

# Module 8: The Multivariate Normal Distribution

Rebecca C. Steorts

Hoff, Section 7.4

# Agenda

- ▶ Motivational reading comprehension case study
- ▶ Introduction/Review of vectors, matrices
- ▶ Population means/covariance matrices
- ▶ General multivariate notation
- ▶ Background on linear algebra (with practice exercises)
- ▶ Determinants, traces, quadratic forms

# Agenda

- ▶ The multivariate normal distribution (MVN)
- ▶ Exercise with the MVN
- ▶ MVN-MVN semi-conjugacy
- ▶ The inverse wishart distribution
- ▶ MVN-inverse wishart semi-conjugacy
- ▶ The MVN-MVN-inverse wishart model
- ▶ Applying a Gibbs sampler
- ▶ How to draw samples from the MVN and inverse wishart distributions
- ▶ Case study on reading comprehension

# Goal

The goal of this module is to be able **to understand how to work with multivariate distributions**, such as the multivariate normal distribution.

We also want to understand how univariate models that we have used in the past translate to the multivariate setting.

Before we can delve in, we must review background on matrices, vectors, and **multivariate notation**. We also must review background on **linear algebra**.

## Example: Reading Comprehension

A sample of 22 children are given reading comprehension tests before and after receiving a particular instructional method.<sup>1</sup>

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 by the vector  $\mathbf{Y}_i$

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

where  $i = 1, \dots, n$  and  $p = 2$ .

---

<sup>1</sup>This example follows Hoff (Section 7.4, p. 112).

## Example: Reading Comprehension

What does this data look like that is observed?

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

- ▶ A row of  $\mathbf{X}_{n \times p}$  represents a covariate we might be interested in, such as age of a person.
- ▶ Denote  $x_i$  ( $p \times 1$ ) as the  $i$ th **row vector** of the  $\mathbf{X}_{n \times p}$  matrix.

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

## Example: Reading Comprehension

We may be interested in the population mean  $\mu_{p \times 1}$ .

$$E[\mathbf{Y}] =: E[\mathbf{Y}_i] = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \boldsymbol{\mu}$$

## Example: Reading Comprehension

We also may be interested in the population covariance matrix,  $\Sigma_{p \times p}$ .

By definition:

$$\Sigma = \text{Cov}(\mathbf{Y}) \quad (1)$$

$$= \begin{pmatrix} E[Y_1^2] - E[Y_1]^2 & E[Y_1 Y_2] - E[Y_1]E[Y_2] \\ E[Y_1 Y_2] - E[Y_1]E[Y_2] & E[Y_2^2] - E[Y_2]^2 \end{pmatrix} \quad (2)$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix} \quad (3)$$

Remark:  $\text{Cov}(Y_1) = \text{Var}(Y_1) = \sigma_1^2$ .       $\text{Cov}(Y_1, Y_2) = \sigma_{1,2}$ .



## How do we expand this beyond our reading comprehension example

We introduced our notation based upon a specific example to reading comprehension.

How can we make this more general and applicable to general case studies and problems?

## General Notation

Assume that  $\mathbf{y}_{p \times 1} \sim (\mu_{p \times 1}, \Sigma_{p \times p})$ .

$$\mathbf{y}_{p \times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}.$$

$$\boldsymbol{\mu}_{p \times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

$$\Sigma_{p \times p} = \text{Cov}(\mathbf{y}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}.$$

# Background

Before proceeding, we need to review some basic concepts from linear algebra:

1. Basic properties of matrices
2. Useful lemmas for working with matrices

# The determinant of a matrix

Assume a matrix  $A_{n \times n}$  is invertible. The

$$\det(A) = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in},$$

where  $A_{ij}$  are the co-factors and are computed from

$$A_{ij} = (-1)^{i+j} \det(M_{ij}).$$

$M_{ij}$  is known as the minor matrix and is the matrix you get if you eliminate row  $i$  and column  $j$  from matrix  $A$ . You must apply this technique recursively.

**We only use this technique when doing such calculations by hand or in proof-based approaches.**

# The determinant of a matrix

- ▶ How on earth do I use the complicated formula on the pervious slide.

**Easy: Use the `det` command in R when faced with an application.**

- ▶ You will also see a determinant in the definition of the multivariate normal distribution.

**Important point: It's just a function and we typically do not need to evalute it in this course!**

# The trace of a matrix

Assume a matrix  $H_{p \times p}$ .

