

# Multivariate Normal Models, Missing Data and Imputation

STA702

Merlise Clyde  
Duke University

<https://sta702-F23.github.io/website/>



# Introduction to Missing Data

- Missing data/nonresponse is fairly common in real data.
  - Failure to respond to survey question
  - Subject misses some clinic visits out of all possible
  - Only subset of subjects asked certain questions
- posterior computation usually depends on the data through  $p(Y | X, \theta)$ , which can be difficult to compute (at least directly) when some of the  $y_i$  (multivariate  $Y$ ) or  $x_i^T$  values are missing.
- Most software packages often throw away all subjects with incomplete data (can lead to bias and precision loss).
- Some individuals impute missing values with a mean or some other fixed value (ignores uncertainty).
- Imputing missing data is actually quite natural in the Bayesian context.

# Missing data mechanisms

- Data are said to be **missing completely at random (MCAR)** if the reason for missingness does not depend on the values of the observed data or missing data.
- For example, suppose
  - you handed out a double-sided survey questionnaire of 20 questions to a sample of participants;
  - questions 1-15 were on the first page but questions 16-20 were at the back; and
  - some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be **MCAR** if they simply did not realize the pages were double-sided; they had no reason to ignore those questions.
- **This is rarely plausible in practice!**

# Missing Data Mechanisms

- Data are said to be **missing at random (MAR)** if, conditional on the values of the observed data, the reason for missingness does not depend on the missing data.
- Using our previous example, suppose
  - questions 1-15 include demographic information such as age and education;
  - questions 16-20 include income related questions; and
  - once again, some participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be **MAR** if younger people are more likely not to respond to those income related questions than old people, where age is observed for all participants. (missingness reason must be independent of income)
- **This is the most commonly assumed mechanism in practice!**

# Missing data mechanisms

- Data are said to be **missing not at random (MNAR or NMAR)** if the reason for missingness depends on the actual values of the missing (unobserved) data.
- suppose again that
- questions 1-15 include demographic information such as age and education;
- questions 16-20 include income related questions; and
- once again, some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be **MNAR** if people who earn more money are less likely to respond to those income related questions than those with lower incomes.
- **This is usually the case in real data, but analysis can be complex!**

# Multivariate Formulation

- Consider the multivariate data scenario with  $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \dots, \mathbf{Y}_{in})^T$ , where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})^T$ , for  $i = 1, \dots, n$ .
- For now, we will assume the multivariate normal model as the sampling model, so that each  $p$  dimensional  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})^T \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$ .

$$p(\mathbf{Y}_i | \boldsymbol{\theta}, \Sigma) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{p/2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\theta}) \right\}$$

- Suppose now that  $\mathbf{Y}$  contains missing values.
- We can separate  $\mathbf{Y}$  into the observed and missing parts so that for for each individual,

$$\mathbf{Y}_i = (\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis})$$

# Mathematical Formulation

- Let
  - $j$  index variables (where  $i$  already indexes individuals),
  - $r_{ij} = 1$  when  $y_{ij}$  is missing,
  - $r_{ij} = 0$  when  $y_{ij}$  is observed.
- Here,  $r_{ij}$  is known as the missingness indicator of variable  $j$  for person  $i$ .
- Also, let
  - $\mathbf{R}_i = (r_{i1}, \dots, r_{ip})^T$  be the vector of missing indicators for person  $i$ .
  - $\mathbf{R} = (\mathbf{R}_1, \dots, \mathbf{R}_n)$  be the matrix of missing indicators for everyone.
  - $\boldsymbol{\psi}$  be the set of parameters associated with  $\mathbf{R}$ .
- Assume  $\boldsymbol{\psi}$  and  $(\boldsymbol{\theta}, \Sigma)$  are distinct.

# Mathematical Formulation

- MCAR:

$$p(\mathbf{R}|\mathbf{Y}, \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = p(\mathbf{R}|\boldsymbol{\psi})$$

- MAR:

$$p(\mathbf{R}|\mathbf{Y}, \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = p(\mathbf{R}|\mathbf{Y}_{obs}, \boldsymbol{\psi})$$

- MNAR:

$$p(\mathbf{R}|\mathbf{Y}, \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = p(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi})$$



# Implications for Likelihood Function

- Each type of mechanism has a different implication on the likelihood of the observed data  $\mathbf{Y}_{obs}$ , and the missing data indicator  $\mathbf{R}$ .
- Without missingness in  $\mathbf{Y}$ , the likelihood of the observed data is

$$p(\mathbf{Y}_{obs} | \boldsymbol{\theta}, \Sigma)$$

- With missingness in  $\mathbf{Y}$ , the likelihood of the observed data is instead

$$p(\mathbf{Y}_{obs}, \mathbf{R} | \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = \int p(\mathbf{R} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta}, \Sigma) d\mathbf{Y}_{mis}$$

- Since we do not actually observe  $\mathbf{Y}_{mis}$ , we would like to be able to integrate it out so we don't have to deal with it and infer  $(\boldsymbol{\theta}, \Sigma)$  using only the observed data.

