GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lecture 6: The Multivariate Normal Model

Hai Shu, PhD

10/17/2022

Topics

- Multivariate normal distribution
- Inverse Wishart distribution
- Multivariate normal data: inferring upon both the mean and the variance

Multivariate normal data

- Most of the examples we have seen so far consisted in measurements of a single variable on a sample of individuals or measurements of a single variable for every run of an experiment.
- In most situations, we have multiple measurements for each individual or for every experiment.
- We need to develop models for multivariate data.
- The most useful model for multivariate data is the multivariate normal model that allows to estimate jointly the population mean, the population variance and the association (quantified by the correlation) among the variables.
- The multivariate normal data example is quite important also because most of the examples we will look at later (for example, linear models) are closely related to the multivariate normal case.

Multivariate data

- We illustrate the multivariate normal model with an example. Suppose we have a sample of 22 children who are given a reading comprehension test before and after receiving a particular instructional method.
- For each student i we have two scores, Y_{1i} and Y_{2i}, the before and after score.
- We denote the pair of scores of student i in a 2×1 vector $\mathbf{Y}_i$:

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

- The population mean is then given by a 2×1 vector θ:

$$\theta = E[\mathbf{Y}_i] = \begin{pmatrix} E[Y_{1i}] \\ E[Y_{2i}] \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

Multivariate data

- The population variance for the first and second score as well as the covariance between the two scores are given by the population covariance matrix $\mathbf{\Sigma}$:

$$\begin{aligned}
\mathbf{\Sigma} &= \text{Cov}(\mathbf{Y}) = \begin{pmatrix} \text{Var}[Y_1] & \text{Cov}[Y_1, Y_2] \\ \text{Cov}[Y_1, Y_2] & \text{Var}[Y_2] \end{pmatrix} \\
&= \begin{pmatrix} E\left[Y_1^2\right] - (E[Y_1])^2 & E[Y_1 \cdot Y_2] - E[Y_1] \cdot E[Y_2] \\ E[Y_1 \cdot Y_2] - E[Y_1] \cdot E[Y_2] & E\left[Y_2^2\right] - (E[Y_2])^2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}
\end{aligned}$$

Multivariate data

- Note that the population covariance matrix $\mathbf{\Sigma}$ has diagonal entries that are positive, and all its elements are function of the first and second moments of $\mathbf{Y}$:

first moments	$E[Y_1], E[Y_2]$
second moments	$E\left[Y_1^2\right], E\left[Y_2^2\right], E[Y_1 \cdot Y_2]$

- Conversely, the population mean θ is function only of the first moments of $\mathbf{Y}$.

Multivariate data

- In analyzing the data for the 22 children, we are interested in inferring about the difference between $E[Y_2] - E[Y_1] = \theta_2 - \theta_1$, that is, the difference in the population mean for the second and the first score.
  This will help assess the effectiveness of the teaching method.
- We might also be interested in inferring upon the population correlation ρ between the two scores:

$$\rho = \frac{E[Y_1 \cdot Y_2] - E[Y_1] \cdot E[Y_2]}{\sqrt{\text{Var}[Y_1] \cdot \text{Var}[Y_2]}} = \frac{\sigma_{1,2}}{\sqrt{\sigma_1^2 \cdot \sigma_2^2}}$$

  This will inform us about the consistency in the reading comprehension test.

Multivariate normal distribution

- The univariate normal distribution $N(\mu, \sigma^2)$ is characterized by the fact that it is completely determined by the two parameters, the population mean, μ, and the population variance, σ^2.
- The analogous model for multivariate data is the multivariate normal distribution $N_p(\mathbf{Y}; \theta, \mathbf{\Sigma})$.
  A p-dimensional vector $\mathbf{Y}$ has a multivariate normal distribution if its density is given by:

$$p(\mathbf{y}|\theta, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\theta)' \mathbf{\Sigma}^{-1} (\mathbf{y}-\theta)\right\}$$

$$\text{where } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \ddots & & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Matrix algebra

- Calculating the multivariate normal distribution involves some matrix algebra.
- Given a $p \times p$ matrix, $\mathbf{A}$, the scalar (i.e. the number) $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$ and tells us how "big" the matrix $\mathbf{A}$ is.
  The determinant of a matrix can be computed in `R` using the command `det(A)`.
- Given a $p \times p$ matrix, $\mathbf{A}$, the $p \times p$ matrix $\mathbf{A}^{-1}$ denotes the inverse of $\mathbf{A}$ and it is the matrix such that $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}_p$, the identity matrix of order p, that is the matrix that has all entries equal to 0 except for 1's on the diagonal.
  The inverse of a matrix can be computed in `R` using the command `solve(A)`.

Matrix algebra

- If $\mathbf{b}$ is a $p \times 1$ vector, $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$, then the transpose of $\mathbf{b}$ is the $1 \times p$ vector $\mathbf{b}'$:

$$\mathbf{b}' = \begin{pmatrix} b_1 & b_2 & \dots & b_p \end{pmatrix}$$

- If $\mathbf{b}$ is a $p \times 1$ vector and $\mathbf{A}$ is a $p \times p$ matrix, then the vector-matrix product $\mathbf{b}'\mathbf{A}$ is the $1 \times p$ vector

$$\begin{pmatrix} \sum_{j=1}^{p} b_j a_{j,1} & \sum_{j=1}^{p} b_j a_{j,2} & \dots & \sum_{j=1}^{p} b_j a_{j,p} \end{pmatrix}$$

- Conversely, the product, also called the quadratic form $\mathbf{b}'\mathbf{A}\mathbf{b}$ is the scalar (i.e. the number)

$$\sum_{j=1}^{p} \sum_{k=1}^{p} b_j a_{j,k} b_k = \sum_{j=1}^{p} \sum_{k=1}^{p} b_j b_k a_{j,k}$$

Multivariate normal distribution

- If $\mathbf{Y}$ is a $p \times 1$ vector with multivariate normal distribution with mean θ and covariance matrix $\mathbf{\Sigma}$ where

