  Sigma <- solve(rwish(nu0 + n, solve(Sn)))
  ## save results
  THETA <- rbind(THETA, theta)
  SIGMA <- rbind(SIGMA, c(Sigma))
}
```

## Posterior inference

Using the samples from the Gibbs sampler, we have generated 5,000 samples

$$(\theta^1, \textit{Sigma}(1), \dots, \theta^1, \textit{Sigma}(1))$$

that approximates  $p(\theta, \Sigma \mid y_1, \dots, y_n)$ .

## Glance at Gibbs sampler

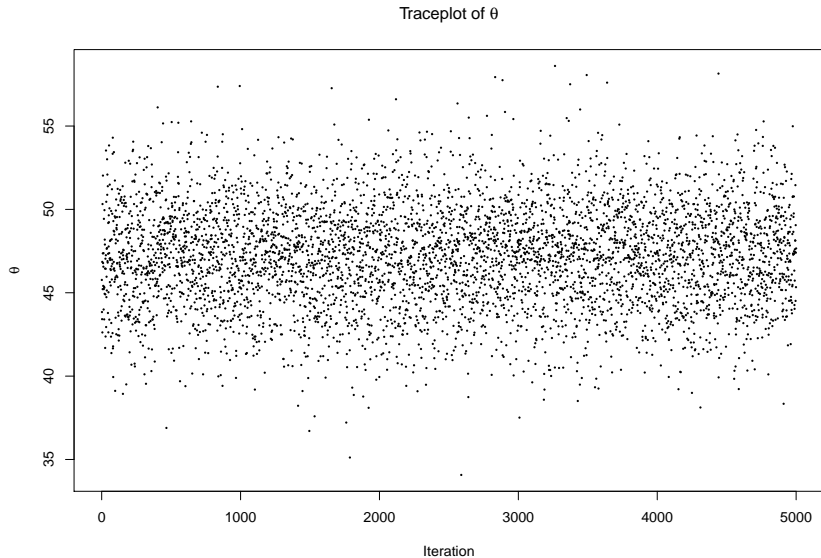
```
head(THETA)
```

```
##           [,1]      [,2]  
## [1,] 45.76871 53.64765  
## [2,] 43.84243 51.80471  
## [3,] 43.41651 51.30521  
## [4,] 46.85067 50.64238  
## [5,] 42.62048 53.71350  
## [6,] 50.32035 58.93397
```

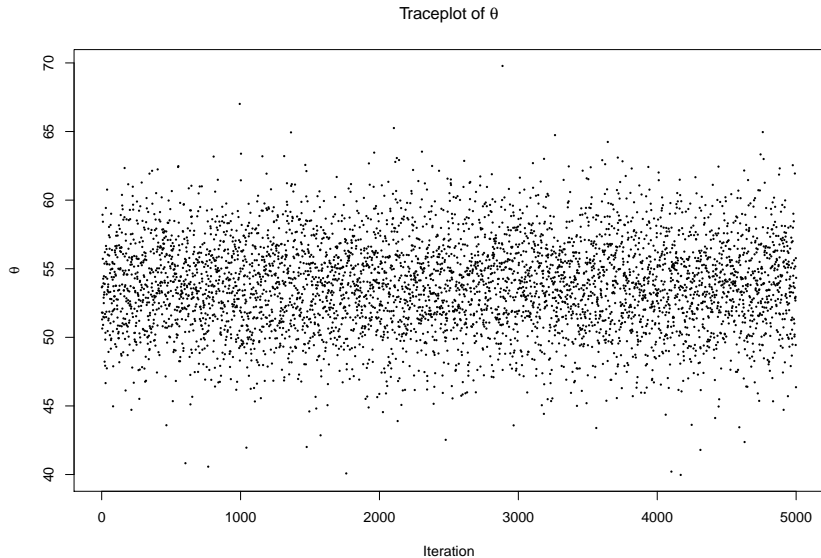
```
head(SIGMA)
```