# Likelihood function: MAR

- Focus on MAR

$$\begin{aligned} p(\mathbf{Y}_{obs}, \mathbf{R} | \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) &= \int p(\mathbf{R} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta}, \Sigma) d\mathbf{Y}_{mis} \\ &= \int p(\mathbf{R} | \mathbf{Y}_{obs}, \boldsymbol{\psi}) \cdot p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta}, \Sigma) d\mathbf{Y}_{mis} \\ &= p(\mathbf{R} | \mathbf{Y}_{obs}, \boldsymbol{\psi}) \cdot \int p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta}, \Sigma) d\mathbf{Y}_{mis} \\ &= p(\mathbf{R} | \mathbf{Y}_{obs}, \boldsymbol{\psi}) \cdot p(\mathbf{Y}_{obs} | \boldsymbol{\theta}, \Sigma). \end{aligned}$$

- For inference on  $(\boldsymbol{\theta}, \Sigma)$ , we only need  $p(\mathbf{Y}_{obs} | \boldsymbol{\theta}, \Sigma)$  in the likelihood function for inference  $(\boldsymbol{\theta}, \Sigma)$ .
- Still is hard, as we need marginal model!

# Bayesian Inference with Missing Data

- For posterior sampling for most models (especially multivariate models), sampling is easier with complete data  $\mathbf{Y}$ 's to update the parameters.
- Think of the missing data as **latent variables** and sample from the **posterior predictive distribution** of the missing data conditional on the observed data and parameters:

$$p(\mathbf{Y}_{mis} | \mathbf{Y}_{obs}, \boldsymbol{\theta}, \Sigma) \propto \prod_{i=1}^n p(\mathbf{Y}_{i,mis} | \mathbf{Y}_{i,obs}, \boldsymbol{\theta}, \Sigma).$$

- In the case of the multivariate normal model, each  $p(\mathbf{Y}_{i,mis} | \mathbf{Y}_{i,obs}, \boldsymbol{\theta}, \Sigma)$  is just a normal distribution, and we can leverage results on conditional distributions for normal models.

# Model for Missing Data

- Rewrite as  $\mathbf{Y}_i$  in block form

$$\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_{i,mis} \\ \mathbf{Y}_{i,obs} \end{pmatrix} \sim \mathcal{N}_p \left[ \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right],$$

- Missing data has a conditional

$$\mathbf{Y}_{i,mis} | \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs} \sim \mathcal{N} \left( \boldsymbol{\theta}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{y}_{i,obs} - \boldsymbol{\theta}_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right).$$

- multivariate normal distribution (or univariate normal distribution if  $\mathbf{Y}_i$  only has one missing entry)
- This sampling technique actually encodes MAR since the imputations for  $\mathbf{Y}_{mis}$  depend on the  $\mathbf{Y}_{obs}$ .

# Semi-Conjugate Prior

- We need prior distributions for  $\boldsymbol{\theta}$  and  $\Sigma$
- Multivariate Normal Prior for  $\boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \Lambda_0^{-1})$
- Analogous to the univariate case, the **inverse-Wishart distribution** is the corresponding conditionally conjugate prior for  $\Sigma$  (multivariate generalization of the inverse-gamma).
- A random variable  $\Sigma \sim \text{IW}_p(\eta_0, \mathbf{S}_0^{-1})$ , where  $\Sigma$  is positive definite and  $p \times p$ , has pdf

$$p(\Sigma) \propto |\Sigma|^{\frac{-(\eta_0+p+1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}_0 \Sigma^{-1}) \right\}$$

- $\eta_0 > p - 1$  is the “degrees of freedom”, and
- $\mathbf{S}_0$  is a  $p \times p$  positive definite matrix.