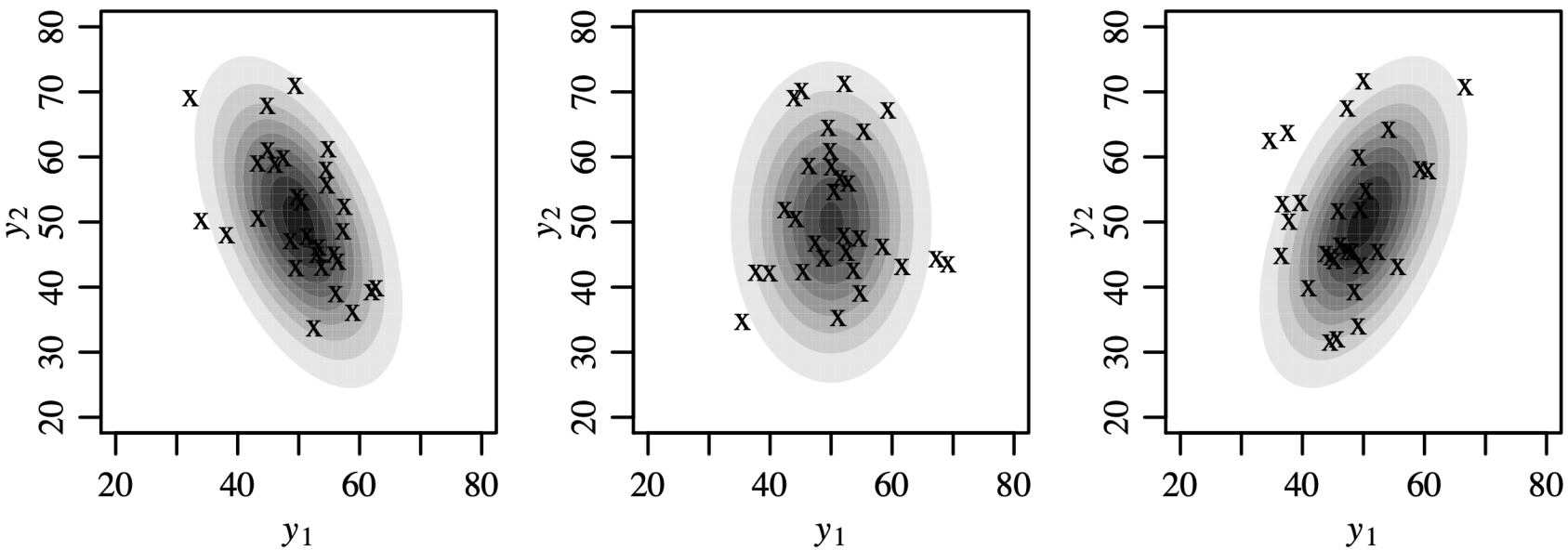
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \ddots & & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_p^2 \end{pmatrix}$$

then, marginally each component Y_j of the $\mathbf{Y}$ vector has a univariate normal distribution with mean θ_j and variance σ_j^2:

$$Y_j \sim N(\theta_j, \sigma_j^2).$$

Multivariate normal distribution

- Scatter and contour plots of three bivariate normal distributions.
  In each bivariate normal density $\theta = \begin{pmatrix} 50 \\ 50 \end{pmatrix}$, $\sigma_1^2 = 64$, $\sigma_2^2 = 144$
  but the covariance parameter $\sigma_{1,2}$ varies from plot to plot, from -48 to 0 to 48. These covariances $\sigma_{1,2}$ correspond to correlation parameters of, respectively, -0.5, 0 and 0.5.



Covariance matrix

- Before we start looking on how to make inference for multivariate normal data, let's look more closely at the covariance matrix.
- Suppose $\mathbf{Y}$ is a p-dimensional random vector with a multivariate normal $N_p(\theta, \mathbf{\Sigma})$ distribution. Then, the covariance matrix $\mathbf{\Sigma}$ is a square $p \times p$ matrix and
  - its diagonal elements are positive, that is $\sigma_j^2 > 0$ for $j = 1, \ldots, p$
  - it is symmetrix, that is $\sigma_{ij} = \sigma_{ji}$ for $i, j = 1, \ldots, p$
  - it is positive definite

Positive definiteness

- A square $p \times p$ matrix $\mathbf{\Sigma}$ is said positive definite if for any **nonzero** p-dimensional vector $\mathbf{x}$

$$\mathbf{x}'\mathbf{\Sigma}\mathbf{x} > 0$$

  i.e. the quadratic form $\mathbf{x}'\mathbf{\Sigma}\mathbf{x}$ is positive.

- Note that the quadratic from $\mathbf{x}'\mathbf{\Sigma}\mathbf{x}$ is equal to

$$\sum_{i=1}^{p}\sum_{j=1}^{p} x_i \sigma_{ij} x_j$$

- The above requirement is quite strong since it has to hold for **ANY** p-dimensional vector $\mathbf{x} \neq \mathbf{0}$.
  In other words, it is not very easy to verify that a square $p \times p$ matrix is positive definite.

Prior distribution for a covariance matrix

- Suppose we have observations $\mathbf{y}_1, \ldots, \mathbf{y}_n | \theta, \mathbf{\Sigma} \overset{iid}{\sim} N_p(\theta, \mathbf{\Sigma})$ and we want to infer upon **both** θ and $\mathbf{\Sigma}$
- Then, we need to place priors on θ and $\mathbf{\Sigma}$.
- A convenient choice of a prior for the p-dimensional vector θ is a multivariate normal distribution, that is, $p(\theta) = N_p(\mu_0, \mathbf{\Lambda}_0)$.
- Given that a covariance matrix is a $p \times p$ positive definite, symmetric matrix, the prior distribution for $\mathbf{\Sigma}$ needs to put mass on the space of square $p \times p$ positive definite, symmetric matrices.
- How do we define such a distribution?

Sum of squares matrix

- Let's consider for a moment the p-dimensional vector $\mathbf{y}_i$ and its transpose $\mathbf{y}_i'$. Then:

$$\mathbf{y}_i \cdot \mathbf{y}_i' = \begin{pmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,p} \end{pmatrix} \cdot \begin{pmatrix} y_{i,1} & y_{i,2} & \cdots & y_{i,p} \end{pmatrix}$$

$$= \begin{pmatrix} y_{i,1}^2 & y_{i,1}y_{i,2} & \cdots & y_{i,1}y_{i,p} \\ y_{i,2}y_{i,1} & y_{i,2}^2 & \cdots & y_{i,2}y_{i,p} \\ \vdots & \vdots & & \vdots \\ y_{i,p}y_{i,1} & y_{i,p}y_{i,2} & \cdots & y_{i,p}^2 \end{pmatrix}$$

that is, $\mathbf{y}_i \cdot \mathbf{y}_i'$ is a $p \times p$ matrix where the diagonal elements are equal to the squared components of the vector $\mathbf{y}_i$ while the off-diagonal elements are equal to the cross-products of the components of $\mathbf{y}_i$.

Sum of squares matrix

- Now, let's consider the sum of all n such $p \times p$ matrices, that is, let's consider

$$\sum_{i=1}^{n} \mathbf{y}_i \cdot \mathbf{y}_i' = \sum_{i=1}^{n} \begin{pmatrix} y_{i,1}^2 & y_{i,1}y_{i,2} & \cdots & y_{i,1}y_{i,p} \\ y_{i,2}y_{i,1} & y_{i,2}^2 & \cdots & y_{i,2}y_{i,p} \\ \vdots & \vdots & & \vdots \\ y_{i,p}y_{i,1} & y_{i,p}y_{i2} & \cdots & y_{i,p}^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} y_{i,1}^2 & \sum_{i=1}^{n} y_{i,1}y_{i,2} & \cdots & \sum_{i=1}^{n} y_{i,1}y_{i,p} \\ \sum_{i=1}^{n} y_{i,2}y_{i,1} & \sum_{i=1}^{n} y_{i,2}^2 & \cdots & \sum_{i=1}^{n} y_{i,2}y_{i,p} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} y_{i,p}y_{i,1} & \sum_{i=1}^{n} y_{i,p}y_{i2} & \cdots & \sum_{i=1}^{n} y_{i,p}^2 \end{pmatrix}$$

This matrix is called the sum of squares matrix.