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

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

# Interactive Exercise

We're going to move into working in groups, so you can work on the following exercises.

## The trace of a matrix

$$H = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

What is  $\text{tr}(H)$ ?

(Take 1 minute to complete this.)



# Linear Algebra Tricks

Suppose that  $A$  is  $n \times n$  matrix and suppose that  $B$  is a  $n \times n$  matrix.

Lemma 1:

$$\text{tr}(AB) = \text{tr}(BA)$$

Proof: Exercise.

(Take 5-7 minutes to complete this.)

# Linear Algebra Tricks

Lemma 2:

Suppose  $\mathbf{x}$  is a vector.  $\mathbf{x}^T A \mathbf{x}$  is called a **quadratic form**.

$$\mathbf{x}^T A \mathbf{x} = \text{tr}(\mathbf{x}^T A \mathbf{x}) = \text{tr}(\mathbf{x} \mathbf{x}^T A) = \text{tr}(A \mathbf{x} \mathbf{x}^T)$$

Proof: Exercise.

# Linear Algebra Tricks

Proof of Lemma 2:

$$\text{tr}(\mathbf{x}^T A \mathbf{x}) = \sum_i (\mathbf{x}^T A \mathbf{x})_{ii} \quad (4)$$

$$= (\mathbf{x}^T (A \mathbf{x})) \quad (5)$$

$$= \text{tr}(A \mathbf{x} \mathbf{x}^T) \text{ (by Lemma 1)} \quad (6)$$

$$\text{tr}(\mathbf{x}^T A \mathbf{x}) = \sum_i (\mathbf{x}^T A \mathbf{x})_{ii} \quad (7)$$

$$= ((\mathbf{x}^T A) \mathbf{x}) \quad (8)$$

$$= \text{tr}(\mathbf{x} \mathbf{x}^T A) \text{ (by Lemma 1)} \quad (9)$$

# Notation

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

Example:  $\text{Cov}(Y_1, Y_2)$  is just one element of the matrix  $\Sigma$ .

# Multivariate Normal

Just as the probability density of a scalar normal is

$$p(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp \left\{ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right\}, \quad (10)$$

the probability density of the multivariate normal is

$$p(\mathbf{x}) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}. \quad (11)$$

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

Lemma 3.

Consider

$$Z_1, \dots, Z_n \stackrel{iid}{\sim} N(0, 1).$$

Show that

$$Z_1, \dots, Z_n \sim MVN(\mathbf{0}, I_{n \times n}).$$

## Proof of Lemma 3

Proof:

$$f_z(z) = \prod_{i=1}^n (2\pi)^{-1/2} e^{-z_i^2/2} \quad (12)$$

$$= (2\pi)^{-n/2} e^{\sum_i -z_i^2/2} \quad (13)$$

$$= (2\pi)^{-n/2} e^{-z^T z/2}. \quad (14)$$

The last line follows since  $\sum_i -z_i^2 = -z^T z$ .

Thus,  $Z_1, \dots, Z_n \sim \text{MVN}(\mathbf{0}, I)$ .

## Conjugate to MVN

Suppose that

$$\mathbf{y} = (y_1 \dots y_n)^T \mid \theta \sim MVN(\theta, \Sigma).$$

Let

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

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



## Prior

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-p/2} \det \Omega^{-1/2} \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \right\} \quad (15)$$

$$\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \right\} \quad (16)$$

$$\propto \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \Omega^{-1} \boldsymbol{\mu} \right\} \quad (17)$$

$$\propto \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Omega^{-1} \boldsymbol{\mu} \right\} \quad (18)$$

$$= \exp -\frac{1}{2} \left\{ \boldsymbol{\theta}^T A_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o \right\} \quad (19)$$

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

## Likelihood

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \Sigma) = \prod_{i=1}^n (2\pi)^{-p/2} \det \Sigma^{-1/2} \exp \left\{ -\frac{1}{2} (y_i - \boldsymbol{\theta})^T \Sigma^{-1} (y_i - \boldsymbol{\theta}) \right\} \quad (20)$$

$$\propto \exp -\frac{1}{2} \left\{ \sum_i y_i^T \Sigma^{-1} y_i - 2 \sum_i \boldsymbol{\theta}^T \Sigma^{-1} y_i + \sum_i \boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta} \right\} \quad (21)$$

$$\propto \exp -\frac{1}{2} \left\{ -2 \boldsymbol{\theta}^T \Sigma^{-1} n \bar{\mathbf{y}} + n \boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta} \right\} \quad (22)$$