# Mean

- For this distribution,  $E[\Sigma] = \frac{1}{\eta_0 - p - 1} \mathbf{S}_0$ , for  $\eta_0 > p + 1$ .
- If we are very confident in a prior guess  $\Sigma_0$ , for  $\Sigma$ , then we might set
  - $\eta_0$ , the degrees of freedom to be very large, and
  - $\mathbf{S}_0 = (\eta_0 - p - 1)\Sigma_0$ .
  - $E[\Sigma] = \frac{1}{\eta_0 - p - 1} \mathbf{S}_0 = \frac{1}{\eta_0 - p - 1} (\eta_0 - p - 1)\Sigma_0 = \Sigma_0$ , and  $\Sigma$  is tightly (depending on the value of  $\eta_0$ ) centered around  $\Sigma_0$ .
- If we are not at all confident but we still have a prior guess  $\Sigma_0$ , we might set
  - $\eta_0 = p + 2$ , so that the  $E[\Sigma] = \frac{1}{\eta_0 - p - 1} \mathbf{S}_0$  is finite.
  - $\mathbf{S}_0 = \Sigma_0$

# Alternatives

- Jeffreys prior (improper limiting case)
- unit-information (data dependent)
- [Sun, D. and Berger, J.O \(2006\)](#) Objective Bayesian Analysis for the Multivariate Normal Model
- [Mulder, J. Pericchi, L.R. \(2018\)](#) The Matrix-F Prior for Estimating and Testing Covariance Matrices.

# Wishart distribution

- Just as we had with the gamma and inverse-gamma relationship in the univariate case, we can also work in terms of the **Wishart distribution** (multivariate generalization of the gamma) instead.
- The **Wishart distribution** provides a conditionally-conjugate prior for the precision matrix  $\Sigma^{-1}$  in a multivariate normal model.
- if  $\Sigma \sim IW_p(\eta_0, \mathbf{S}_0)$ , then  $\Phi = \Sigma^{-1} \sim W_p(\eta_0, \mathbf{S}_0^{-1})$ .
- A random variable  $\Phi \sim W_p(\eta_0, \mathbf{S}_0^{-1})$ , where  $\Phi$  has dimension  $(p \times p)$ , has pdf

$$f(\Phi) \propto |\Phi|^{\frac{\eta_0 - p - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}_0 \Phi) \right\}.$$

- Here,  $E[\Phi] = \eta_0 \mathbf{S}_0$ .



# Conditional posterior for $\Sigma$

$$\begin{aligned} Y_i | \boldsymbol{\theta}, \Sigma &\stackrel{ind}{\sim} N(\boldsymbol{\theta}, \Sigma) \\ \Sigma &\sim \text{IW}_p(\eta_0, \mathbf{S}_0^{-1}) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Psi}_0^{-1}) \end{aligned}$$

- The conditional posterior (full conditional)  $\Sigma | \boldsymbol{\theta}, \mathbf{Y}$ , is then

$$\Sigma | \boldsymbol{\theta}, \mathbf{Y} \sim \text{IW}_p \left( \eta_0 + n, \left( \mathbf{S}_0 + \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})^T \right)^{-1} \right)$$

- posterior sample size  $\eta_0 + n$
- posterior sum of squares  $\mathbf{S}_0 + \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})^T$

# Posterior Derivation

- The conditional posterior (full conditional)  $\Sigma \mid \boldsymbol{\theta}, \mathbf{Y}$ , is

$$\begin{aligned}\pi(\Sigma \mid \boldsymbol{\theta}, \mathbf{Y}) &\propto p(\Sigma) \cdot p(\mathbf{Y} \mid \boldsymbol{\theta}, \Sigma) \\ &\propto |\Sigma|^{\frac{-(\eta_0 + p + 1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}_0 \Sigma^{-1}) \right\} \cdot \prod_{i=1}^n |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(\mathbf{Y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\theta})] \right\}\end{aligned}$$

$$\Sigma \mid \boldsymbol{\theta}, \mathbf{Y} \sim \text{IW}_p \left( \eta_0 + n, \left( \mathbf{S}_0 + \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})^T \right)^{-1} \right)$$

# Gibbs sampler with missing data

At iteration  $s + 1$ , do the following

1. Sample  $\boldsymbol{\theta}^{(s+1)}$  from its multivariate normal full conditional
 
$$p(\boldsymbol{\theta}^{(s+1)} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(s)}, \boldsymbol{\Sigma}^{(s)})$$
2. Sample  $\boldsymbol{\Sigma}^{(s+1)}$  from its inverse-Wishart full conditional
 
$$p(\boldsymbol{\Sigma}^{(s+1)} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(s)}, \boldsymbol{\theta}^{(s+1)})$$
3. For each  $i = 1, \dots, n$ , with at least one “1” value in the missingness indicator vector  $\mathbf{R}_i$ , sample  $\mathbf{Y}_{i,mis}^{(s+1)}$  from the full conditional

$$\mathbf{Y}_{i,mis}^{(s+1)} | \mathbf{Y}_{i,obs}, \boldsymbol{\theta}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)} \sim \mathcal{N}(\boldsymbol{\theta}_1^{(s+1)} + \boldsymbol{\Sigma}_{12}^{(s+1)} \boldsymbol{\Sigma}_{22}^{(s+1)-1} (\mathbf{Y}_{i,obs} - \boldsymbol{\theta}_2^{(s+1)}), \boldsymbol{\Sigma}_{11}^{(s+1)} - \boldsymbol{\Sigma}_{12}^{(s+1)} \boldsymbol{\Sigma}_{22}^{(s+1)-1} \boldsymbol{\Sigma}_{21}^{(s+1)})$$