Sum of squares matrix

- It easy to show that the sum of square matrix $\sum_{i=1}^{n} \mathbf{y}_i \cdot \mathbf{y}_i'$ is also equal to the matrix $\mathbf{Y}' \cdot \mathbf{Y}$ where

$$\mathbf{Y}' = \begin{pmatrix} y_{1,1} & y_{2,1} & \cdots & y_{n,1} \\ y_{1,2} & y_{2,2} & \cdots & y_{n,2} \\ \vdots & \vdots & & \vdots \\ y_{1,p} & y_{2,p} & \cdots & y_{n,p} \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p} \\ \vdots & \vdots & & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,p} \end{pmatrix}$$

Thus, $\mathbf{Y}'$ is a $p \times n$ matrix whose j-th column is the $p \times 1$ vector $\mathbf{y}_j$.

Sample covariance matrix

- Now, note that if $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are samples from the multivariate normal distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$, then:
  - $\frac{1}{n}[\mathbf{Y}'\mathbf{Y}]_{k,k} = \frac{1}{n}\sum_{i=1}^{n} y_{i,k}^2$
    provides an estimate of the variance σ_k^2, for $k = 1, \ldots, p$
  - $\frac{1}{n}[\mathbf{Y}'\mathbf{Y}]_{k,l} = \frac{1}{n}\sum_{i=1}^{n} y_{i,k} \cdot y_{i,l}$
    provides an estimate of the covariance σ_{kl}, for $k, l = 1, \ldots, p$
- Therefore, $\frac{1}{n}\sum_{i=1} \mathbf{y}_i \mathbf{y}_i' = \frac{1}{n}\mathbf{Y}'\mathbf{Y}$ is the sample covariance matrix.
- Additionally, if $n > p$, that is, if we have more observations than variables, and the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are linearly independent, then $\mathbf{Y}'\mathbf{Y}$ is positive definite and symmetric.
- If we knew what is the distribution of the sum of squares matrix $\mathbf{Y}'\mathbf{Y}$ when $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are $\overset{\text{iid}}{n}$ p-dimensional random vectors following a multivariate normal $N_p(\mathbf{0}, \mathbf{\Sigma})$, then we would know what can be a convenient choice of a prior for a covariance matrix.

Wishart distribution

- It can be shown that if $\mathbf{y}_1, \ldots, \mathbf{y}_{\nu_0}$ are $\overset{\text{iid}}{\nu_0}$ p-dimensional random vectors that follow a multivariate normal $N_p(\mathbf{0}, \mathbf{\Phi}_0)$ distribution, then the $p \times p$ matrix $\mathbf{Y}'\mathbf{Y}$ follows a Wishart distribution with parameters $(\nu_0, \mathbf{\Phi}_0)$.
- If a $p \times p$ random matrix $\mathbf{\Sigma}$ has a Wishart $(\nu_0, \mathbf{\Phi}_0)$ distribution, then:
  1. with probability 1, $\mathbf{\Sigma}$ is positive definite
  2. with probability 1, $\mathbf{\Sigma}$ is symmetric
  3. $E(\mathbf{\Sigma}) = \nu_0 \mathbf{\Phi}_0$
- The Wishart distribution is the multivariate analog of the Gamma distribution.
- In the univariate case, we used the Gamma distribution as a prior distribution for the precision$=\frac{1}{\text{(variance)}}$ while we used an Inverse Gamma distribution as prior distribution for the variance. In the multivariate case, we do the same.
- Thus, a convenient choice of a prior for the covariance matrix $\mathbf{\Sigma}$ is the Inverse Wishart distribution.

Inverse Wishart distribution

- A $p \times p$ random matrix $\mathbf{\Sigma}$ has an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution if its density is given by:

$$\begin{aligned} p(\mathbf{\Sigma}) &= \left[2^{\frac{\nu_0 p}{2}} \Gamma_p\left(\frac{\nu_0}{2}\right) |\mathbf{\Phi_0}|^{-\frac{\nu_0}{2}}\right]^{-1} \\ &\times |\mathbf{\Sigma}|^{-\frac{(\nu_0+p+1)}{2}} \exp\left(-\frac{\mathrm{tr}(\mathbf{\Phi}_0 \mathbf{\Sigma}^{-1})}{2}\right) \end{aligned}$$

  where $\Gamma_p(\cdot)$ is the multivariate p-dimensional Gamma function and tr stands for trace of the square $p \times p$ matrix $\mathbf{\Phi}_0 \mathbf{\Sigma}^{-1}$.

- If $\mathbf{\Sigma}$ is $p \times p$ square matrix, its trace is the sum of its diagonal elements, that is:

$$\mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{n} [\mathbf{\Sigma}]_{ii}$$

Inverse Wishart distribution

- If a $p \times p$ random matrix $\mathbf{\Sigma}$ has an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution, then:

$$E[\mathbf{\Sigma}] = \frac{\mathbf{\Phi}_0}{\nu_0 - p - 1} \qquad \text{Mode}[\mathbf{\Sigma}] = \frac{\mathbf{\Phi}_0}{\nu_0 + p + 1}$$

- The parameter ν_0 is called the degrees of freedom and it should be a real number greater than $p-1$, that is $\nu_0 > p-1$.
- Given the form of the expectation of an Inverse Wishart distribution, in constructing a prior distribution for $\mathbf{\Sigma}$ we can proceed as follows:
  - if we have a prior idea on what $\mathbf{\Sigma}$ should be, say $\mathbf{\Sigma}_0$, then we can take ν_0 large and $\mathbf{\Phi}_0 = (\nu_0 - p - 1) \cdot \mathbf{\Sigma}_0$. This will make the distribution of $\mathbf{\Sigma}$ concentrated around $\mathbf{\Sigma}_0$
  - if we choose $\nu_0 = p + 2$ and $\mathbf{\Phi}_0 = \mathbf{\Sigma}_0$, then we will have a distribution that is loosely centered around $\mathbf{\Sigma}_0$.

Wishart distribution

- Finally, if the $p \times p$ random matrix $\mathbf{\Sigma}$ has an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution, then its inverse, $\mathbf{\Sigma}^{-1}$ has a Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution and

$$E\left[\mathbf{\Sigma}^{-1}\right] = \nu_0 \mathbf{\Phi}_0^{-1} \qquad \text{Mode}\left[\mathbf{\Sigma}^{-1}\right] = (\nu_0 - p - 1)\mathbf{\Phi}_0^{-1}$$

- Now, suppose that we have a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of size n from the multivariate normal distribution $N_p(\theta, \mathbf{\Sigma})$ and we want to infer upon both the mean θ and the covariance matrix $\mathbf{\Sigma}$.
- Suppose we place a multivariate normal prior on θ: $p(\theta) = N_p(\mu_0, \mathbf{\Lambda}_0)$ and we place an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ prior on $\mathbf{\Sigma}$, that is, $p(\mathbf{\Sigma}) = \text{Inverse Wishart}(\nu_0, \mathbf{\Phi}_0^{-1})$.

