Bayesian Statistics

Chapter 7. Multivariate Normal Model

Hojin Yang

Department of Statistics
Pusan National University

Introduction

- Up until now all of our statistical models have been univariate models, that is, models for a single measurement on each member of a sample of individuals or each run of a repeated experiment
- This chapter covers what is perhaps the most useful model for multivariate data, the multivariate normal model, which allows us to jointly estimate population means, variances and correlations of a collection of variables

7.1. Multivariate Normal Density

- Let $Y_{i,1}$ and $Y_{i,2}$ be two pre- and post-scores for the ith student

- Denote each student's pair of scores as a 2×1 vector $\mathbf{Y}_i$

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

- the population mean $\boldsymbol{\theta}$

$$\mathrm{E}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{E}[Y_{i,1}] \\ \mathrm{E}[Y_{i,2}] \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

- the covariance matrix Σ

$$\Sigma = \mathrm{Cov}[Y] = \begin{pmatrix} \mathrm{E}[Y_1^2] - \mathrm{E}[Y_1]^2 & \mathrm{E}[Y_1 Y_2] - \mathrm{E}[Y_1]\mathrm{E}[Y_2] \\ \mathrm{E}[Y_1 Y_2] - \mathrm{E}[Y_1]\mathrm{E}[Y_2] & \mathrm{E}[Y_2^2] - \mathrm{E}[Y_2]^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

- Having information about $\boldsymbol{\theta}$ and Σ may help us in assessing the effectiveness of the teaching method, for instance, $\theta_1 - \theta_2$
- The correlation coefficient $\rho_{1,2} = \sigma_{1,2}/\sqrt{\sigma_1^2\sigma_2^2}$
- Notice that $\boldsymbol{\theta}$ and Σ are both functions of population moments

$$
\begin{array}{ll}
\text{first-order moments:} & \mathrm{E}[Y_1], \mathrm{E}[Y_2] \\
\text{second-order moments:} & \mathrm{E}[Y_1^2], \mathrm{E}[Y_1 Y_2], \mathrm{E}[Y_2^2]
\end{array}
$$

- **Y** has a multivariate normal distribution if its sampling density is given by

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-(\boldsymbol{y}-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\theta})/2\}$$

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

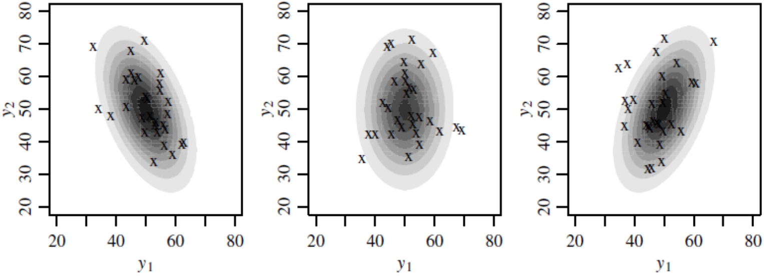
- For a $p \times 1$ vector b and $p \times p$ matrix A,

$$b^T A = (\sum_{j=1}^{p} b_j a_{j,1}, \ldots, \sum_{j=1}^{p} b_j a_{j,p})$$

- $b^T Ab$ is the single number

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

- Figure gives contour plots and 30 samples from each of three different two-dimensional multivariate normal densities



- $\boldsymbol{\theta} = (50, 50)^T$, $\sigma_1^2 = 64$, $\sigma_2^2 = 144$ but $\sigma_{1,2}$ varies from plot to plot, with -48, 0, 48 (giving correlations of $-.5$, 0 and .5 respectively)

- The marginal distribution of each variable Y_j is a univariate normal distribution, with mean θ_j and variance σ_j^2

