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

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

- A column of x represents a particular covariate we might be interested in, such as age of a person.
- Denote x_i as the ith row vector of the $X_{n\times p}$ matrix.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Distribution of MVN

We assume that the population mean is $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma = \text{Var}(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$, 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() commend 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 $\boldsymbol{X} \sim \mathcal{MVN}(\boldsymbol{\mu}, \Sigma)$.
- The $(i, j)^{\text{th}}$ component of Σ is the covariance between X_i and X_j (so the diagonal of Σ gives the component variances).

Example: $Cov(X_1, X_2)$ is just one element of the matrix Σ.

Multivariate Normal

Just as the probability density of a scalar normal is

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

the probability density of the multivariate normal is

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

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

$$Z_1, \ldots, Z_n \overset{iid}{\sim} MVN(0, I)$$

$$\begin{aligned} f_z(z) &= \prod_{i=1}^{n} \frac{1}{2\pi} e^{-z_i^2/2} && (3) \\ &= (2\pi)^{-n} e^{z^T z/2} && (4) \end{aligned}$$

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

Conjugate to MVN

Suppose that

$$X_1 \ldots X_n \mid \theta \overset{iid}{\sim} MVN(\theta, \Sigma).$$

Let

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

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

Prior

$$
\begin{aligned}
\pi(\boldsymbol{\theta}) &= (2\pi)^{-p/2} \det \Omega^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} && (5) \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} && (6) \\
&\propto \exp -\frac{1}{2}\left\{\boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Omega^{-1} \mu + \mu^T \Omega^{-1} \mu\right\} && (7) \\
&\propto \exp -\frac{1}{2}\left\{\boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Omega^{-1} \mu\right\} && (8) \\
&= \exp -\frac{1}{2}\left\{\boldsymbol{\theta}^T A_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o\right\} && (9)
\end{aligned}
$$

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

Likelihood

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

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

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

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

where

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

and

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

Full conditional

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{X}, \Sigma) &\propto p(\mathbf{X} \mid \boldsymbol{\theta}, \Sigma) \times p(\boldsymbol{\theta}) && (14)\\
&\propto \exp -\frac{1}{2}\Big\{-2\boldsymbol{\theta}^T b_1 + \boldsymbol{\theta}^T A_1 \boldsymbol{\theta}\Big\} && (15)\\
&\times \exp -\frac{1}{2}\Big\{\boldsymbol{\theta}^T A_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o\Big\} && (16)\\
&\propto \exp\{\boldsymbol{\theta}^T b_1 - \frac{1}{2}\boldsymbol{\theta}^T A_1 \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T A_o \boldsymbol{\theta} + \boldsymbol{\theta}^T b_o\} && (17)\\
&\propto \exp\{\boldsymbol{\theta}^T (b_o + b_1) - \frac{1}{2}\theta^T (A_o + A_1)\theta\} && (18)
\end{aligned}
$$

Full conditional

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

Using the kernel of the multivariate normal, we can now find the posterior mean and the posterior covariance:

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

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

$$\begin{aligned} \mu_n = A_n^{-1} b_n &= [\Omega^{-1} + n\Sigma^{-1}]^{-1}(b_o + b_1) && (19) \\ &= [\Omega^{-1} + n\Sigma^{-1}]^{-1}(\Omega^{-1}\mu + \Sigma^{-1} n\bar{x}) && (20) \end{aligned}$$

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

inverse Wishart distribution

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

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

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

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

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

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

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

Continued

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

$$
\begin{aligned}
& p(\boldsymbol{X} \mid \boldsymbol{\theta}, \Sigma) && (22)\\
& \propto \det(\Sigma)^{-n/2} \exp\{-\sum_i (\mathbf{X}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\theta})/2\} && (23)\\
& \propto \det(\Sigma)^{-n/2} \exp\{-tr(\sum_i (\mathbf{X}_i - \boldsymbol{\theta})(\mathbf{X}_i - \boldsymbol{\theta})^T \Sigma^{-1}/2)\} && (24)\\
& \propto \det(\Sigma)^{-n/2} \exp\{-\mathrm{tr}(S_\theta \Sigma^{-1}/2)\} && (25)
\end{aligned}
$$

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

Fact:

$$\sum_k b_k^T A b_k = tr(BB^T A),$$

where B is the matrix whose kth row is b_k.