- derived from the original sampling model but with the updated parameters,
 
$$\mathbf{Y}_i^{(s+1)} = (\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis}^{(s+1)})^T \sim \mathcal{N}_p(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)}).$$

# Reading example from Hoff with missing data

```
      pretest posttest
[1,]      59      77
[2,]      43      39
[3,]      34      46
[4,]      32     NA
[5,]     NA      38
[6,]      38     NA
[7,]      55     NA
[8,]      67      86
[9,]      64      77
[10,]     45      60
[11,]     49      50
[12,]     72      59

      pretest posttest
0.1363636 0.2272727
```

# MCMC Summary for $\Sigma$

Iterations = 1:20000  
 Thinning interval = 1  
 Number of chains = 1  
 Sample size per chain = 20000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

|          | Mean  | SD    | Naive SE | Time-series SE |
|----------|-------|-------|----------|----------------|
| sigma_11 | 194.0 | 63.08 | 0.4460   | 0.4947         |
| sigma_12 | 152.1 | 60.75 | 0.4295   | 0.4665         |
| sigma_21 | 152.1 | 60.75 | 0.4295   | 0.4665         |
| sigma_22 | 248.7 | 82.70 | 0.5018   | 0.6084         |

# Compare to inference from full data

- With missing data:

|         | theta_1  | theta_2  |
|---------|----------|----------|
| Min.    | 30.45459 | 38.29322 |
| 1st Qu. | 43.65988 | 51.96991 |
| Median  | 45.60829 | 54.19592 |
| Mean    | 45.63192 | 54.20408 |
| 3rd Qu. | 47.61896 | 56.48918 |
| Max.    | 58.81206 | 70.49105 |

- Based on true data:

|         | theta_1  | theta_2  |
|---------|----------|----------|
| Min.    | 34.88365 | 37.80999 |
| 1st Qu. | 45.29473 | 51.47834 |
| Median  | 47.28229 | 53.65172 |
| Mean    | 47.26301 | 53.64100 |
| 3rd Qu. | 49.21423 | 55.81819 |
| Max.    | 60.94924 | 69.92354 |

- Very similar for the most part.

# Compare to inference from full data

- With missing data:

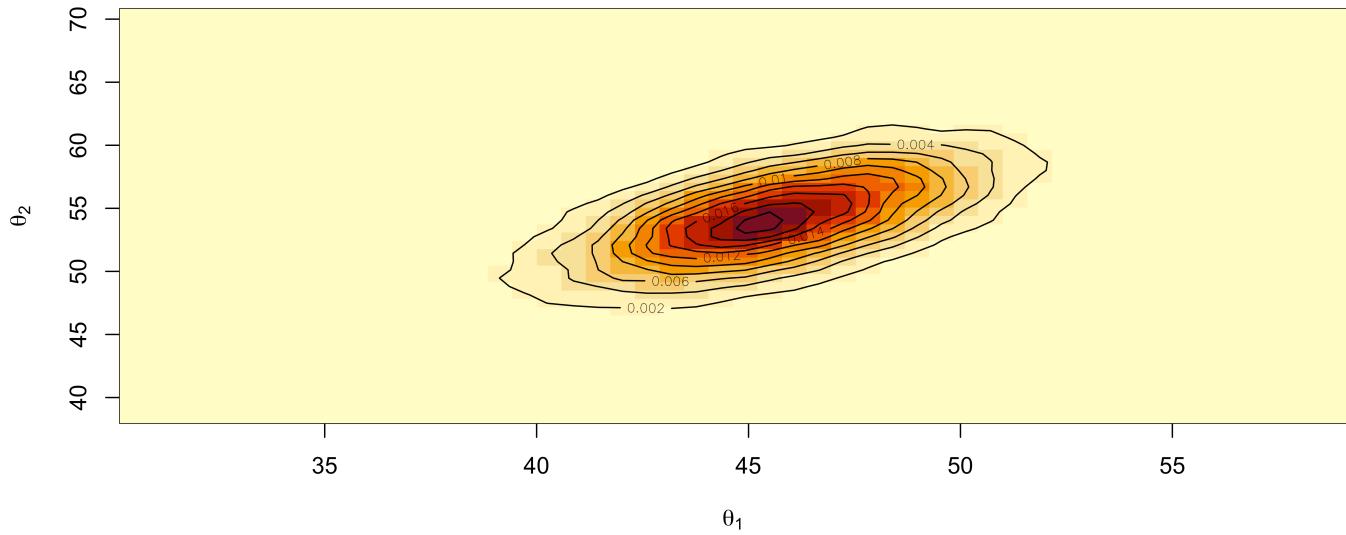
|         | sigma_11 | sigma_12  | sigma_21  | sigma_22  |
|---------|----------|-----------|-----------|-----------|
| Min.    | 64.0883  | -20.39204 | -20.39204 | 82.55346  |
| 1st Qu. | 149.8338 | 109.84218 | 109.84218 | 190.25962 |
| Median  | 182.4496 | 142.34686 | 142.34686 | 233.43312 |
| Mean    | 193.9803 | 152.12898 | 152.12898 | 248.67527 |
| 3rd Qu. | 224.0994 | 182.75082 | 182.75082 | 289.47663 |
| Max.    | 734.8704 | 668.77332 | 668.77332 | 981.99916 |