7.2. Semiconjugate Prior Distribution for Mean

- Convenient prior distribution for the multivariate mean $\boldsymbol{\theta}$ is a multivariate normal (MN) distribution

$$p(\boldsymbol{\theta}) \sim MN(\boldsymbol{\mu}_0, \Lambda_0)$$

- We need full conditional dist of $\boldsymbol{\theta}$, given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and Σ
- Let us examine the prior distribution as a function of $\boldsymbol{\theta}$

$$\begin{aligned}
p(\boldsymbol{\theta}) &= (2\pi)^{-p/2}|\Lambda_0|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T\Lambda_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)\} \\
&= (2\pi)^{-p/2}|\Lambda_0|^{-1/2}\exp\{-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0-\frac{1}{2}\boldsymbol{\mu}_0^T\Lambda_0^{-1}\boldsymbol{\mu}_0\} \\
&\propto \exp\{-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0\} \\
&= \exp\{-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}_1\boldsymbol{\theta}+\boldsymbol{\theta}^T b_1\},
\end{aligned}$$

where $\mathbf{A}_1 = \Lambda_0^{-1}$ and $\mathbf{b}_1 = \Lambda_0^{-1}\boldsymbol{\mu}_0$

- Conversely, it says if θ has a density that is proportional to $\exp\{-\theta^T \mathbf{A}\theta/2 + \theta^T \mathbf{b}\}$ for some $\mathbf{A}$ and $\mathbf{b}$, then θ must have a MVN with covariance $\mathbf{A}^{-1}$ and mean $\mathbf{A}^{-1}\mathbf{b}$

- If our sampling model $\mathbf{Y}_1, \ldots, \mathbf{Y}_n | \theta, \Sigma \sim MVN(\theta, \Sigma)$

- Then similar calculations show that the joint sampling density of the observed vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is

$$
\begin{aligned}
p(y_1, \ldots, y_n | \theta, \Sigma) &= \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-(y_i - \theta)^T \Sigma^{-1} (y_i - \theta)/2\} \\
&= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{1}{2} \sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)\} \\
&\propto \exp\{-\frac{1}{2} \theta^T \mathrm{A}_1 \theta + \theta^T b_1\},
\end{aligned}
$$

where $\mathbf{A}_1 = n\Sigma^{-1}$, $\mathbf{b}_1 = n\Sigma^{-1}\bar{\mathbf{y}}$ and $\bar{\mathbf{y}} = (\frac{1}{n}\sum_{i=1}^{n} y_{i,1}, \ldots, \frac{1}{n}\sum_{i=1}^{n} y_{i,p})^T$

- Combining Equations likelihood and prior gives

$$
\begin{aligned}
p(\theta|y_1,\ldots,y_n,\Sigma) &\propto \exp\{-\frac{1}{2}\theta^T \mathrm{A}_0\theta + \theta^T b_0\} \times \exp\{-\frac{1}{2}\theta^T \mathrm{A}_1\theta + \theta^T b_1\} \\
&= \exp\{-\frac{1}{2}\theta^T \mathrm{A}_n\theta + \theta^T b_n\}, \text{ where} \\
\mathrm{A}_n &= \mathrm{A}_0 + \mathrm{A}_1 = \Lambda_0^{-1} + n\Sigma^{-1} \text{ and} \\
b_n &= b_0 + b_1 = \Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}
\end{aligned}
$$

- This implies that the conditional dist of θ must be MVN($\mathbf{A}_n^{-1}\mathbf{b}_n, \mathbf{A}_n^{-1}$). So,

$$
\begin{aligned}
\mathrm{Cov}[\theta|y_1,\ldots,y_n,\Sigma] &= \Lambda_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1} \\
\mathrm{E}[\theta|y_1,\ldots,y_n,\Sigma] &= \mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) \\
p(\theta|y_1,\ldots,y_n,\Sigma) &= \text{multivariate normal}(\mu_n, \Lambda_n).
\end{aligned}
$$