Continued

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

$$
\begin{aligned}
& p(\Sigma \mid \boldsymbol{X}, \boldsymbol{\theta}) && (26)\\
& \quad = p(\Sigma) \times p(\boldsymbol{X} \mid \boldsymbol{\theta}, \Sigma) && (27)\\
& \quad \propto \det(\Sigma)^{-(\nu_o + p + 1)/2} \times \exp\{-\mathrm{tr}(S_o \Sigma^{-1})/2\} && (28)\\
& \qquad \times \det(\Sigma)^{-n/2} \exp\{-\mathrm{tr}(S_\theta \Sigma^{-1})/2\} && (29)\\
& \quad \propto \det(\Sigma)^{-(\nu_o + n + p + 1)/2} \exp\{-\mathrm{tr}((S_o + S_\theta)\Sigma^{-1})/2\} && (30)
\end{aligned}
$$

This implies that

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

Continued

Suppose that we wish now to take

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

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

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

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 μ_n and Σ_n from $\boldsymbol{X}$ and $\Sigma^{(s)}$
   b) Sample $\boldsymbol{\theta}^{(s+1)} \sim MVN(\mu_n, \Sigma_n)$
2. Sample $\Sigma^{(s+1)}$ from its full conditional:
   a) Compute S_n from $\boldsymbol{X}$ 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,] 1.692983 1.726209
```

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.

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

```
##              [,1]        [,2]
## [1,] -1.31973359  0.32891729
## [2,] -0.05824105 -1.02829590
## [3,] -0.18654563  0.06951028
```

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.
- Each row is a different sample from the Wishart distribution.

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

```
##           [,1]      [,2]
## [1,]  3.479015 -2.228020
## [2,] -2.228020  1.445725
```

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 $\boldsymbol{Y}_i$

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

Model set up

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

$$\boldsymbol{\theta}_i \sim MVN(\boldsymbol{\mu_0}, \Lambda_0)$$

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

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

Prior settings

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

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

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

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

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

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

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

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 θ_1 and θ_2 to be

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

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

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)

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

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

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

What about the prior settings for Σ?

We take S_o to be about the same as Λ_o.

We will center Σ 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

```r
# 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.

## Loading required package: coda

## Loading required package: MASS

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

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

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

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

Gibbs sampler

```r
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) %*% ybar)
  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

$$(\boldsymbol{\theta}^{(1)}, \Sigma^{(1)}, \ldots, \boldsymbol{\theta}^{(5000)}, \Sigma^{(5000)})$$

that approxmiates $p(\boldsymbol{\theta}, \Sigma \mid y_1, \ldots, y_n)$.

Glance at Gibbs sampler

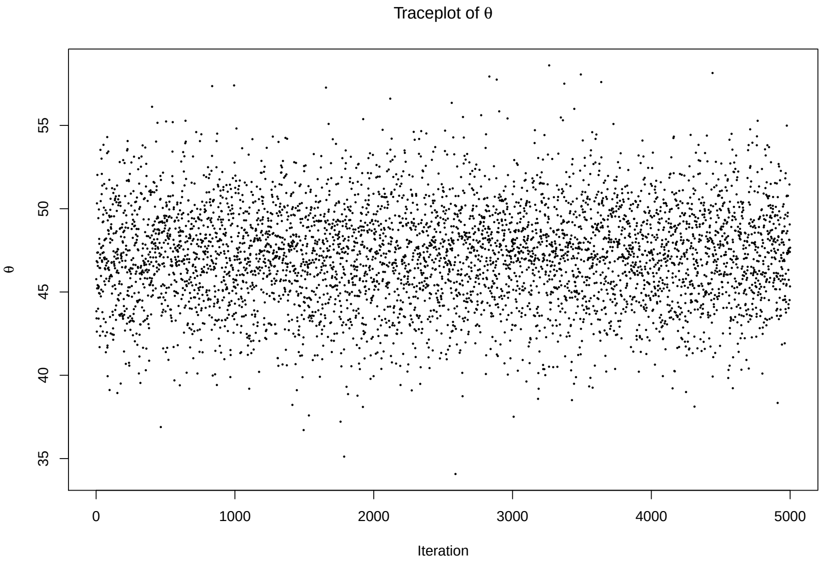
```r
head(THETA)
```

