Lecture 13. Missing Data and Ignorability

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

§1 数据缺失模式

1. A Simple Sketch

1. Complete-data model

- **Complete-data model:**
  - $Y_1, \cdots, Y_n$ are *i.i.d* draws from multivariate distribution $P_\theta \in \mathcal{P} = \{P_\eta : \eta \in \Theta\}$, i.e.,

$$Y_i = (Y_{i1}, \cdots, Y_{ip})^T \sim P_\theta.$$

  - Data Pattern: E.g., when sample size $n = 3$ and dimension $p = 5$, the complete data matrix can be written as the following way,

$$\begin{array}{cccc} & Y_1 & Y_2 & Y_3 \\ 1 & Y_{11} & Y_{21} & Y_{31} \\ 2 & Y_{12} & Y_{22} & Y_{32} \\ 3 & Y_{13} & Y_{23} & Y_{33} \\ 4 & Y_{14} & Y_{24} & Y_{34} \\ 5 & Y_{15} & Y_{25} & Y_{35} \end{array}$$

  - Consider the inference problem of estimating θ, two common ways in the parametric inference setting would be
    (i) Likelihood approach $\rightsquigarrow$ maximum likelihood estimator, MLE.
    (ii) Bayesian approach $\rightsquigarrow$ Bayes estimator.

    Here, say we have the distribution function P_θ together with a loss function $L(\theta, \delta)$ and a prior $\Lambda(\theta)$, then the Bayes estimator δ_Λ of θ with respect to prior $\Lambda(\theta)$ is defined as

$$\delta_\Lambda \triangleq \arg\min_\delta r(\Lambda, \delta)$$

where $r(\Lambda, \delta)$ is called the Bayes risk function of a estimator (statistic) δ,

$$r(\Lambda, \delta) \triangleq \int R(\theta, \delta) d\Lambda(\theta) = \int \left(\int L(\theta, \delta(x)) dP_\theta(x) \right) d\Lambda(\theta)$$

– Since this is an random sample, i.e., independent and identical observations, so the likelihood function of the complete data is given by

$$f_Y(y|\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i|\theta) \, ,$$

where $f_{Y_i}(\cdot|\theta)$ is the density of P_θ.

2. Missing-data model

- **Missing-data model:**

– Problem arised in Practice, some data Y_{ij} might be missing. For example, we collect some data of the math grade of some middle school students. Say the data is

(i) Math grade was collect for 14 students together with their gender if they filled in the choice item in the survey.

(ii) For each student, the math grade is collected for each year from year 1 to year 6.

(iii) For 5 of the 14 students, who studied for the 7th year, an additional 7th year math grade is collected for them.

Id	Gender	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
1	F	85	89	86	87	90	98	·
2	M	45	55	53	60	58	62	67
3	F	85	88	89	78	76	80	77
4	F	80	75	88	83	85	87	·
5	F	95	97	93	92	91	97	·
6	·	86	84	93	92	86	89	·
7	F	85	86	82	90	87	89	·
8	M	54	49	53	47	44	48	56
9	M	70	74	75	72	67	64	71
10	F	85	94	92	86	89	89	·
11	F	95	87	93	89	86	80	·
12	·	65	54	63	58	61	59	63
13	·	85	94	82	84	91	87	·
14	M	95	100	93	97	99	99	·

– In order to dealing with the missing data, we create certain format to describe missingness of Y_{ij} by introducing the indicator variable R_{ij}, s.t.,

$$R_{ij} = \begin{cases} 1 & Y_{ij} \text{ has been observed} \\ 0 & Y_{ij} \text{ is missing} \end{cases}$$

- Hence the data pattern for multivariate dataset with missing valuses is given by

	Y_1	Y_2	Y_3	R_1	R_2	R_3
1	Y_{11}	Y_{21}	Y_{31}	1	1	1
2	–	Y_{22}	Y_{32}	0	1	1
3	Y_{13}	–	Y_{33}	1	0	1
4	Y_{14}	Y_{24}	–	1	1	0
5	Y_{15}	–	Y_{35}	1	0	1

- And the above data consisted our Observed-data model, which contain two parts, one, the observed values $Y_{ij} = y_{ij}$, and two, the values r_{ij} to indicating which values are missing. We adopt the following commonly used notations. The complete data under our experiments is $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} are the observed variables and Y_{mis} are the missing variables.

