

Module 9: The Multivariate Normal Distribution and Missing Data

Rebecca C. Steorts

Agenda

- ▶ Review of model
- ▶ Introduction to Pima Indian data set
- ▶ Approach to handling missing at random data
- ▶ Application to Pima Indian data set

Model set up

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

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

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

Pima Indian heritage data

We consider a dataset involving health-related measurements on 200 women of Pima Indian heritage living near Phoenix, Arizona (Smith et al, 1988).

The four variables are `glu` (blood plasma glucose concentration), `bp` (diastolic blood pressure), `skin` (skin fold thickness) and `bmi` (body mass index).

Missing data

- ▶ The NA's stand for "not available," and so some data for some individuals are "missing."
- ▶ Missing data are fairly common in survey data and other data sets.

```
head(Y)
```

```
##    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
```

Simple Approaches

1. Many software packages either throw away all subjects with incomplete data.
 - ▶ We don't want to throw away data!
2. Others impute missing values with a population mean or some other fixed value, then proceed with the analysis.
 - ▶ This approach is statistically incorrect, as it says we are certain about the values of the missing data when in fact we have not observed them.

Notation

Let $\mathbf{O}_i = (O_1, \dots, O_p)^T$ be a binary vector of 0's and 1's.

Specifically, $O_{ij} = 1$ if Y_{ij} is observed and not missing.

$O_{ij} = 0$ if Y_{ij} is missing.

Our observed information about subject i is $\mathbf{O}_i = \mathbf{o}_i$ and $Y_{ij} = y_{ij}$ for variable j such that $o_{ij} = 1$.

Missing at Random

- ▶ We assume the data are missing at random, meaning that \mathbf{O}_i and \mathbf{Y}_i are independent and the distribution of \mathbf{O}_i does not depend on θ or Σ .
- ▶ For modeling the data in a way that missing not a random, refer to Chapter 21 of Gelman et al (2004).

Missing at Random

The sampling probability for data for subject i is:

$$p(\mathbf{o}_i, \{y_{ij} : o_{ij} = 1\} \mid \boldsymbol{\theta}, \Sigma) \quad (1)$$

$$= p(\mathbf{o}_i) \times p(\{y_{ij} : o_{ij} = 1\} \mid \boldsymbol{\theta}, \Sigma) \quad (2)$$

$$= p(\mathbf{o}_i) \times \int \left\{ p(\{y_{i,1}, \dots, y_{ip} \mid \boldsymbol{\theta}, \Sigma) \prod_{y_{ij}: o_{ij}=0} dy_{ij} \right\} \quad (3)$$

Missing at Random

Let's look at a special case so that the example is more concrete.

Let $\mathbf{y}_i = (y_{i1}, NA, y_{i3}, NA)^T$.

Then $\mathbf{o}_i = (1, 0, 1, 0)^T$.

So, when data are missing at random we integrate over the missing data to obtain the marginal probability of the observed data:

$$p(\mathbf{o}_i, y_{i1}, y_{i3} \mid \boldsymbol{\theta}, \Sigma) = p(\mathbf{o}_i) p(y_{i1}, y_{i3} \mid \boldsymbol{\theta}, \Sigma) \quad (4)$$

$$= p(\mathbf{o}_i) \int p(\mathbf{y}_i \mid \boldsymbol{\theta}, \Sigma) dy_2 dy_4 \quad (5)$$

Notation

Let $\mathbf{Y}_{n \times p}$ be the matrix of all potential data in which $o_{ij} = 1$ if Y_{ij} is observed and $o_{ij} = 0$ if Y_{ij} is missing.

We can think of the matrix as consisting of two parts:

1.

$$\mathbf{Y}_{obs} = \{y_{ij} : o_{ij} = 1\}$$

2.

$$\mathbf{Y}_{miss} = \{y_{ij} : o_{ij} = 0\}$$

From the observed data, we want to obtain $p(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \mathbf{Y}_{miss} \mid \mathbf{Y}_{obs})$.