```
##          [,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
```

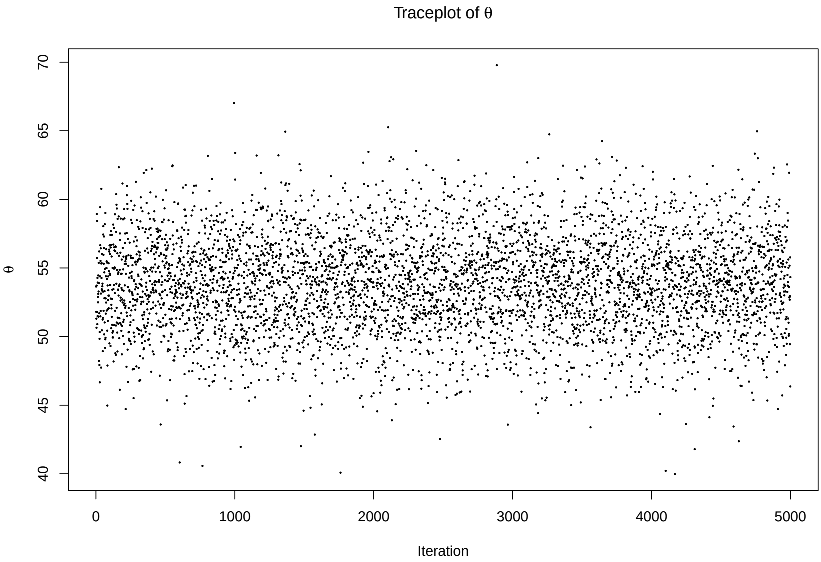
```r
head(SIGMA)
```