7.3. Inverse-Wishart Distribution

- Just as a variance σ^2 must be positive, a variance-covariance matrix Σ must be positive definite, meaning that

$$x^T \Sigma x > 0 \;\; \text{for all vectors not equal to zero}$$

- Another requirement of our covariance matrix is that it is symmetric

- The sum of squares matrix of a collection of multivariate vectors $z_1, \ldots, z_n$ is given by

$$\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T = \mathbf{Z}^T \mathbf{Z}$$

  where $\mathbf{Z}$ is the $n \times p$ matrix whose ith row is z_i^T

- Since z_i is a $p \times 1$ vector, $z_i z_i^T$ can be thought of

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

- If the z_i's are samples from a population with zero mean, we can think of the matrix $z_i z_i^T / n$ as the contribution of vector z_i to the estimate of the covariance matrix

$$\begin{aligned} \tfrac{1}{n}[\mathbf{Z}^T\mathbf{Z}]_{j,j} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n z_{i,j}^2 = s_{j,j} = s_j^2 \\ \tfrac{1}{n}[\mathbf{Z}^T\mathbf{Z}]_{j,k} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n z_{i,j}z_{i,k} = s_{j,k} \ . \end{aligned}$$

- If $n > p$ and the z_i's are linearly independent, then $\mathbf{Z}^T\mathbf{Z}$ will be positive definite and symmetric

- This suggests the following construction of a “random” covariance matrix
- For a given positive integer ν_0 and a $p \times p$ covariance matrix Φ_0

  1. sample $z_1, \ldots, z_{\nu_0} \sim$ i.i.d. multivariate normal$(\mathbf{0}, \Phi_0)$;
  2. calculate $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{\nu_0} z_i z_i^T$.

- We can repeat this procedure over and over again, generating matrices $\mathbf{Z}_1^T\mathbf{Z}_1, \ldots, \mathbf{Z}_S^T\mathbf{Z}_S$
- The population distribution of these sum of squares matrices is called a Wishart distribution with (ν_0, Φ_0)

- If $\mathbf{Z}^T\mathbf{Z} \sim W(\nu_0, \Phi_0)$, the following properties hold

  If $\nu_0 > p$, then $\mathbf{Z}^T\mathbf{Z}$ is positive definite with probability 1
  $\mathbf{Z}^T\mathbf{Z}$ is symmetric with probability 1.
  $\mathrm{E}[\mathbf{Z}^T\mathbf{Z}] = \nu_0\Phi_0$.

- Wishart distribution is a multivariate analogue of the gamma distribution

- Likewise $\sigma^2 \sim IG(a, b)$ and $1/\sigma^2 \sim G(a, b)$, we model the covariance matrix $\Sigma \sim IW(\nu_0, \Phi_0)$ called inverse-Wishart distribution, whereas the precision matrix $\Sigma^{-1} \sim W(\nu_0, \Phi_0)$

- We consider reparameterization, to sample a covariance matrix from an inverse Wishart

  1. sample $z_1, \ldots, z_{\nu_0} \sim$ i.i.d. multivariate normal$(\mathbf{0}, \mathbf{S}_0^{-1})$;
  2. calculate $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{\nu_0} z_i z_i^T$;
  3. set $\Sigma = (\mathbf{Z}^T\mathbf{Z})^{-1}$.

- Under this, $\Sigma^{-1} \sim W(\nu_0, S_0^{-1})$ and $\Sigma \sim IW(\nu_0, S_0^{-1})$
- Their expectations are

$$\mathrm{E}[\Sigma^{-1}] = \nu_0 \mathbf{S}_0^{-1}$$

$$\mathrm{E}[\Sigma] = \frac{1}{\nu_0 - p - 1}(\mathbf{S}_0^{-1})^{-1} = \frac{1}{\nu_0 - p - 1}\mathbf{S}_0$$