Gibbs sampler

Suppose the Gibbs sampler is at iteration s .

1. Sample $\theta^{(s+1)}$ from it's full conditional:
 - a) Compute μ_n and Σ_n from \mathbf{Y}_{obs} , \mathbf{Y}_{miss} 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}_{obs} , \mathbf{Y}_{miss} , and $\theta^{(s+1)}$
 - b) Sample $\Sigma^{(s+1)} \sim \text{inverseWishart}(\nu_n, S_n^{-1})$
3. Sample

$$\mathbf{Y}_{miss}^{s+1} \sim p(\mathbf{Y}_{miss} \mid \mathbf{Y}_{obs}, \theta^{(s+1)}, \Sigma^{(s+1)})$$

In steps 1–2, Note the fixed value of \mathbf{Y}_{obs} combines with the current value of $\mathbf{Y}_{miss}^{(s)}$ to form a current version of a complete data matrix $\mathbf{Y}^{(s)}$ with no missing values.

Step 3 of Gibbs sampler

$$p(\mathbf{Y}_{miss} \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}, \Sigma) \propto p(\mathbf{Y}_{miss} \mathbf{Y}_{obs} \mid \boldsymbol{\theta}, \Sigma) \quad (6)$$

$$= \prod_{i=1}^n p(\mathbf{y}_{i,miss}, \mathbf{y}_{i,obs} \mid \boldsymbol{\theta}, \Sigma) \quad (7)$$

$$\propto \prod_{i=1}^n p(\mathbf{y}_{i,miss} \mid \mathbf{y}_{i,obs} \mid \boldsymbol{\theta}, \Sigma) \quad (8)$$

We can compute the above quantity using a fact from multivariate methods.

Multivariate fact

Let $\mathbf{y}_{[b]}$ and $\mathbf{y}_{[a]}$ correspond to the elements of \mathbf{y} corresponding to the indices of a and b . Let $\Sigma_{[a,b]}$ be a matrix with rows a and columns b .

Knowing information about partitioned matrices, one can show that

$$\mathbf{y}_{[b]} \mid \mathbf{y}_{[a]}, \boldsymbol{\theta}, \Sigma \sim \text{MVN}(\boldsymbol{\theta}_{b|a}, \Sigma_{b|a}),$$

where

$$\boldsymbol{\theta}_{b|a} = \boldsymbol{\theta}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\mathbf{y}_{[a]} - \boldsymbol{\theta}_{[a]})$$

and

$$\Sigma_{[b,a]} = \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}.$$

Application to Pima Indian data set

We first talk about hyper-parameter selection and then implement the Gibbs sampler.

Recall the full model is

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

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

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

Hyper-parameter selection

The prior mean of $\mu_0 = (120, 64, 26, 26)^T$ is taken from national averages.

The corresponding prior variances are based primarily on keeping most of the prior mass on values that are above zero.

These prior distributions are likely much more diffuse than more informed prior distributions that could be provided by an expert in this field of study.

The data set

```
head(Y)
```

```
##    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
```

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

```
p <- dim(Y)[2]
```

Prior parameter specification

```
mu0 <- c(120,64,26,26)
(sd0 <- mu0/2)
```

```
## [1] 60 32 13 13
```

```
(L0 <- matrix(0.1, p,p))
```

```
##      [,1] [,2] [,3] [,4]
## [1,] 0.1  0.1  0.1  0.1
## [2,] 0.1  0.1  0.1  0.1
## [3,] 0.1  0.1  0.1  0.1
## [4,] 0.1  0.1  0.1  0.1
```

```
diag(L0) <- 1
L0 <- L0*outer(sd0,sd0)
nu0 <- p + 2
S0 <- L0
```

Starting values

```
Sigma <- S0
Y.full <- Y
# put a -1 for observed values
# 0 for NA's
O <- -1*(!is.na(Y))

# replace the NA values with #average of all the observed
# values in column j

for(j in 1:p){
Y.full[is.na(Y.full[,j]),j]<- mean(Y.full[,j],na.rm=TRUE)
}
```