$$\propto \exp -\frac{1}{2} \left\{ -2 \boldsymbol{\theta}^T \mathbf{b}_1 + \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} \right\}, \quad (23)$$

where

$$\mathbf{b}_1 = \Sigma^{-1} n \bar{\mathbf{y}}, \quad \mathbf{A}_1 = n \Sigma^{-1}$$

and

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

## Full conditional

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \Sigma) \propto p(\mathbf{y} \mid \boldsymbol{\theta}, \Sigma) \times p(\boldsymbol{\theta}) \quad (24)$$

$$\propto \exp - \frac{1}{2} \left\{ -2\boldsymbol{\theta}^T \mathbf{b}_1 + \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} \right\} \quad (25)$$

$$\times \exp - \frac{1}{2} \left\{ \boldsymbol{\theta}^T \mathbf{A}_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \mathbf{b}_o \right\} \quad (26)$$

$$\propto \exp \left\{ \boldsymbol{\theta}^T \mathbf{b}_1 - \frac{1}{2} \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} - \frac{1}{2} \boldsymbol{\theta}^T \mathbf{A}_o \boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{b}_o \right\} \quad (27)$$

$$\propto \exp \left\{ \boldsymbol{\theta}^T (\mathbf{b}_o + \mathbf{b}_1) - \frac{1}{2} \boldsymbol{\theta}^T (\mathbf{A}_o + \mathbf{A}_1) \boldsymbol{\theta} \right\} \quad (28)$$

## Full conditional

From the previous slide, recall that

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \Sigma) \propto \exp\{\boldsymbol{\theta}^T(b_o + b_1) - \frac{1}{2}\boldsymbol{\theta}^T(A_o + A_1)\boldsymbol{\theta}\}$$

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

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

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

# Interpretations

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

$$\mu_n = A_n^{-1}b_n = [\Omega^{-1} + n\Sigma^{-1}]^{-1}(b_o + b_1) \quad (29)$$

$$= [\Omega^{-1} + n\Sigma^{-1}]^{-1}(\Omega^{-1}\mu + \Sigma^{-1}n\bar{y}) \quad (30)$$

$$\Sigma_n = A_n^{-1} = [\Omega^{-1} + n\Sigma^{-1}]^{-1} \quad (31)$$

## inverse Wishart distribution

Let us now consider a prior distribution on  $\Sigma_{p \times p}$ , which must be a positive definite matrix, meaning that  $\mathbf{x}^T \Sigma \mathbf{x} > 0$  for all  $\mathbf{x}$ .

Suppose  $\Sigma_{p \times p} \sim \text{inverseWishart}(\nu_o, S_o^{-1})$  where  $\nu_o$  is a scalar and  $S_o^{-1}$  is a matrix, where  $\nu_o > p - 1$  and  $S_o$  must be positive definite.

It can be shown that

$$E[\Sigma] = \frac{1}{\nu_o - p - 1} S_o.$$

(See Hoff, p. 110.)

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

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

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

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

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

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

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

## Continued

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

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \Sigma) \quad (32)$$

$$\propto \det(\Sigma)^{-n/2} \exp\left\{-\sum_i (\mathbf{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta})/2\right\} \quad (33)$$

$$\propto \det(\Sigma)^{-n/2} \exp\left\{-\text{tr}\left(\sum_i (\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})^T \Sigma^{-1}/2\right)\right\} \quad (34)$$

$$\propto \det(\Sigma)^{-n/2} \exp\left\{-\text{tr}(S_{\boldsymbol{\theta}} \Sigma^{-1}/2)\right\} \quad (35)$$

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

Note that

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

where B is the matrix whose  $k$ th row is  $b_k$ . (Here we are applying Lemma 2.)



## Continued

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

$$p(\Sigma \mid \mathbf{y}, \boldsymbol{\theta}) \tag{36}$$

$$= p(\Sigma) \times p(\mathbf{y} \mid \boldsymbol{\theta}, \Sigma) \tag{37}$$

$$\propto \det(\Sigma)^{-(\nu_o + p + 1)/2} \times \exp\{-\text{tr}(S_o \Sigma^{-1})/2\} \tag{38}$$

$$\times \det(\Sigma)^{-n/2} \exp\{-\text{tr}(S_\theta \Sigma^{-1})/2\} \tag{39}$$

$$\propto \det(\Sigma)^{-(\nu_o + n + p + 1)/2} \exp\{-\text{tr}((S_o + S_\theta) \Sigma^{-1})/2\} \tag{40}$$

This implies that

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

## Continued

Suppose that we wish now to take

$$\boldsymbol{\theta} \mid \mathbf{y}, \Sigma \sim \text{MVN}(\boldsymbol{\mu}_n, \Sigma_n)$$

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

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

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

# Gibbs sampler

Suppose the Gibbs sampler is at iteration  $s$ .