3. 缺失模式

- **Missing data mechanism:** The statistical model for missing data is defined by

$$\mathcal{F}_0 = \left\{P_\xi(R = r|Y = y) = f_{R|Y}(r|y, \xi) : \xi \in \Xi\right\}.$$

and further, by separating $\mathcal{F}_0$ into subsets, we can define three classical missing data mechanism,

Definition 1.1 (♣ **Missing completely at random, MCAR**). The observation are missing completely at random if for $\forall \xi \in \Xi$

$$P_\xi(R = r|Y = y) = P_\xi(R = r) \quad \text{or equivalently} \quad f_{R|Y}(r|y, \xi) = f_R(r|\xi)\ .$$

that is, R and Y are independent.

Definition 1.2 (♣ **Missing at random, MAR**). The observation are missing at random if for $\forall \xi \in \Xi$

$$P_\xi(R = r|Y = y) = P_\xi(R = r|Y_{obs}) \quad \text{or equivalently}$$
$$f_{R|Y}(r|y, \xi) = f_{R|Y_{obs}}(r|y_{obs}, \xi)\ .$$

that is, knowledge about Y_{mis} does not provide any additional information about R if Y_{obs} is already known. In other words, the value of the observation which is missing does not effect why it's missing when we have the observed data. (R and Y_{mis} are conditionally independent given Y_{obs}).

Definition 1.3 (♣ **Missing not at random, MNAR**). The observation are missing not at random if the above assumption for MAR is not fulfilled.

2. Several Examples

We can try to understand different type of missing data mechanism by looking at following exammples.

- *Example* 2.1. **Bernoulli selection**. Say our complete data are $Y_1, \cdots, Y_n$ and the corresponding response indicators are $R_1, \cdots, R_n$. Now, if Y_i is missing with probability ξ, then this is a MCAR case.

Answer. Notice that since Y_i is missing with probability ξ, in other words, $R_i \sim_{i.i.d} Bernoulli(\xi)$, so

$$\boxed{f_{R|Y}(r|y,\xi) = f_R(r|\xi) = \prod_{i=1}^{n} \xi^{r_i}(1-\xi)^{1-r_i} \cdot I_{\{r_i \in \{0,1\}\}}}$$

hence this is MCAR case. □

- *Example* 2.2. **Simple random sample**. Suppose we can only afford to take m observations in a size n sample, such like, testing gene expression of a person. And we are choosing those m genes equally out of total n genes, then it's a MCAR case.

Answer. Based on the content, we have

$$\boxed{f_{R|Y}(r|y,\xi) = f_R(r|\xi) = \binom{n}{m}^{-1} \cdot I_{\{\sum_{i=1}^{n} r_i = m\}}}$$

so we have this is the MCAR case, note that $\xi = m$ is a parameter of the distribution of R. □

- *Example* 2.3. **Double sampling**. Suppose we testing blood pressure for n patients in the hospital. If one has high blood pressure (say blood pressure great than 169/99), then we would further do a cholesterol level test, then it's a MAR case.