Full Conditional Distribution of Covariance Matrix

- $IW(\nu_0, S_0^{-1})$ density is given by

$$p(\Sigma) = \left[2^{\nu_0 p/2} \pi^{\binom{p}{2}/2} |\mathrm{S}_0|^{-\nu_0/2} \prod_{j=1}^{p} \Gamma([\nu_0 + 1 - j]/2) \right]^{-1} \times$$
$$|\Sigma|^{-(\nu_0+p+1)/2} \times \exp\{-\mathrm{tr}(\mathrm{S}_0 \Sigma^{-1})/2\}$$

- We now need to combine the above prior distribution with the sampling distribution for $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$

$$p(y_1, \ldots, y_n | \theta, \Sigma) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)/2\}$$

- An interesting result from matrix algebra is that the sum $\sum_{k=1}^{K} \mathbf{b}_k^T \mathbf{A} \mathbf{b}_k = tr(\mathbf{B}^T \mathbf{A} \mathbf{B})$, where $\mathbf{B}^T$ is the matrix whose kth row is $\mathbf{b}_k^T$
- This means that

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) = \mathrm{tr}(\mathrm{S}_\theta \Sigma^{-1}), \text{ where}$$

$$\mathrm{S}_\theta = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T$$

- The matrix S_θ is the residual sum of squares matrix for the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ when θ was population mean

- Conditional distribution of Σ can be shown as

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

- Thus we have

$$\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\boldsymbol{\theta} \sim IW(\nu_0+n,[S_0+S_\theta]^{-1})$$

- We can think of ν_0+n, as the posterior sample size

- $[S_0+S_\theta]$ can be thought of as the prior residual sum of squares plus the residual sum of squares from the data

- Conditional expectation of Σ

$$\mathrm{E}[\Sigma|y_1,\ldots,y_n,\theta] = \frac{1}{\nu_0+n-p-1}(\mathrm{S}_0+\mathrm{S}_\theta)$$
$$= \frac{\nu_0-p-1}{\nu_0+n-p-1}\frac{1}{\nu_0-p-1}\mathrm{S}_0 + \frac{n}{\nu_0+n-p-1}\frac{1}{n}\mathrm{S}_\theta$$

- Conditional expectation can be seen as a weighted average of the prior expectation and the unbiased estimator

- Because it can be shown that S_θ converges to the true population covariance matrix, the posterior expectation of Σ is a consistent estimator of the population covariance

7.4. Gibbs Sampling of θ and Σ

- We showed that

$$\boldsymbol{\theta}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma \sim MVN(\boldsymbol{\mu}_n, \Lambda_n)$$
$$\Sigma|\mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\theta} \sim IW(\nu_n, S_n^{-1})$$

- We have results for $\{\boldsymbol{\mu}_n, \Lambda_n\}$, $\nu_n = \nu_0 + n$, $S_n = S_0 + S_\theta$
- Full conditional distributions can be used to construct a Gibbs sampler, providing us with an MCMC approximation to the joint posterior distribution $p(\boldsymbol{\theta}, \Sigma|\mathbf{y}_1, \ldots, \mathbf{y}_n)$
- Given a starting value $\Sigma^{(0)}$, the Gibbs sampler generates $\{\boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)}\}$ from $\{\boldsymbol{\theta}^{(s)}, \Sigma^{(s)}\}$ via the following two steps

- We generate

  step 1: Sample $\boldsymbol{\theta}^{(s+1)}$from its full conditional distribution:

  step 1.1: Compute $\boldsymbol{\mu}_n$ and Λ_n from $\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma^{(s)}$

  step 1.2: Sample $\boldsymbol{\theta}^{(s+1)} \sim MVN(\boldsymbol{\mu}_n, \Lambda_n)$

  step 2: Sample $\Sigma^{(s+1)}$from its full conditional distribution:

  step 2.1: Compute S_n from $\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma^{(s)}$

  step 2.2: Sample $\Sigma^{(s+1)} \sim IW(\nu_0 + n, S_n^{-1})$

- Note that $\{\boldsymbol{\mu}_n, \Lambda_n\}$ depend on Σ, and S_n depends on $\boldsymbol{\theta}$
- Hence, these quantities need to be recalculated at every iteration of the sampler