Inferring upon both θ and Σ

- As for univariate normal data, we can show that if we place independent priors on θ and $\mathbf{\Sigma}$ of the type just considered, that is, if we place the following prior $p(\theta, \mathbf{\Sigma})$

$$p(\theta, \mathbf{\Sigma}) = p(\theta) \cdot p(\mathbf{\Sigma}) = N_p(\mu_0, \mathbf{\Lambda}_0) \cdot \text{Inverse Wishart}(\nu_0, \mathbf{\Phi}_0^{-1})$$

  then, the posterior distribution $p(\theta, \mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_n)$ does not have a closed recognizable form.

- However, we can approximate it by using a Gibbs sampling algorithm that generates a Markov chain with stationary distribution the joint posterior distribution $p(\theta, \mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_n)$ by sampling sequentially from the full conditionals:
  1. $p(\theta | \mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{\Sigma})$
  2. $p(\mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_n, \theta)$

Full conditional for θ

- Before deriving the full conditional for θ, let's first note that we can write the prior distribution of θ as:

$$\begin{aligned} p(\theta) &= \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Lambda}_0|^{\frac{1}{2}}} \cdot \exp\left\{-\tfrac{1}{2}(\theta-\mu_0)'\mathbf{\Lambda}_0^{-1}(\theta-\mu_0)\right\} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Lambda}_0|^{\frac{1}{2}}} \cdot \exp\left\{-\tfrac{1}{2}\theta'\mathbf{\Lambda}_0^{-1}\theta + \theta'\mathbf{\Lambda}_0^{-1}\mu_0 - \tfrac{1}{2}\mu_0'\mathbf{\Lambda}_0^{-1}\mu_0\right\} \\ &\propto \exp\left\{-\tfrac{1}{2}\theta'\mathbf{\Lambda}_0^{-1}\theta + \theta'\mathbf{\Lambda}_0^{-1}\mu_0\right\} \\ &= \exp\left\{-\tfrac{1}{2}\theta'\mathbf{A}_0\theta + \theta'\mathbf{b}_0\right\} \end{aligned}$$

  where the $\mathbf{A}_0 = \mathbf{\Lambda}_0^{-1}$ and $\mathbf{b}_0 = \mathbf{\Lambda}_0^{-1}\mu_0$.

- Note that this result holds for any $p \times 1$ random vector $\mathbf{Y}$!
  If $\mathbf{Y}$ has a density that is proportional to $\exp\left\{-\frac{1}{2}\mathbf{Y}'\mathbf{A}\mathbf{Y} + \mathbf{Y}'\mathbf{b}\right\}$ then $\mathbf{Y}$ has a multivariate normal distribution with covariance matrix $\mathbf{A}^{-1}$ and mean $\mathbf{A}^{-1} \cdot \mathbf{b}$.

Full conditional for θ

- If we have multivariate normal data $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n | \theta, \mathbf{\Sigma} \overset{iid}{\sim} N_p(\theta, \mathbf{\Sigma})$, then the likelihood is:

$$
\begin{aligned}
p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \theta, \mathbf{\Sigma}) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \cdot \exp\left\{ -\tfrac{1}{2} (\mathbf{y}_i - \theta)' \mathbf{\Sigma}^{-1} (\mathbf{y}_i - \theta) \right\} \\
&= \frac{1}{(2\pi)^{\frac{np}{2}}} \cdot \frac{1}{|\mathbf{\Sigma}|^{\frac{n}{2}}} \cdot \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \theta)' \mathbf{\Sigma}^{-1} (\mathbf{y}_i - \theta) \right\} \\
&\propto \exp\left\{ -\tfrac{1}{2} \theta' \mathbf{A}_1 \theta + \theta' \mathbf{b}_1 \right\}
\end{aligned}
$$

where $\mathbf{A}_1 = n\mathbf{\Sigma}^{-1}$ and $\mathbf{b}_1 = n\mathbf{\Sigma}^{-1} \cdot \bar{\mathbf{y}}$ with

$$
\bar{\mathbf{y}} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} y_{i,1} \\ \frac{1}{n} \sum_{i=1}^{n} y_{i,2} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} y_{i,p} \end{pmatrix}
$$

Full conditional for θ

- Therefore the full conditional for θ given the data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mathbf{\Sigma}$ is given by:

$$\begin{aligned} p(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{\Sigma}) &\propto p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\theta, \mathbf{\Sigma}) \cdot p(\theta) \\ &\propto \exp\left\{-\tfrac{1}{2}\theta'\mathbf{A}_1\theta + \theta'\mathbf{b}_1\right\} \cdot \exp\left\{-\tfrac{1}{2}\theta'\mathbf{A}_0\theta + \theta'\mathbf{b}_0\right\} \\ &= \exp\left\{-\tfrac{1}{2}\theta'\mathbf{A}_n\theta + \theta'\mathbf{b}_n\right\} \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_n &= \mathbf{A}_1 + \mathbf{A}_0 = n\mathbf{\Sigma}^{-1} + \mathbf{\Lambda}_0^{-1} \\ \mathbf{b}_n &= \mathbf{b}_1 + \mathbf{b}_0 = n\mathbf{\Sigma}^{-1} \cdot \bar{\mathbf{y}} + \mathbf{\Lambda}_0^{-1}\mu_0 \end{aligned}$$

Full conditional for θ

- This implies that the full conditional for (or the conditional posterior distribution of) θ given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\boldsymbol{\Sigma}$ is a multivariate normal distribution:

$$\theta | \mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\Sigma} \sim N_p\left(\mathbf{A}_n^{-1}\mathbf{b}_n, \mathbf{A}_n^{-1}\right).$$

- In other words:

$$\begin{aligned}
\text{Cov}\left[\theta | \mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\Sigma}\right] &= \boldsymbol{\Lambda}_n = \mathbf{A}_n^{-1} = \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_0^{-1}\right)^{-1} \\
\text{E}\left[\theta | \mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\Sigma}\right] &= \mu_n \\
&= \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_0^{-1}\right)^{-1} \cdot \left(n\boldsymbol{\Sigma}^{-1} \cdot \bar{\mathbf{y}} + \boldsymbol{\Lambda}_0^{-1}\mu_0\right) \\
p\left(\theta | \mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\Sigma}\right) &= N_p\left(\mu_n, \boldsymbol{\Lambda}_n\right)
\end{aligned}$$

Full conditional for θ

- These results are the multivariate analog of what we have seen in the univariate case when we want to infer on the mean and we assume that the variance is known.
- The inverse of the covariance matrix is the precision matrix.
- Then we have that the posterior precision matrix $\boldsymbol{\Lambda}_n^{-1}$ is equal to:

$$\boldsymbol{\Lambda}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_0^{-1}$$

that is, is the sum of the prior precision matrix $\boldsymbol{\Lambda}_0^{-1}$ and the data precision matrix $n\boldsymbol{\Sigma}^{-1}$.

- Additionally, the posterior mean μ_n is equal to:

$$\begin{aligned} \mu_n &= \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_0^{-1}\right)^{-1} \cdot \left(n\boldsymbol{\Sigma}^{-1} \cdot \bar{\mathbf{y}} + \boldsymbol{\Lambda}_0^{-1}\mu_0\right) \\ &= \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_0^{-1}\right)^{-1} n\boldsymbol{\Sigma}^{-1} \cdot \bar{\mathbf{y}} \\ &+ \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_0^{-1}\right)^{-1} \boldsymbol{\Lambda}_0^{-1}\mu_0 \end{aligned}$$

that is, it is the weighted average of the sample mean and the prior mean with weights that are proportional to the data precision and the prior precision, respectively.