Answer. On one hand, notice that,

$$f_{R_{i1}|Y}(1|y,\xi) = f_R(1|\xi) = 1\ , \quad \text{for } \forall\, i \in \{1, \cdots, n\} \quad \text{(对于 blood pressure)}$$

in other words, we will always test for i-th patient's blood pressure. Therefore, ① $\boxed{\{Y_{i1}\}_{1\le i\le n} \subseteq Y_{obs}}$, on the other hand,

$$② \boxed{f_{R_{i2}|Y}(1|y,\xi) = I_{\{y_{i1}>c\}} = f_{R_{i2}|Y_{obs}}(1|y_{obs},\xi)} \quad \text{(对于 cholesterol level)}$$

so overall, we have $R|Y \sim R|Y_{obs}$, hence this is a MAR case and not MCAR actually. □

- *Example* 2.4. **sampling with nonresponse followup.** Suppose we are conducting MRI for n patients, and for those patients who couldn't do the MRI (such like the patient has a tattoo), we conduct CT instead. If we denote our data for i-th patient as $Y_i = (Y_{i1}, Y_{i2})$=(MRI result, CT result) and assume that the patient couldn't do MRI with probability p_i for some personal reasons, then it's a MAR case.

Answer. Notice that

$$\boxed{f_{R_{i1}|Y}(0|y,\xi) = f_{R_{i1}}(0|\xi) = p_i}$$

in other words, we have i-th patient's MRI result is MCAR. Now after we have finish the MRI test, we have

$$\boxed{f_{R_{i2}|Y}(1|y,\xi) = 1 - I_{\{y_{i1} \in y_{obs}\}} = f_{R_{i2}|Y_{obs}}(1|y_{obs},\xi)}$$

so overall, we have $R|Y \sim R|Y_{obs}$, hence this is a MAR case and not MCAR actually. □

- *Example* 2.5. Consider n individuals are treated for high blood pressure, s.t.

$$Y_{ij} = \text{blood pressure of } i-th \text{ individual on } j-th \text{ day}$$

and the response indicator is

$$R_{ij} = \begin{cases} 1 & i-th \text{ individual appears for measurement on } j-th \text{ day.} \\ 0 & \text{otherwise} \end{cases}$$

suppose our statistical model for Y and R is given by

$$Y_{ij} \sim_{i.i.d} N(\mu_i, \sigma^2), \quad \text{and} \quad R_{ij} \sim_{i.i.d} Bernoulli\left(1 - \frac{\exp(Y_{ij})}{1+\exp(Y_{ij})}\right)$$

then it's a MNAR case.

Answer. Notice that the missing mechanism R_{ij} depend on the value of Y_{ij} for each of them could be a missing value, so this is MNAR case. As a matter of fact, notice that the probability of R_{ij} being missing is large if Y_{ij} is large, i.e., individuals with high blood pressure are less likely to turn up for a measurement. □

- *Example* 2.6. **Randomly Censored Data.** Consider n components life time $T_1, \cdots, T_n \sim_{i.i.d} f(\cdot|\theta)$, with survivor function $S(t) = P(T_1 > t)$. Now, if we denote the censoring times associated with $T_i's$ are

$$C_1, \cdots, C_n \sim_{i.i.d} g(\cdot|\xi) \quad \text{and} \quad Y_i = min(T_i, C_i), \quad R_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{otherwise} \end{cases}$$

then it's a MNAR case.

Answer. Notice that

$$f_{R_i,C_i|Y}(1, c_i|y, \xi) = I_{\{y_i \leq c_i\}} \cdot g(c_i|\xi)$$

i.e.,

$$\boxed{f_{R_i|Y}(1|y, \xi) = \int I_{\{y_i \leq c_i\}} \cdot g(c_i|\xi) dc_i}$$

so the response indicator R_i depend on the value of Y_i while Y_i could be a missing value, so this is MNAR case. □

3. Primarily Study of Missing Data

3.1. Observed-Data Likelihood

Based on the notations introduced eariler, $Y = (Y_{obs}, Y_{mis})$ and the response indicator R, we have the likelihood of complete observations (Y_{obs}, R) is given by

$$\begin{aligned} L(\theta, \xi|Y_{obs}, R) &= f_{Y_{obs},R}(Y_{obs}, R|\theta, \xi) \\ &= \int f_{Y,R}(Y_{obs}, y_{mis}, R|\theta, \xi) dy_{mis} \\ &= \int f_{R|Y}(R|Y_{obs}, y_{mis}, \xi) f_Y(Y_{obs}, y_{mis}|\theta) dy_{mis} \end{aligned}$$