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

# Simulating Data

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

```
# Each row corresponds to a sample  
# Here we have one sample (one row)  
rmvnorm(n = 1, mean = rep(0, 2), sigma = diag(2))
```

```
##           [,1]      [,2]  
## [1,] 0.2613613 1.912427
```

# 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,] -0.3004900 -0.1311393
## [2,]  0.4553081 -0.2442609
## [3,]  0.7046704 -1.0748051
```

## 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,] 1.4065771 0.5531162  
## [2,] 0.5531162 0.2212803
```



# 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, where  $i = 1, \dots, n$ .

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

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

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

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

Let  $\boldsymbol{\theta} = (\theta_1, \theta_2)$ .

$i = 1, \dots, n$ .

## Prior settings

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

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

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

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

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

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

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

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)

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

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

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

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  $\nu_0 = p + 2 = 4$ , which is the mean of the inverse Wishart distribution (Hoff, p. 100).

## 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))
```

```
## pretest posttest
```

```
## 47.18182 53.86364
```

## 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

```
## Loading required package: coda
## Loading required package: MASS
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
## ## Copyright (C) 2003-2020 Andrew D. Martin, Kevin M. Quinn
## ##
## ## Support provided by the U.S. National Science Foundation
## ## (Grants SES-0350646 and SES-0350613)
## ##
```



# Gibbs sampler (review)

Suppose the Gibbs sampler is at iteration  $s$ .

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

## 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) %*% 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

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

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

# Glance at Gibbs sampler

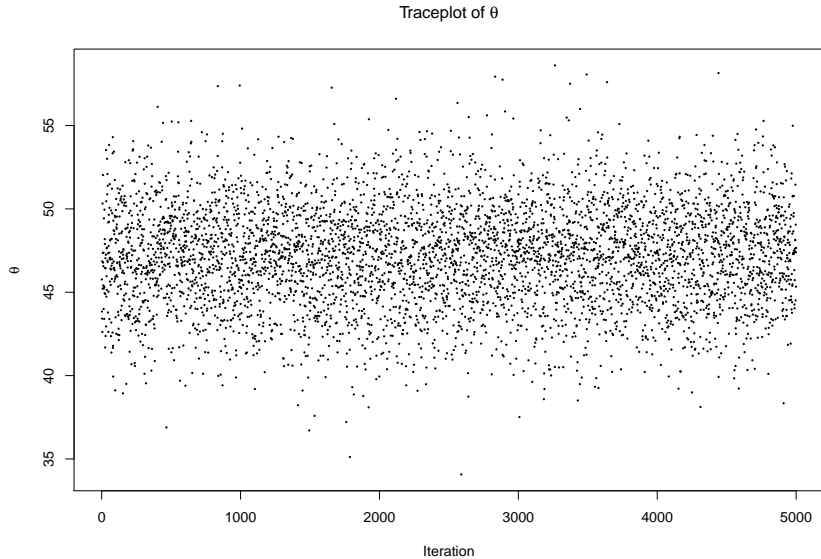
```
head(THETA)
```

```
##           [,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
```

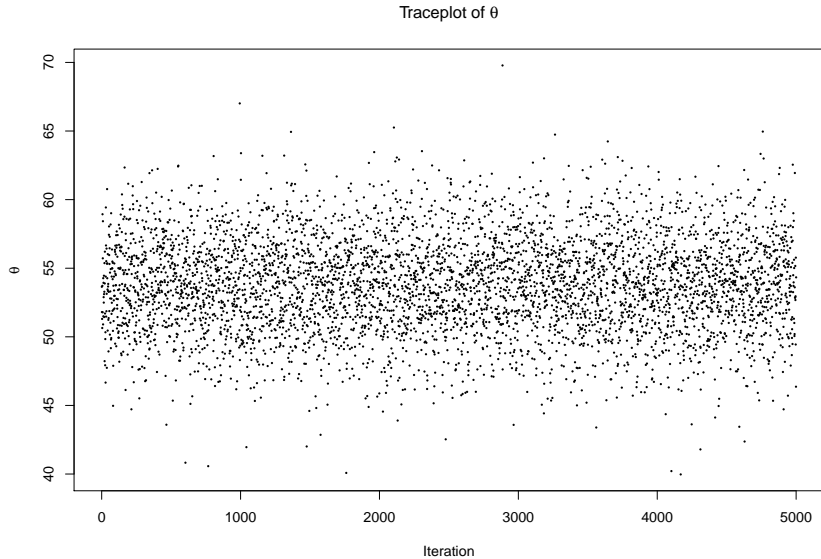
```
head(SIGMA)
```