Full conditional for Σ

- Let's derive now the full conditional for $\boldsymbol{\Sigma}$.
- Note that the likelihood is:

$$\begin{aligned} p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \theta, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \theta)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \theta)\right\} \\ &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{y}_i - \theta)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \theta)\right\} \end{aligned}$$

- From matrix algebra, we know that if $\mathbf{b}_1, \ldots, \mathbf{b}_n$ are n $p \times 1$ vectors and $\mathbf{A}$ is a square $p \times p$ matrix, then:

$$\sum_{i=1}^{n} \mathbf{b}_i' \mathbf{A} \mathbf{b}_i = \text{tr}(\mathbf{B}'\mathbf{B}\mathbf{A})$$

  where $\mathbf{B}'$ is the $p \times n$ matrix whose i-th column is the $p \times 1$ vector $\mathbf{b}_i$.
- Let's denote with $(\mathbf{Y} - \theta)'$ the $p \times n$ matrix whose i-th column is the $p \times 1$ vector $(\mathbf{y}_i - \theta)$, then

$$\sum_{i=1}^{n} (\mathbf{y}_i - \theta)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \theta) = \text{tr}\left[(\mathbf{Y} - \theta)' (\mathbf{Y} - \theta) \boldsymbol{\Sigma}^{-1}\right]$$

Full conditional for Σ

- $\sum_{i=1}^{n}(\mathbf{y}_i-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\theta) = \text{tr}\left[(\mathbf{Y}-\theta)'(\mathbf{Y}-\theta)\boldsymbol{\Sigma}^{-1}\right]$
- If we denote with $\mathbf{S}_\theta$ the $p \times p$ sum of squares matrix

$$\mathbf{S}_\theta = (\mathbf{Y}-\theta)'(\mathbf{Y}-\theta)$$

then, the likelihood can be written as:

$$\begin{aligned} p(\mathbf{y}_1,\ldots,\mathbf{y}_n|\theta,\boldsymbol{\Sigma}) &= (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\mathbf{y}_i-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\theta)\right\} \\ &= (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left(-\tfrac{1}{2}\text{tr}\left[\mathbf{S}_\theta\boldsymbol{\Sigma}^{-1}\right]\right) \end{aligned}$$

- The matrix $\mathbf{S}_\theta$ is called the residual sum of squares matrix for the vectors $\mathbf{y}_1,\ldots,\mathbf{y}_n$ if the population mean is θ.
- Note that, **conditional** on θ, $\frac{1}{n}\mathbf{S}_\theta$ provides an unbiased estimate of the covariance matrix $\boldsymbol{\Sigma}$.

Full conditional for Σ

- Let's derive the full conditional for $\boldsymbol{\Sigma}$.
- We have:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}|\mathbf{y}_1,\ldots,\mathbf{y}_n,\theta) &\propto p(\mathbf{y}_1,\ldots,\mathbf{y}_n|\theta,\boldsymbol{\Sigma})\cdot p(\boldsymbol{\Sigma}) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left(-\tfrac{1}{2}\text{tr}\left[\mathbf{S}_\theta\boldsymbol{\Sigma}^{-1}\right]\right) \\
&\cdot\ |\boldsymbol{\Sigma}|^{-\frac{(\nu_0+p+1)}{2}}\exp\left(-\tfrac{1}{2}\text{tr}\left[\boldsymbol{\Phi}_0\boldsymbol{\Sigma}^{-1}\right]\right) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{(n+\nu_0+p+1)}{2}}\exp\left(-\tfrac{1}{2}\text{tr}\left[\mathbf{S}_\theta\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Phi}_0\boldsymbol{\Sigma}^{-1}\right]\right) \\
&= |\boldsymbol{\Sigma}|^{-\frac{(n+\nu_0+p+1)}{2}}\exp\left(-\tfrac{1}{2}\text{tr}\left[(\mathbf{S}_\theta+\boldsymbol{\Phi}_0)\boldsymbol{\Sigma}^{-1}\right]\right)
\end{aligned}
$$

that is, $p(\boldsymbol{\Sigma}|\mathbf{y}_1,\ldots,\mathbf{y}_n,\theta) = \text{Inverse Wishart}(\nu_0+n,(\mathbf{S}_\theta+\boldsymbol{\Phi}_0)^{-1})$

Full conditional for Σ

- Since the prior distribution on $\mathbf{\Sigma}$ is Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ with ν_0 prior degrees of freedom and the full conditional for (or, the conditional posterior distribution of) $\mathbf{\Sigma}$ is $p(\mathbf{\Sigma}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \theta) = \text{Inverse Wishart}(\nu_0 + n, (\mathbf{S}_\theta + \mathbf{\Phi}_0)^{-1})$ it follows that:
  - the posterior degrees of freedom, ν_n, is equal to $\nu_0 + n$, i.e. the sum of the prior degrees of freedom and the sample size.
    This suggests that we can interpret the prior degrees of freedom ν_0 as the prior sample size.
  - The posterior center of the distribution, $\mathbf{\Phi}_n$, is equal to $\mathbf{S}_\theta + \mathbf{\Phi}_0$, i.e. the sum of $\mathbf{\Phi}_0$ and $\mathbf{S}_\theta$, the residual sum of squares matrix.
    This suggests that we can interpret $\mathbf{\Phi}_0$ as the prior residual sum of squares.

unscaled

Inferring upon both θ and Σ

- Thus, the full conditional distributions are:
  1. $p(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{\Sigma})$
  2. $p(\mathbf{\Sigma}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \theta)$
- A Gibbs sampling algorithm in this case will proceed as follows
  - Choose a number S of iterations
  - Repeat the following two steps for $j = 1, \ldots, S$
    - sample $\theta^{(j)} \sim p(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{\Sigma}^{(j-1)})$
    - sample $\mathbf{\Sigma}^{(j)} \sim p(\mathbf{\Sigma}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \theta^{(j)})$

Checking for convergence

- This will generate a Markov chain of values $\left((\theta^{(1)}, \mathbf{\Sigma}^{(1)}), (\theta^{(2)}, \mathbf{\Sigma}^{(2)}), \ldots, (\theta^{(S)}, \mathbf{\Sigma}^{(S)})\right)$
- Before starting to make inference, we need to check that the Markov chain has reached its stationary distribution, the joint posterior distribution $p(\theta, \mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_n)$.
  To check for convergence, we should do the following:
  - make trace plots for each parameter
  - look at the autocorrelation function to evaluate the mixing of the Markov chain
  - compute the effective sample size: this will also inform us on the autocorrelation in the Markov chain
  - ideally, we will run the Gibbs sampling algorithm m times using m different and dispersed initial values.