where $\theta \in \Theta$ is the parameter invovled in modeling the complete data Y, and $\xi \in \Xi$ is the parameter invovled in describing the missing mechanism, i.e., in modeling the response indicator R. One interesting fact is that, if the missing observations are missing at random, MAR, then the first factor $f_{R|Y}(R|Y_{obs}, y_{mis}, \xi) = f_{R|Y_{obs}}(R|Y_{obs}, \xi)$ does not depend on y_{mis} and thus can be take out of the integral with respect to y_{mis}, therefore, for independent observation Y, we have

$$\begin{aligned} L(\theta, \xi|Y_{obs}, R) &= \int f_{R|Y}(R|Y_{obs}, y_{mis}, \xi) f_Y(Y_{obs}, y_{mis}|\theta) dy_{mis} \\ &= f_{R|Y_{obs}}(R|Y_{obs}, \xi) \int f_Y(Y_{obs}, y_{mis}|\theta) dy_{mis} \\ &= f_{R|Y_{obs}}(R|Y_{obs}, \xi) L(\theta|Y_{obs}) \end{aligned}$$

If the parameter ξ does not depend on θ, which is of our interest, then we can conduct inference for θ in the way that

$$\begin{aligned} \hat{\theta}_{MLE}(Y_{obs}, R) &= \arg\max_{\theta \in \Theta} L(\theta, \xi|Y_{obs}, R) \\ &= \arg\max_{\theta \in \Theta} f_{R|Y_{obs}}(R|Y_{obs}, \xi) L(\theta|Y_{obs}) \\ &= \arg\max_{\theta \in \Theta} L(\theta|Y_{obs}) = \hat{\theta}_{MLE}(Y_{obs}) \end{aligned}$$

which means, in this case, the MLE of θ based on the (Y_{obs}, R) is the same as the MLE based only on the observed data Y_{obs}. This is a desirable property in practice as a matter of fact, in classical regression as well as most other models, R automatically excludes all cases in which any of the inputs are missing, (This can limit the amount of information available in the analysis, especially if the model includes many inputs with potential missingness), this approach is called a complete-case analysis.

3.2. Ignorability

Definition 3.1 (♣ **Observed-data likelihood**). The likelihood function $L(\theta|Y_{obs}) = f_{Y_{obs}}(Y_{obs}|\theta)$ is called the "likelihood ignoring the missing data mechanism", or sometimes for short, "observed-data likelihood".

Definition 3.2 (♣ **Distinct**). We say the parameter ξ in the missing mechanism and θ in the data model are distinct if they don't depend on each other, in other words, ξ and θ are distinct in the sense that the joint parameter space of (θ, ξ) is the product of the parameter spaces Θ and Ξ.

Definition 3.3 (♣ **Ignorability**). A missing data mechanism is "ignorable for likelihood inference" if • the observations are MAR, and • the parameter ξ and θ are distinct.

• *Example* 3.4. **Non-distinct parameters.** Suppose that $Y_i \sim_{i.i.d} Bernoulli(\theta)$ and $R_i \sim_{i.i.d} Bernoulli(\theta)$, and we have $Y_{obs} = (Y_1, \cdots, Y_m)^T$ after some proper reordering. Please find the MLE for θ based on (Y_{obs}, R) and find the MLE of θ based on complete-case analysis.

Answer: First we try to find the MLE for θ based on (Y_{obs}, R). Notice the joint likelihood of (Y_{obs}, R) is

$$\begin{aligned} L(\theta|Y_{obs} = y_{obs}, R = r) &= \prod_{i=1}^{n} \theta^{r_i}(1-\theta)^{1-r_i}\left(\theta^{y_i}(1-\theta)^{1-y_i}\right)^{r_i} \\ &= \theta^{m+\sum_{i=1}^{m} y_i}(1-\theta)^{n-m+m-\sum_{i=1}^{m} y_i}. \end{aligned}$$