```
##           [,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
## [6,] 289.4997 246.9184 246.9184 319.9786
```

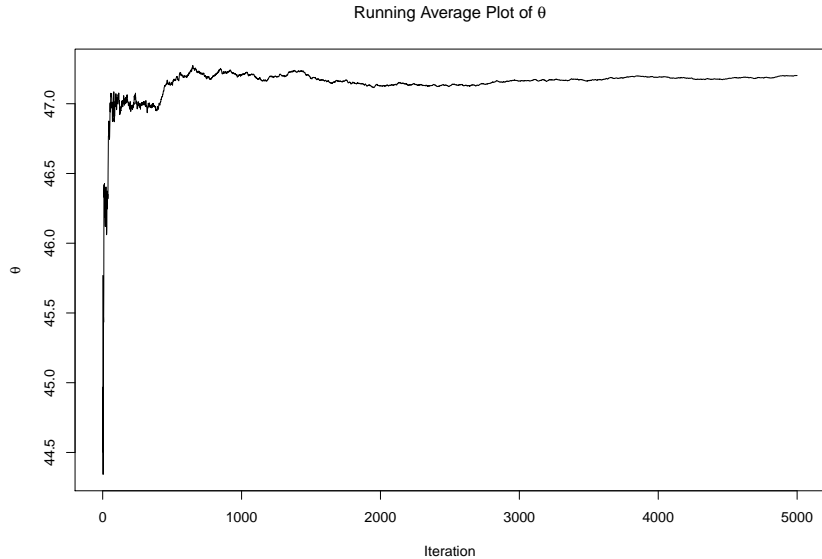
# Traceplot of $\theta_1$



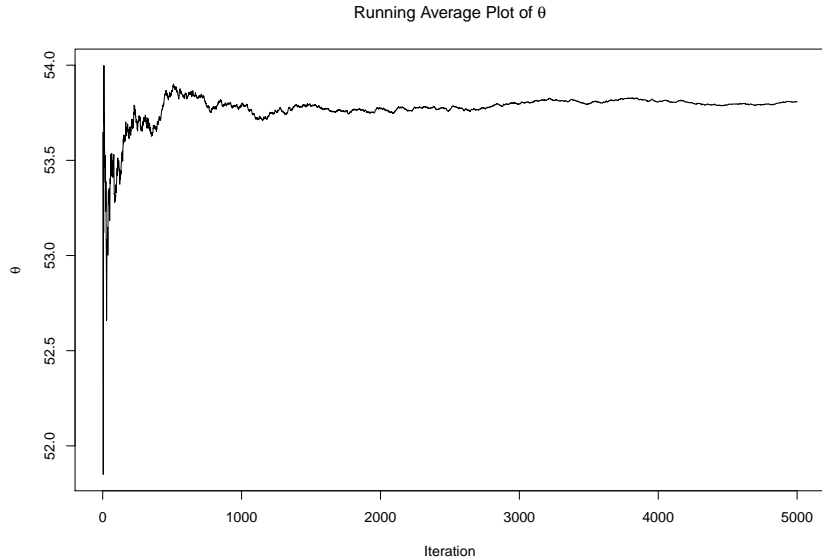
## Traceplot of $\theta_2$



# Running average plot of $\theta_1$

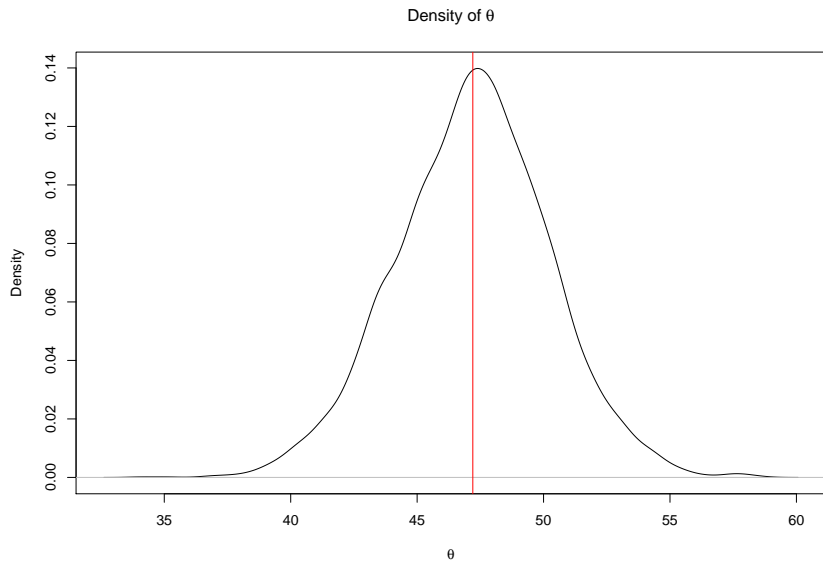


## Running average plot of $\theta_2$

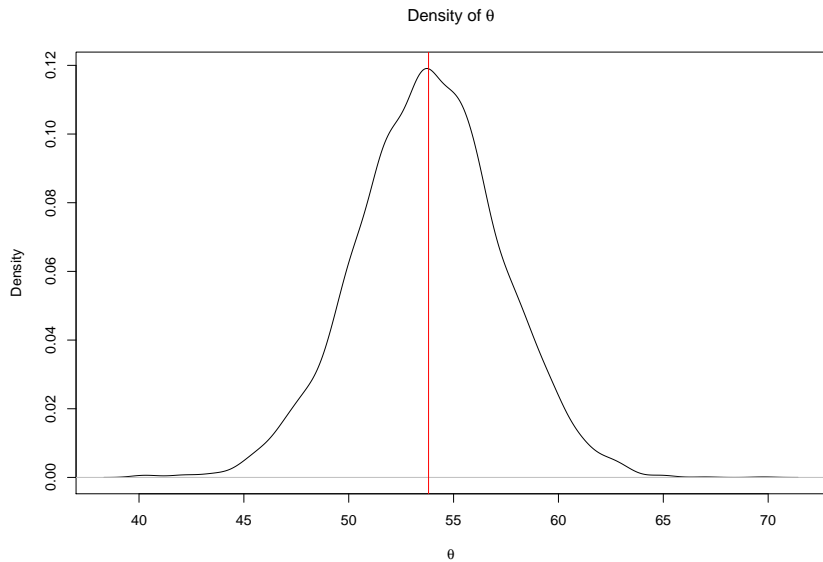




# Estimated density of $\theta_1$



## Estimated density of $\theta_2$



## Traceplots and running average plots

The traceplots don't tell us much of anything, so this is why we examine the running average plots. Specifically, the traceplots indicate that the chain has not failed to converged.

The running average plots indicate that the sampler appears to be mixing well by 5,000 iterations and that the chain has not failed to converged.

## 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
```