Checking for convergence

- Assume we have run m parallel Markov chains, then for each parameter in the m Markov chains, we can look at the difference among the within and between chains variance.
- Precisely, suppose we are making inference on r parameters $\theta_1, \ldots, \theta_r$ and we have run m parallel Markov chains. Let's consider only the parameter θ_1 at the moment.
  We can compute the sample variance for θ_1 within the m chains, that is, we can compute $s^2_{\theta_1,1}, s^2_{\theta_1,2}, \ldots, s^2_{\theta_1,m}$ where

$$s^2_{\theta_1,k} = \frac{1}{S-1} \sum_{j=1}^{S} (\theta^{(j)}_{1,k} - \bar{\theta}_{1,k})^2$$

with $\theta^{(j)}_{1,k}$ j-th sampled value for parameter θ_1 in the k-th Markov chain and $\bar{\theta}_{1,k}$ sample mean for parameter θ_1 in the k-th chain.

Checking for convergence

- If $\bar{\bar{\theta}}_1$ is the overall sample mean for θ_1 across the m Markov chains, then we can compute the sample variance for θ_1 between the m Markov chains as

$$s^2_{\theta_1,B} = \frac{S}{m-1}\sum_{k=1}^{m}(\bar{\theta}_{1,k} - \bar{\bar{\theta}}_1)^2$$

If we call $s^2_{\theta_1,W}$ the average of the m sample variances within each of the chains, then

$$s^2_{\theta_1,W} = \frac{1}{m}\sum_{k=1}^{m} s^2_{\theta_1,k}$$

- Then, we can estimate the marginal variance of the parameter θ_1 as:

$$\widehat{\text{Var}}(\theta_1|\text{data}) = \frac{S-1}{S}s^2_{\theta_1,W} + \frac{1}{S}s^2_{\theta_1,B}$$

Checking for convergence

- Gelman and Rubin suggest that if the ratio

$$\sqrt{\frac{\widehat{\text{Var}}(\theta_1|\text{data})}{s^2_{\theta_1,W}}} \to 1 \qquad \text{as } S \to \infty$$

  then we can assume that the Markov chain for θ_1 has reached its stationary distribution.
- If we have multiple parameters, we need to assess that this holds for each parameter.
- This statistic is computed in the `R` package `coda`: look at the function `gelman.diag`.

Multivariate normal data: example

- Let's consider the reading comprehension example: 22 children were given two tests, one before and one after they received a particular type of instructional method.
- We model these $n = 22$ pairs of scores as an i.i.d. sample from a bivariate normal distribution with mean the 2×1 vector θ and with covariance matrix the 2×2 matrix $\mathbf{\Sigma}$.
- We want to infer upon both θ and $\mathbf{\Sigma}$.
- We place independent priors on θ and $\mathbf{\Sigma}$.
- Precisely, we place a bivariate normal prior on θ, $N_2(\mu_0, \mathbf{\Lambda}_0)$ where

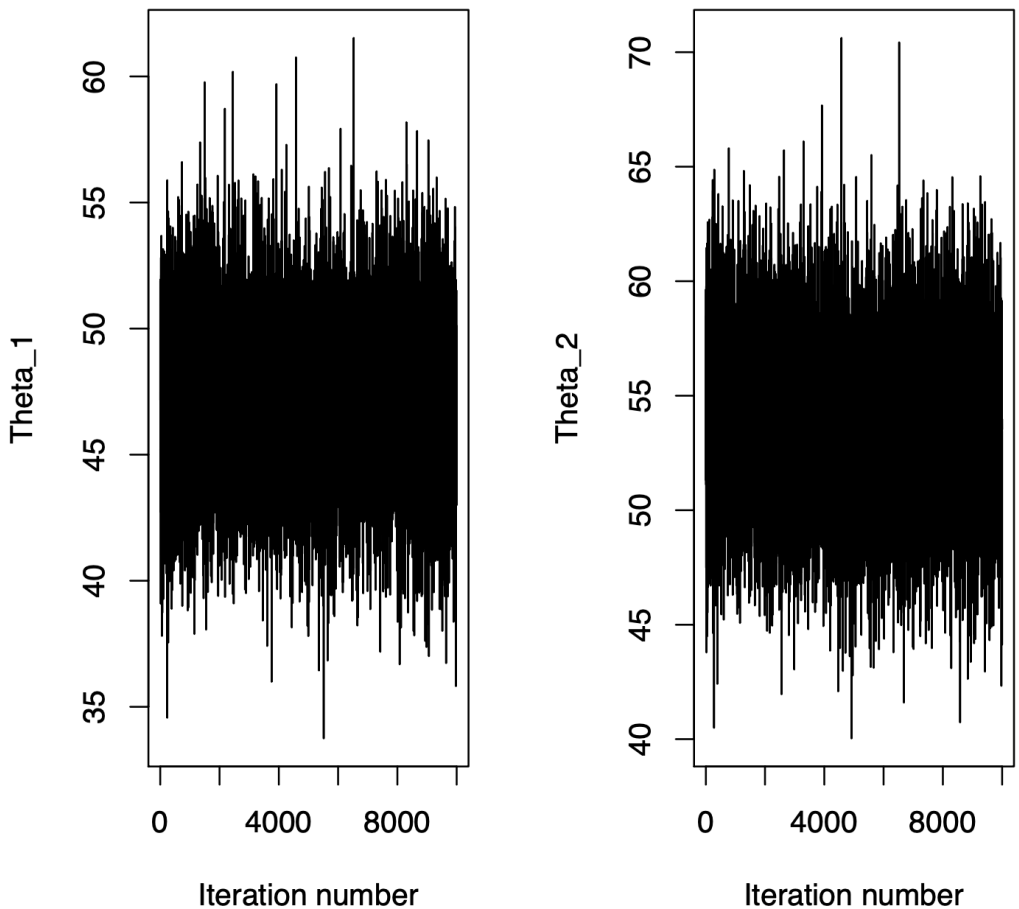
$$\mu_0 = \begin{pmatrix} 50 \\ 50 \end{pmatrix} \qquad \mathbf{\Lambda}_0 = \begin{pmatrix} (\frac{50}{2})^2 = 625 & 312.5 \\ 312.5 & 625 \end{pmatrix}$$

Multivariate normal data: example

- We chose the diagonal elements of $\mathbf{\Lambda}_0$ equal to 625 because we want the scores to be within the range $0-100$, so to ensure that we have a 95% probability that the scores on both test are within the range $0-100$.
- As for the off-diagonal elements of $\mathbf{\Lambda}_0$, we determined them by assuming that there is a correlation of 0.5 between the two scores. This means that the covariance between the two scores is equal to $0.5 \cdot \frac{50}{2} \cdot \frac{50}{2} = 312.5$.
- Finally, we chose the prior on $\mathbf{\Sigma}$ to be an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ where we took $\mathbf{\Phi}_0$ to be equal to $\mathbf{\Lambda}_0$ and $\nu_0 = 4$ so that the prior is loosely concentrated around $\mathbf{\Phi}_0$.
- We ran the Gibbs sampling algorithm for $S = 10,000$ iterations using as initial values for θ and $\mathbf{\Sigma}$, respectively, the 2×1 vector $\bar{\mathbf{y}}$ with the sample means on the two tests, $\bar{\mathbf{y}} = \begin{pmatrix} 47.18 \\ 53.86 \end{pmatrix}$, and the 2×2 matrix with the sample variances and covariance, $\begin{pmatrix} 182.16 & 148.41 \\ 148.41 & 243.65 \end{pmatrix}$.