Notice that $\sum_{i=1}^{n} r_i = m$, hence that

$$\ell(\theta|Y_{obs} = y_{obs}, R = r) = \left(m + \sum_{i=1}^{m} y_i\right)\log\theta + \left(n - \sum_{i=1}^{m} y_i\right)\log(1-\theta).$$

Thus, by letting

$$\begin{aligned} \frac{\partial \ell(\theta|Y_{obs} = y_{obs}, R = r)}{\partial\theta} = 0 \quad \Rightarrow \\ \hat{\theta}_{MLE}(Y_{obs}, R) = \frac{m + \sum_{i=1}^{m} Y_i}{m+n} = \frac{\sum_{i=1}^{n} R_i + \sum_{i=1}^{m} R_i Y_i}{\sum_{i=1}^{m} R_i + n}. \end{aligned}$$

by checking that $\hat{\theta}_{MLE}$ is indeed the globle maximum of the likelihood function, we have it is the MLE for θ based on (Y_{obs}, R).

Second, for the MLE of θ based on complete-case analysis, notice that the observed-data likelihood is

$$L(\theta, Y_{obs}) = \prod_{i=1}^{m} \theta^{y_i}(1-\theta)^{1-y_i}$$

by letting $\partial\ell(\theta, y_{obs})/\partial\theta = 0$, and checking the solution is indeed the globle maximum of the likelihood, we find the MLE

$$\hat{\theta}_{MLE}(Y_{obs}) = \frac{\sum_{i=1}^{m} Y_i}{m}$$

which still is a valid estimator (due to the fact that this is a MCAR case) but with a lower efficiency.

And as can be seen in the above example, for the univariate case of missing data, such as our observations is given by

	1	2	$\cdots$	m	$m+1$	$\cdots$	n
Y	Y_1	Y_2	$\cdots$	Y_m	$-$	$\cdots$	$-$
R	1	1	$\cdots$	1	0	$\cdots$	0

suppose that $Y_1, \cdots, Y_n$ is an random sample, only $Y_{obs} = (Y_1, \cdots, Y_m)^T$ have been observed and $Y_{mis} = (Y_{m+1}, \cdots, Y_n)^T$ are MAR, then the observed-data likelihood is

$$\begin{aligned} L(\theta|Y_{obs}) = f_{Y_{obs}}(Y_{obs}|\theta) &= \int f_Y(Y_{obs}, y_{mis}|\theta)dy_{mis} \\ &= \int \cdots \int \prod_{i=1}^{m} f_{Y_i}(Y_i|\theta) \prod_{i=m+1}^{n} f_{Y_i}(y_i|\theta)dy_{m+1}\cdots dy_n \\ &= \prod_{i=1}^{m} f_{Y_i}(Y_i|\theta) \int \cdots \int \prod_{i=m+1}^{n} f_{Y_i}(y_i|\theta)dy_{m+1}\cdots dy_n \\ &= \prod_{i=1}^{m} f_{Y_i}(Y_i|\theta) \end{aligned}$$

so the observed-data likelihood is a complete-case data likelihood, i.e., it's the likelihood based on the reduced sample $(Y_1, \cdots, Y_m)$.

- **Some features about ignorability**
  - The first condition in defining ignorability (MAR) is typically regarded as the more important condition.

- If MAR does not hold, then the MLE based on the observed-data likelihood can be seriously biased.
- If the data are MAR but distinctness does not hold, inference based on the observed-data likelihood $L(\theta|Y_{obs})$ is still valid from the frequentist perspective, but usually is relatively less efficiency compare to the MLE induced from the likelihood based on (Y_{obs}, R).