Example

- Data: 22 children were given two exams, one before a certain type of instruction and one after
- Model these 22 pairs of scores as samples from MVN
- 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 expectation
- Since the true mean cannot be below 0 or above 100, it is desirable to use a prior variance for $\boldsymbol{\theta}$ 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 $p(\theta_i \notin [0, 100]) = 0.05$

- Finally, since the two exams are measuring similar things, it is probable that θ_1 and θ_2 are close, which prior correlation of 0.5, so that $\lambda_{1,2} = 312.5$

- We'll take S_0 to be the same as Λ_0 and take $\nu_0 = p + 2 = 4$

```
mu0<-c(50,50)
L0<-matrix(c(625,312.5,312.5,625),nrow=2,ncol=2)

nu0<-4
S0<-matrix(c(625,312.5,312.5,625),nrow=2,ncol=2)
```

- We observed $\mathbf{y} = (47.18, 53.86)^T$, $s_1^2 = 182.16$, $s_2^2 = 243.65$, $s_{1,2}/(s_1 s_2) = 0.70$

- Let's use the Gibbs sampler described above to combine this sample information with our prior distributions to obtain estimates and confidence intervals for the population parameters

- We begin by setting $\Sigma^{(0)}$ equal to the sample covariance matrix, and iterating from there

```
data(chapter7) ; Y<-Y.reading
n<-dim(Y)[1] ; ybar<-apply(Y,2,mean)
Sigma<-cov(Y) ; 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(1, nu0+n, solve(Sn)) )
  ###

  ### save results
  THETA<-rbind(THETA,theta) ; SIGMA<-rbind(SIGMA,c(Sigma))
  ###

}
```

- $\{(\boldsymbol{\theta}^{(1)}, \Sigma^{(1)}), \ldots, (\boldsymbol{\theta}^{(5000)}, \Sigma^{(5000)})\}$ are generated
- From these samples we can approximate posterior probabilities and confidence regions of interest.

```
> quantile(   THETA[,2]-THETA[,1], prob=c(.025,.5,.975) )
     2.5%       50%     97.5%
 1.513573  6.668097 11.794824


> mean( THETA[,2]>THETA[,1])
[1] 0.9942
```

- The posterior probability $p(\theta_2 > \theta_1 | \mathbf{y}_1, \ldots, \mathbf{y}_n,) = 0.99$ indicates strong evidence that, the average score on the second exam would be higher than that on the first
- There is a “highly significant difference” in exam scores before and after the instruction

- What is the probability that a randomly selected child will score higher on the second exam than on the first?
- We can compute $p(\tilde{Y}_2 > \tilde{Y}_1 | \mathbf{y}_1, \ldots, \mathbf{y}_n,) = 0.71$
- This says that almost a third of the students will get a lower score on the second exam
- Be careful about the difference between these two probabilities

7.5. Missing Data and Imputation

- The NA's stand for "not available," and so some data for some individuals are "missing"

	glu	bp	skin	bmi
1	86	68	28	30.2
2	195	70	33	NA
3	77	82	NA	35.8
4	NA	76	43	47.9
5	107	60	NA	NA
6	97	76	27	NA
7	NA	58	31	34.3
8	193	50	16	25.9
9	142	80	15	NA
10	128	78	NA	43.3

- Let $\mathbf{O}_i = (O_1, \ldots, O_p)$ be a binary vector such that

$$O_{i,j} = 1 \text{ if } Y_{i,j} \text{ is observed}$$
$$O_{i,j} = 0 \text{ if } Y_{i,j} \text{ is missing}$$

- Therefore, we have $\mathbf{O}_i = \mathbf{o}_i$ and $Y_{i,j} = y_{i,j}$ for the ith subject and the jth variable $o_{i,j}$

- Assume that missing data are missing at random (MAR), meaning that $\mathbf{O}_i$ and $\mathbf{Y}_i$ are statistically independent