```
##           [,1]      [,2]      [,3]      [,4]  
## [1,] 270.7381 175.9276 175.9276 213.0155  
## [2,] 237.3720 191.0999 191.0999 266.0570  
## [3,] 245.6029 183.9140 183.9140 248.4452  
## [4,] 169.6788 114.1658 114.1658 200.8390  
## [5,] 247.0899 197.0802 197.0802 295.1981
```

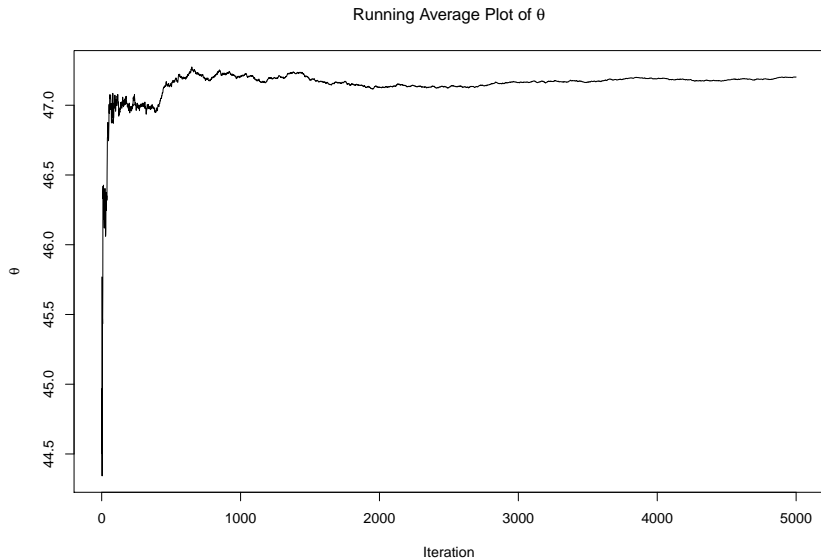
# Traceplot of $\theta_1$



## Traceplot of $\theta_2$

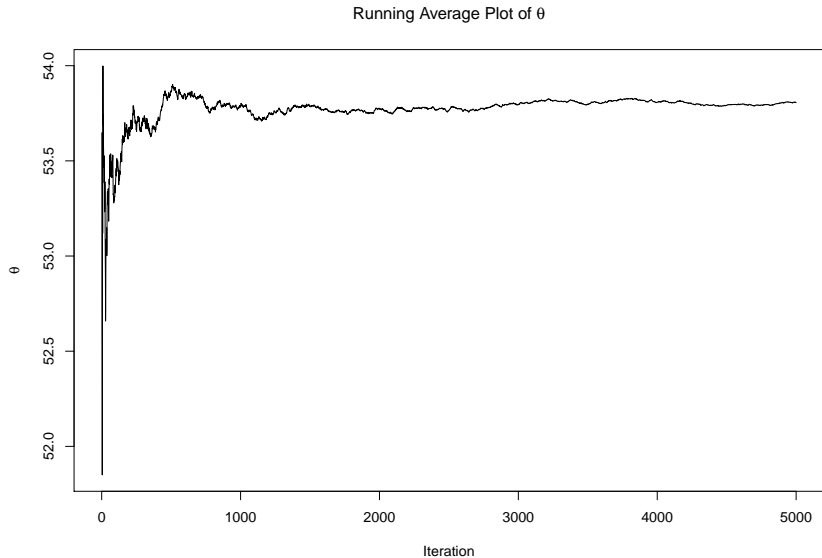


# Running average plot of $\theta_1$

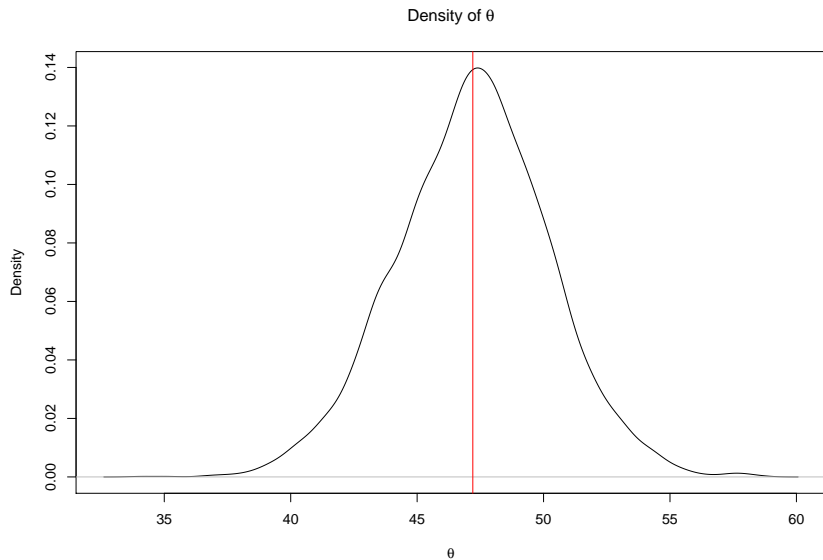




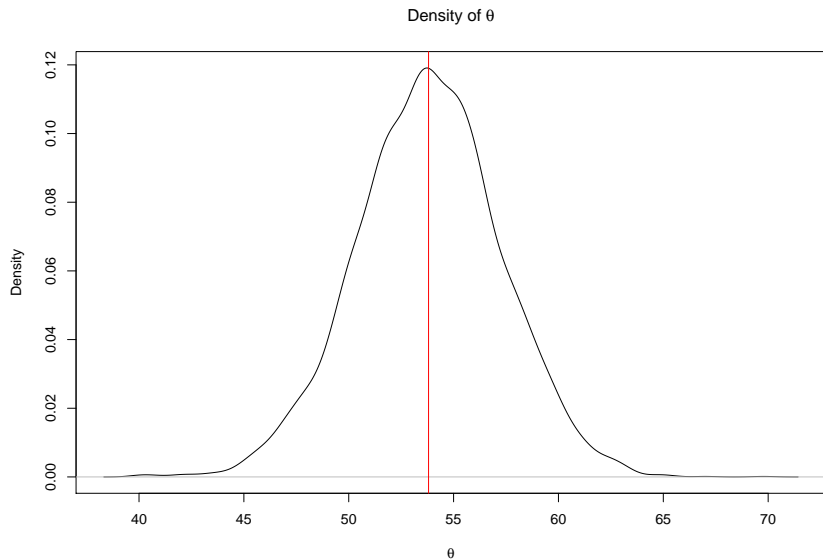
# Running average plot of $\theta_2$



# Estimated density of $\theta_1$



## Estimated density of $\theta_2$



## Traceplots and running average plots of $\sigma$

Examine the trace plots and running average plots of  $\Sigma$  on your own.

## Return to posterior inference

Given our samples from our Gibbs sampler, we can approximate posterior probabilities and confidence regions.

## Confidence regions

```
quantile(THETA[,2] - THETA[,1], prob=c(0.025,0.5,0.975))
```

```
##          2.5%          50%          97.5%  
## 1.356260  6.614818 11.667128
```

## Posterior inference

Suppose we were to give the exams/instruction to a large population, then would the average score on the second exam be higher than the first second?

We can quantify this by calculating

$$Pr(\theta_2 > \theta_1 \mid y_1, \dots, y_n) = 0.99$$

```
mean(THETA[,2] > THETA[,1])
```

```
## [1] 0.9926
```