4. Simulation study of a Bivariate normal missing data model

• *Example* 4.1. Say we consider the bivariate data with one variable subject to nonresponse, we assume the first m observations (i.e, $Y_1, \cdots, Y_m$) are complete case and the late $n-m$ observations are missing the second coordinate, so the data pattern may looks like

	1	2	$\cdots$	m	$m+1$	$\cdots$	n
Y	Y_{11}	Y_{21}	$\cdots$	Y_{m1}	$Y_{(m+1)1}$	$\cdots$	Y_{n1}
	Y_{12}	Y_{22}		Y_{m2}	$-$		$-$
R	1	1	$\cdots$	1	1	$\cdots$	1
	1	1		1	0		0

Suppose $Y_i = (Y_{i1}, Y_{i2}) \sim_{i.i.d} N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

for some $\rho \in (0,1)$, and for simplicity, assume $\sigma_1 = \sigma_2 = 1$. Notice that

$$Y_{i2}|Y_{i1} \sim N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(Y_{i1} - \mu_1), (1-\rho^2)\sigma_2^2\right)$$

so we have the observed-data likelihood is given by

$$\begin{aligned} L(\theta|Y_{obs}) &= \prod_{i=1}^{m} f_{Y_i}(Y_i|\mu,\Sigma) \prod_{i=m+1}^{n} f_{Y_{i1}}(Y_{i1}|\mu_1,\sigma_1^2) \\ &= \prod_{i=1}^{m} f_{Y_{i2}|Y_{i1}}(Y_{i2}|\mu,\Sigma) \prod_{i=1}^{n} f_{Y_{i1}|\mu_1,\sigma_1^2} \\ &= C \cdot \exp\left(-\frac{m}{2}\log(1-\rho^2) - \frac{\sum_{\substack{i=1,\cdots,m \\ j=1,2}} (Y_{ij} - \mu_j)^2}{2(1-\rho^2)}\right) \end{aligned}$$

$$\cdot \exp\left(-\frac{\rho \sum_{i=1}^{m} (Y_{i1}-\mu_1)(Y_{i2}-\mu_2)}{(1-\rho^2)} - \frac{\sum_{i=m+1}^{n} (Y_{i1}-\mu_1)^2}{2} \right)$$

By letting $\bigtriangledown \ell(\theta|Y_{obs}) = 0$, we have

$$\begin{cases} \dfrac{\partial \ell}{\partial \mu_1} = \dfrac{\sum_{i=1}^{m}(Y_{i1}-\mu_1)}{1-\rho^2} + \dfrac{\rho\sum_{i=1}^{m}(Y_{i2}-\mu_2)}{1-\rho^2} + \displaystyle\sum_{i=m+1}^{n}(Y_{i1}-\mu_1) = 0 \\ \dfrac{\partial \ell}{\partial \mu_2} = \dfrac{\sum_{i=1}^{m}(Y_{i2}-\mu_2)}{1-\rho^2} + \dfrac{\rho\sum_{i=1}^{m}(Y_{i1}-\mu_1)}{1-\rho^2} = 0 \\ \dfrac{\partial \ell}{\partial \rho} = \dfrac{m\rho}{1-\rho^2} - \dfrac{\rho\sum_{\substack{i=1,\cdots,m \\ j=1,2}}(Y_{ij}-\mu_j)^2}{(1-\rho^2)^2} - \dfrac{1+\rho^2}{(1-\rho^2)^2}\displaystyle\sum_{i=1}^{m}(Y_{i1}-\mu_1)(Y_{i2}-\mu_2) = 0 \end{cases}$$

and get

$$\hat{\mu}_1 = \bar{Y}_1 \ , \quad \hat{\mu}_2 = \widetilde{Y}_2 + \hat{\rho}(\widetilde{Y}_1 - \bar{Y}_1) \ ,$$