- Assume that $\mathbf{O}_i$ does not θ and Σ

- If MAR is not satisfied, modeling the relationship between $\mathbf{O}_i$ and $\mathbf{Y}_i$ are required

- If MAR is satisfied, the sampling probability for the *i*th subject is

$$p(o_i, \{y_{i,j} : o_{i,j} = 1\} | \theta, \Sigma) = p(o_i) \times p(\{y_{i,j} : o_{i,j} = 1\} | \theta, \Sigma)$$
$$= p(o_i) \times \int \left\{ p(y_{i,1}, \ldots, y_{i,p} | \theta, \Sigma) \prod_{y_{i,j} : o_{i,j} = 0} dy_{i,j} \right\}$$

- Our sampling probability for data from subject i is $p(\mathbf{O}_i)$ multiplied by the marginal probability of the observed variables, after integrating out the missing variables

- Suppose $\mathbf{y}_i = (y_{i,1}, NA, y_{i,3}, NA)^T$, $\mathbf{o}_i = (1, 0, 1, 0)^T$

$$\begin{aligned} p(\boldsymbol{o}_i, y_{i,1}, y_{i,3} | \boldsymbol{\theta}, \Sigma) &= p(\boldsymbol{o}_i) \times p(y_{i,1}, y_{i,3} | \boldsymbol{\theta}, \Sigma) \\ &= p(\boldsymbol{o}_i) \times \int p(\boldsymbol{y}_i | \boldsymbol{\theta}, \Sigma)\, dy_2\, dy_4 \end{aligned}$$

- We can consider $p(y_{i,1}, y_{i,3} | \boldsymbol{\theta}, \Sigma)$ as *MVN* with mean $(\theta_1, \theta_3)^T$ and covariance consisting of $(\sigma_1^2, \sigma_{1,3}, \sigma_3^2)$
- The parameters $\boldsymbol{\theta}$ and Σ are unknown as usual, but the missing data are also an unknown but key component in Bayesian inference

- The $n \times p$ matrix $\mathbf{Y}$ consists of two parts:

$$\begin{aligned} \mathrm{Y}_{\text{obs}} &= \{y_{i,j} : o_{i,j} = 1\}, \text{ the data that we do observe, and} \\ \mathrm{Y}_{\text{miss}} &= \{y_{i,j} : o_{i,j} = 0\}, \text{ the data that we do not observe.} \end{aligned}$$

- We want $p(\theta, \Sigma, \mathbf{Y}_{miss} | \mathbf{Y}_{obs})$, as the posterior distribution of unknown and unobserved quantities

- Gibbs sampling scheme: Given $\{\Sigma^{(0)}, \mathbf{Y}_{miss}^{(0)}\}$, we generate $\{\theta^{(s+1)}, \Sigma^{(s+1)}, \mathbf{Y}_{miss}^{(s+1)}\}$ from $\{\theta^{(s)}, \Sigma^{(s)}, \mathbf{Y}_{miss}^{(s)}\}$ by

  1. sampling $\theta^{(s+1)}$ from $p(\theta | \mathrm{Y}_{\text{obs}}, \mathrm{Y}_{\text{miss}}^{(s)}, \Sigma^{(s)})$;
  2. sampling $\Sigma^{(s+1)}$ from $p(\Sigma | \mathrm{Y}_{\text{obs}}, \mathrm{Y}_{\text{miss}}^{(s)}, \theta^{(s+1)})$;
  3. sampling $\mathrm{Y}_{\text{miss}}^{(s+1)}$ from $p(\mathrm{Y}_{\text{miss}} | \mathrm{Y}_{\text{obs}}, \theta^{(s+1)}, \Sigma^{(s+1)})$.