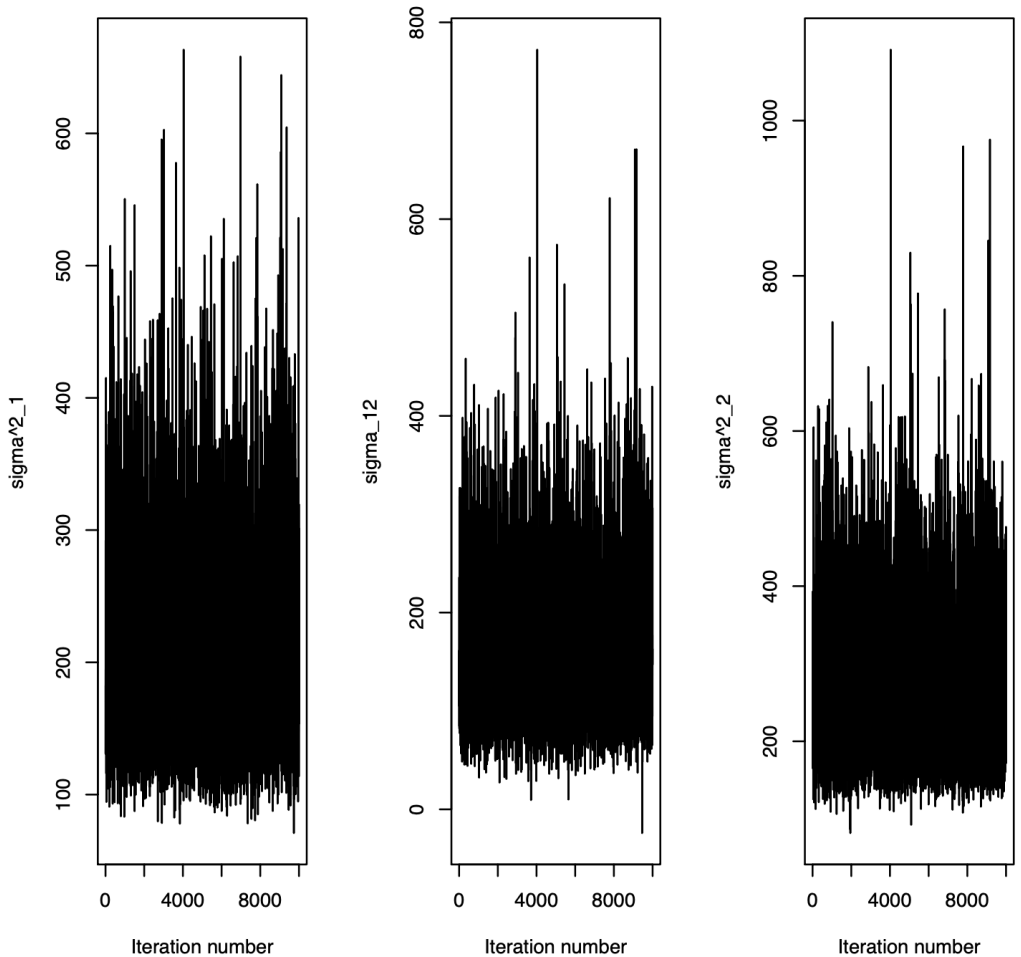
Multivariate normal data: example

- We have 5 parameters to monitor: $\theta_1, \theta_2, \sigma_1^2, \sigma_{12}, \sigma_2^2$.
- Trace plots for θ_1, θ_2
- From the traceplots, it looks as if either the chains are mixing poorly or they have converged very quickly.



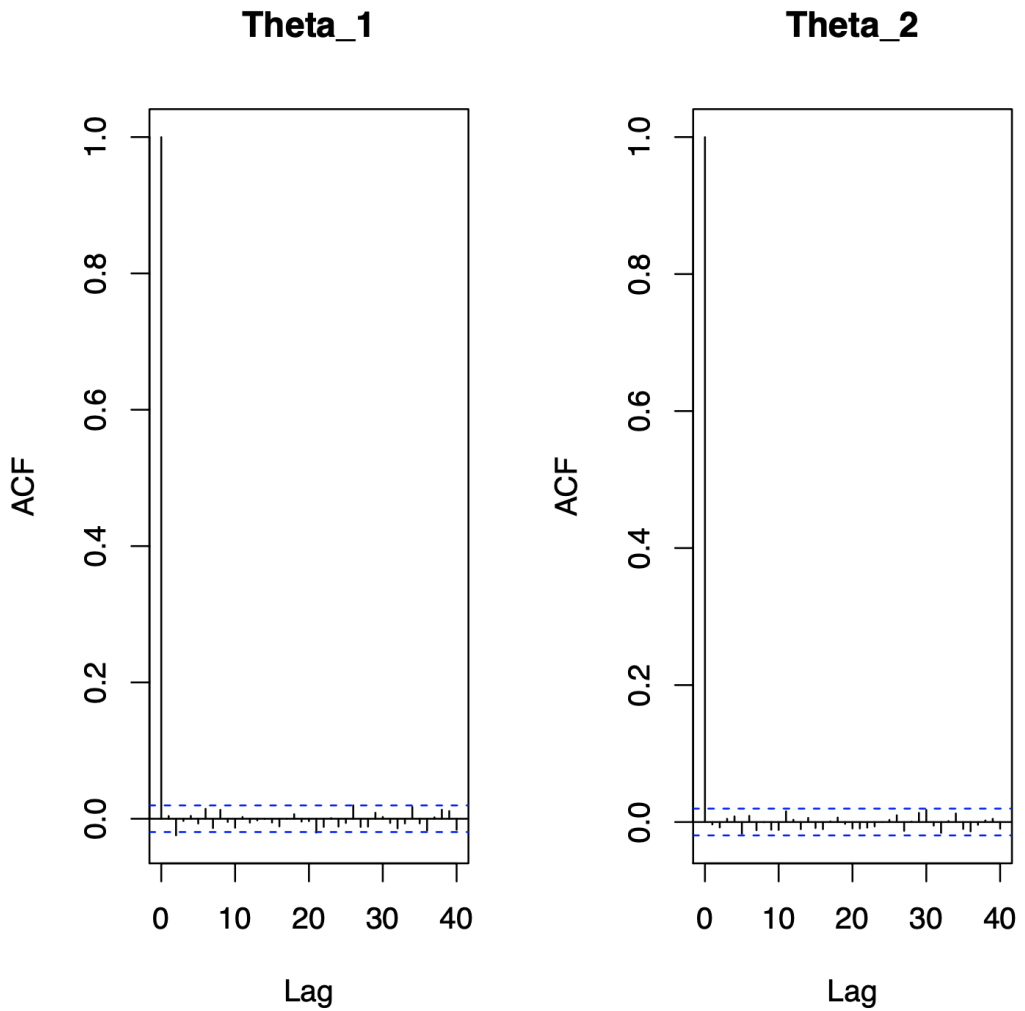
Multivariate normal data: example

- Trace plots for $\sigma_1^2, \sigma_2^2, \sigma_{12}$
- We make the same conclusions as for θ.