# Detailed Takeaways on Background

- ▶ Understanding vectors, matrices and notation
- ▶ Understanding how to write multivariate notation for a conceptual problem
- ▶ Understanding how to write general multivariate notation
- ▶ Background on linear algebra
- ▶ Determinants, traces, quadratic forms
- ▶ Knowing how to do simple proofs such as the exercises from class



# Detailed Takeaways on Multivariate Normal Models

- ▶ The multivariate normal distribution (MVN)
- ▶ Exercise with the MVN
- ▶ MVN-MVN semi-conjugacy
- ▶ The inverse wishart distribution
- ▶ MVN-inverse wishart semi-conjugacy
- ▶ The MVN-MVN-inverse wishart model
- ▶ Applying a Gibbs sampler
- ▶ How to draw samples from the MVN and inverse wishart distributions
- ▶ Case study on reading comprehension

# Proof of Lemma 1

$$\text{tr}(AB) = \text{tr}(BA)$$

Proof: Suppose that  $A_{n \times n}$  and  $B_{n \times n}$ .

Recall that by definition  $\text{tr}(A) = \sum_{i=1}^n a_{ii}$ . By definition

$$\text{tr}(AB) = \sum_{i=1}^n (AB)_{ii} \tag{41}$$

$$\sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ji} \tag{42}$$

$$\sum_{i=1}^n \sum_{j=1}^n b_{ji} a_{ij} \tag{43}$$

$$= \sum_{i=1}^n (BA)_{ii} = \text{tr}(BA) \tag{44}$$