- Note that in steps 1 and 2, we used fixed $\mathbf{Y}_{obs}$ and the current value of $\mathbf{Y}_{miss}^{(s)}$ as a complete data $\mathbf{Y}^{(s)}$

- Step 3 is a bit more complicated

$$
\begin{aligned}
p(\mathbf{Y}_{\text{miss}}|\mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}, \Sigma) &\propto p(\mathbf{Y}_{\text{miss}}, \mathbf{Y}_{\text{obs}}|\boldsymbol{\theta}, \Sigma) \\
&= \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}}, \boldsymbol{y}_{i,\text{obs}}|\boldsymbol{\theta}, \Sigma) \\
&\propto \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}}|\boldsymbol{y}_{i,\text{obs}}, \boldsymbol{\theta}, \Sigma)
\end{aligned}
$$

- For each *i* we need to sample the missing elements of the data vector conditional on the observed elements
- This is made possible via the following result about multivariate normal distributions

- Let $y \sim MVN(\theta, \Sigma)$, two sets $a \subset \{1, 2, \ldots, p\}$, $b = a^c$
- If $p = 4$, then perhaps, $a = \{1, 2\}$ and $b = \{3, 4\}$
- We can consider

$$\{y_{[b]}|y_{[a]}, \theta, \Sigma\} \sim \text{multivariate normal}(\theta_{b|a}, \Sigma_{b|a}), \text{ where}$$
$$\theta_{b|a} = \theta_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(y_{[a]} - \theta_{[a]})$$
$$\Sigma_{b|a} = \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}$$

- θ_b refers to the elements of θ corresponding to the indices in b
- $\Sigma_{[a,b]}$ refers to the matrix made up of the elements that are in rows a and columns b of Σ

- In our case, $y = (\text{glu, bp, skin, bmi})$
- If we have glu and bp data for someone ($a = \{1, 2\}$) but are missing skin and bmi measurements ($b = \{3, 4\}$)
- We want to generate $y_{[b]}|y_{[a]}$ from $MVN(\theta_{b|a}, \Sigma_{b|a})$

```
data(chapter7) ; Y<-Y.pima.miss
### prior parameters
n<-dim(Y)[1] ; p<-dim(Y)[2]
mu0<-c(120,64,26,26)
sd0<-(mu0/2)
L0<-matrix(.1,p,p) ; diag(L0)<-1 ; L0<-L0*outer(sd0,sd0)
nu0<-p+2 ; S0<-L0
###

### starting values
Sigma<-S0
Y.full<-Y
O<-1*(!is.na(Y))
for(j in 1:p)
{
  Y.full[is.na(Y.full[,j]),j]<-mean(Y.full[,j],na.rm=TRUE)
}
###
```

- These calculations can be done with R

```
### Gibbs sampler
THETA<-SIGMA<-Y.MISS<-NULL
set.seed(1)
for(s in 1:1000)
{

  ###update theta
  ybar<-apply(Y.full,2,mean)
  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.full)-c(theta) )%*%t( t(Y.full)-c(theta) )
  Sigma<-solve( rwish(1, nu0+n, solve(Sn)) )
  ###

  ###update missing data
  for(i in 1:n)
  {
    b <- ( O[i,]==0 )
    a <- ( O[i,]==1 )
    iSa<- solve(Sigma[a,a])
    beta.j <- Sigma[b,a]%*%iSa
    Sigma.j   <- Sigma[b,b] - Sigma[b,a]%*%iSa%*%Sigma[a,b]
    theta.j<- theta[b] + beta.j%*%(t(Y.full[i,a])-theta[a])
    Y.full[i,b] <- rmvnorm(1,theta.j,Sigma.j )
  }
```

```
  ### save results
  THETA<-rbind(THETA,theta) ; SIGMA<-rbind(SIGMA,c(Sigma))
  Y.MISS<-rbind(Y.MISS, Y.full[O==0] )
  ###
}
###
```

- Thereby, the Monte Carlo approximation of $E[\theta|y_1, ..., y_n]$ is available as $(123.46, 71.03, 29.35, 32.18)$