- Based on true data:

|         | sigma_11 | sigma_12  | sigma_21  | sigma_22  |
|---------|----------|-----------|-----------|-----------|
| Min.    | 76.4661  | -38.75561 | -38.75561 | 93.65776  |
| 1st Qu. | 157.5870 | 113.32529 | 113.32529 | 203.69192 |
| Median  | 190.6578 | 145.08962 | 145.08962 | 246.08696 |
| Mean    | 201.9547 | 155.20374 | 155.20374 | 260.11361 |
| 3rd Qu. | 233.5809 | 186.36991 | 186.36991 | 300.70840 |
| Max.    | 664.8241 | 577.99100 | 577.99100 | 947.39333 |

- Also very similar. A bit more uncertainty in dimension of  $Y_{i2}$  because we have more missing data there

# Posterior distribution of the mean





# Missing data vs predictions for new observations

- How about predictions for completely new observations?
- That is, suppose your original dataset plus sampling model is  $\mathbf{y}_i = (y_{i,1}, y_{i,2})^T \sim \mathcal{N}_2(\boldsymbol{\theta}, \Sigma), i = 1, \dots, n$ .
- Suppose now you have  $n^*$  new observations with  $y_2^*$  values but no  $y_1^*$ .
- How can we predict  $y_{i,1}^*$  given  $y_{i,2}^*$ , for  $i = 1, \dots, n^*$ ?
- Well, we can view this as a “train  $\rightarrow$  test” prediction problem rather than a missing data problem on an original data.

# Missing data vs predictions for new observations

- That is, given the posterior samples of the parameters, and the test values for  $\mathbf{y}_{i,2}^*$ , draw from the posterior predictive distribution of  $(\mathbf{y}_{i,1}^* | \mathbf{y}_{i,2}^*, \{(\mathbf{y}_{1,1}, \mathbf{y}_{1,2}), \dots, (\mathbf{y}_{n,1}, \mathbf{y}_{n,2})\})$ .
- To sample from this predictive distribution, think of compositional sampling.
- for each posterior sample of  $(\boldsymbol{\theta}, \Sigma)$ , sample from  $(\mathbf{y}_{i,1} | \mathbf{y}_{i,2}, \boldsymbol{\theta}, \Sigma)$ , which is just from the form of the sampling distribution.
- In this case,  $(\mathbf{y}_{i,1} | \mathbf{y}_{i,2}, \boldsymbol{\theta}, \Sigma)$  is just a normal distribution derived from  $(\mathbf{y}_{i,1}, \mathbf{y}_{i,2} | \boldsymbol{\theta}, \Sigma)$ , based on the conditional normal formula.
- No need to incorporate the prediction problem into your original Gibbs sampler!

# MNAR Likelihood function:

- For MNAR, we have:

$$p(\mathbf{Y}_{obs}, \mathbf{R} | \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = \int p(\mathbf{R} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta}, \Sigma) d\mathbf{Y}_{mis}$$

- The likelihood under MNAR cannot simplify any further.
- In this case, we cannot ignore the missing data when making inferences about  $(\boldsymbol{\theta}, \Sigma)$ .
- We must include the model for  $\mathbf{R}$  and also infer the missing data  $\mathbf{Y}_{mis}$ .
- So how can we tell the type of mechanism we are dealing with?
- In general, we don't know!!!
- Rare that data are MCAR (unless planned beforehand); more likely that data are MNAR or MNAR.