```
##          [,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
```

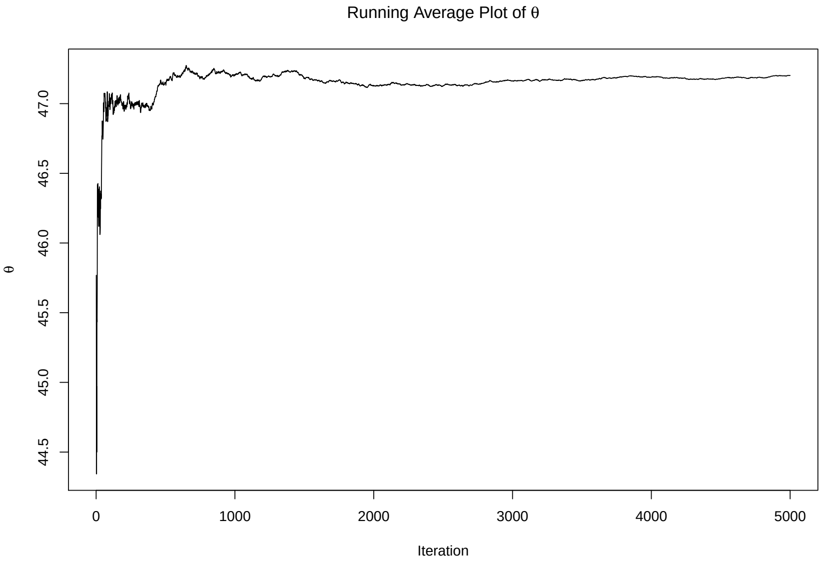
Traceplot of θ_1



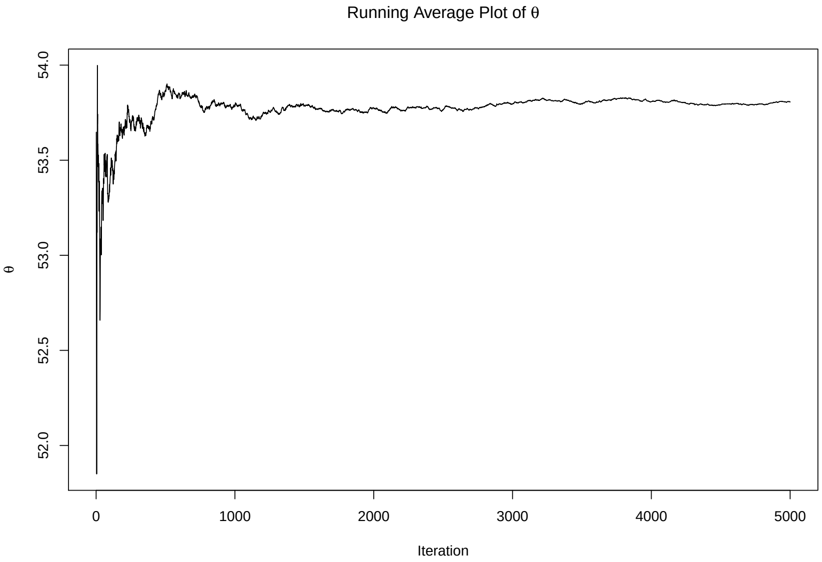
Traceplot of θ_2



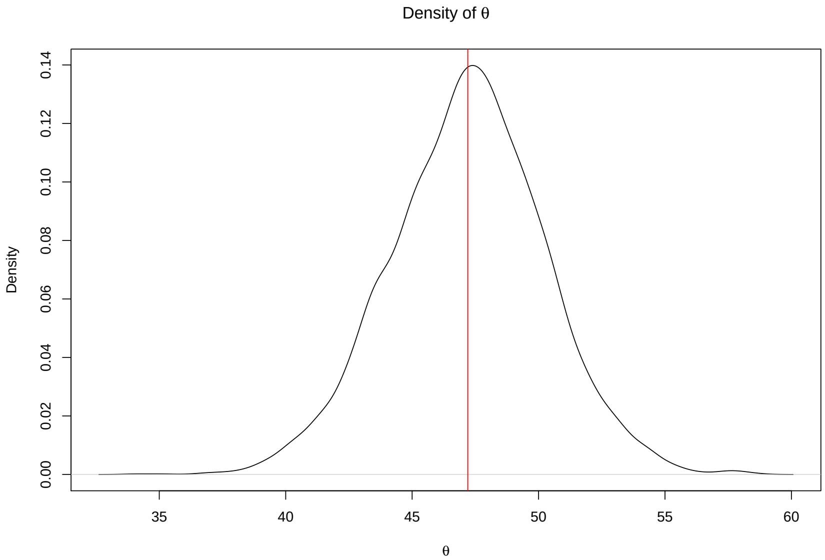
Running average plot of θ_1



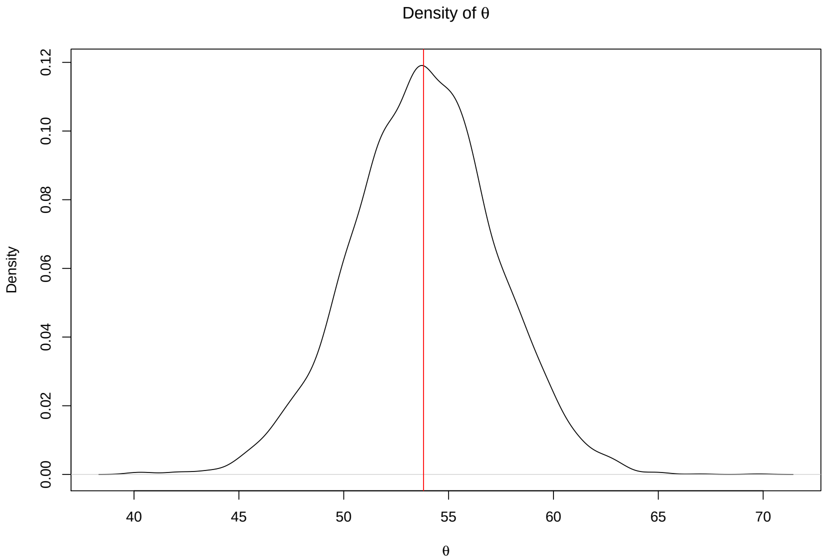
Running average plot of θ_2



Estimated density of θ_1



Estimated density of θ_2



Traceplots and running average plots of σ

Examine the trace plots and running average plots of Σ on your own.

Return to posterior inference

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

Confidence regions

```r
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 quanify this by calculating

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

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

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