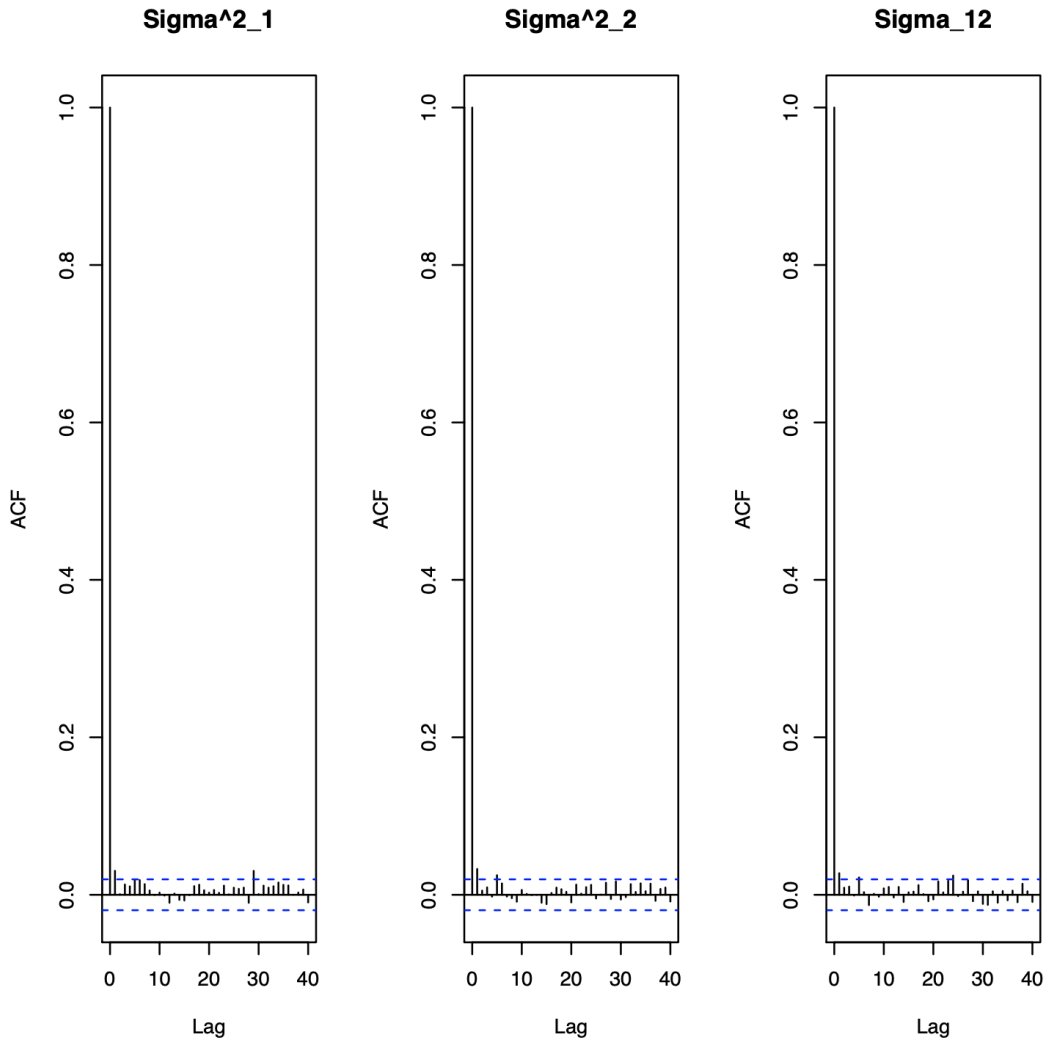
Multivariate normal data: example

- Plots of the autocorrelation function for θ_1, θ_2
- The samples for θ_1, θ_2 are basically independent.



Multivariate normal data: example

- Plots of the autocorrelation function for $\sigma_1^2, \sigma_2^2, \sigma_{12}$
- The autocorrelation is not significantly different from 0 at lags greater than 2.



Multivariate normal data: example

- From the autocorrelation plots, we can conclude that the effective sample size for each parameter is going to be approximately $S = 10,000$, the number of iterations.
- Taking a conservative approach, we discard the first $5,000$ iterations for burn-in and we use the second half for inference.
- We are interested in looking at: (i) the difference between $\theta_2 - \theta_1$ and (ii) the correlation $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \cdot \sigma_2^2}}$ between the test scores before and after the "treatment".

Multivariate normal data: example

- Below are posterior summaries for the difference in the means $\theta_2 - \theta_1$ and for the correlation $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \cdot \sigma_2^2}}$

Parameter	Posterior Mean	Posterior Median	95% Credible Interval	Posterior standard deviation
$\theta_2 - \theta_1$	6.52	6.51	(1.35; 11.66)	2.6
ρ	0.67	0.68	(0.41; 0.85)	0.11

Posterior predictive distribution

- If we want to predict the vector $\tilde{\mathbf{y}}$ with the before and after score for another child **given** the observed scores $\mathbf{y}_1, \ldots, \mathbf{y}_n$ for the 22 children, we would have to use the posterior predictive distribution $p(\tilde{\mathbf{y}}|\mathbf{y}_1, \ldots, \mathbf{y}_n)$.
- This can be evaluated using the same approach as before, that is:

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{y}_1, \ldots, \mathbf{y}_n) &= \int p(\tilde{\mathbf{y}}, \theta, \boldsymbol{\Sigma}|\mathbf{y}_1, \ldots, \mathbf{y}_n) d\theta d\boldsymbol{\Sigma} \\ &= \int p(\tilde{\mathbf{y}}|\theta, \boldsymbol{\Sigma}, \mathbf{y}_1, \ldots, \mathbf{y}_n) \cdot p(\theta, \boldsymbol{\Sigma}|\mathbf{y}_1, \ldots, \mathbf{y}_n) d\theta d\boldsymbol{\Sigma} \\ &= \int p(\tilde{\mathbf{y}}|\theta, \boldsymbol{\Sigma}) \cdot p(\theta, \boldsymbol{\Sigma}|\mathbf{y}_1, \ldots, \mathbf{y}_n) d\theta d\boldsymbol{\Sigma} \end{aligned}$$

Posterior predictive distribution

- In reality we are not going to evaluate the integral that defines the posterior predictive distribution and we approximate it with the kernel density of a sample of $\tilde{\mathbf{y}}^{(1)}, \ldots, \tilde{\mathbf{y}}^{(S)}$ drawn using composition sampling.
- Thus, within our Gibbs sampling algorithm, we are going to add an extra step and the algorithm is now going to look as follows:
  - Choose a number S of iterations
  - Repeat the following three steps for $j = 1, \ldots, S$
    - sample $\theta^{(j)} \sim p(\theta | \mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{\Sigma}^{(j-1)})$
    - sample $\mathbf{\Sigma}^{(j)} \sim p(\mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_n, \theta^{(j)})$
    - sample $\tilde{\mathbf{y}}^{(j)} \sim N_p(\theta^{(j)}, \mathbf{\Sigma}^{(j)})$