and $\hat{\rho}$ is the solution to the following equation of order 3,

$$m\rho(1-\rho^2) - \rho(S_{1m}^2 + S_{2m}^2) - (1+\rho^2)S_{12} = 0$$

where we define

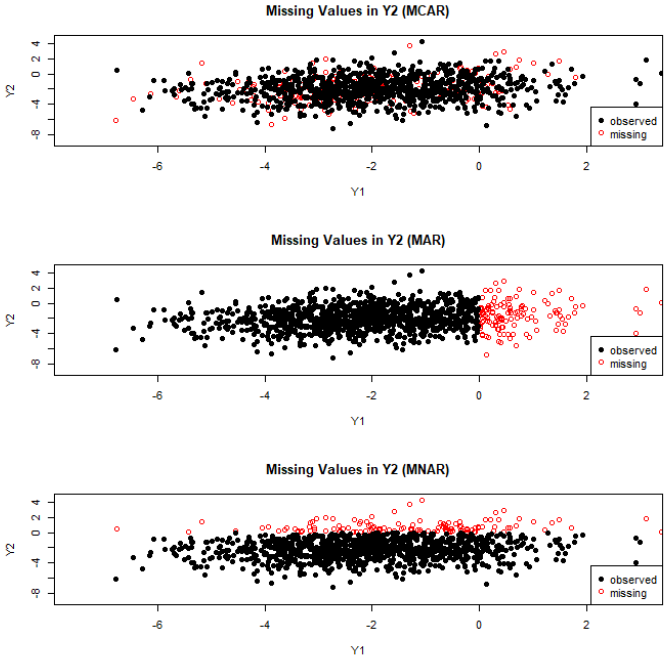

Figure 1: Bivariate Normal Data with Different Missing Mechanism

$$\bar{Y}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_{i1}, \quad \widetilde{Y}_1 = \frac{1}{m}\sum_{i=1}^{m} Y_{i1}, \quad \widetilde{Y}_2 = \frac{1}{m}\sum_{i=1}^{m} Y_{i2}, \quad S_{1m}^2 = \sum_{i=1}^{m}(Y_{i1} - \widetilde{Y}_1)^2$$

$$S_{2m}^2 = \sum_{i=1}^{m}(Y_{i2} - \widetilde{Y}_2)^2, \quad S_{12m}\sum_{i=1}^{m}(Y_{i1} - \widetilde{Y}_1)(Y_{i2} - \widetilde{Y}_2)$$

and based on the MLE for exponential family, we know that the above $(\hat{\mu}_1, \hat{\mu}_2, \hat{\rho})^T$ is the MLE for θ based on the observed-data likelihood. For more about this estimator, one may refer to Dahiya and Korwar (1980).

Further, we can put different missing mechanism, such as

- Missing completely at random (MCAR): if we put
$$R_{i2} \sim_{i.i.d} Bernoulli(p_0)$$
- Missing at random (MAR): if we put
$$R_{i2} = I_{\{Y_{i1} \le c_1\}}$$
- Missing not at random (MNAR): if we put
$$R_{i2} = I_{\{Y_{i2} \le c_2\}}$$

Here the picture is a small simulation for bivariate normal data under different missing mechanism described above, with

$$\mu = \begin{pmatrix} -3 \\ -3 \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} 2.5 & 0.5 \\ 0.5 & 2.5 \end{pmatrix}, \quad p_0 = 0.8\,, \quad \text{and} \quad c_1 = c_2 = 0\,.$$

5. Missing data in Bayesian Setting

Let's look at the last question in Assignment 5 but with certain difference in the setting. Let $X_1, \cdots, X_n \sim_{i.i.d} \theta Exp(1)$ with θ has the prior $\Lambda(\theta) = \xi \exp(-\xi\theta) \cdot I_{\{\theta>0\}}$. But some of the X_i are missing and they are recorded as 0s if they are missing. Let

$$Y_i = \begin{cases} X_i & \text{if } R_i = 1; \\ 0 & \text{if } R_i = 0. \end{cases} \quad \text{where} \quad R_i = \begin{cases} 1 & \text{if } X_i \text{ is observed;} \\ 0 & \text{if } X_i \text{ is missing.} \end{cases}$$

for $i = 1, \cdots, n$. Further let $\{(X_i, R_i)^T\}_{1\le i\le n}$ are independent pairs of random variables such that for each i, $R_i|X_i \sim Bernoulli(\pi_i)$, where $\pi_i = \pi_i(X_i) \in [0, 1]$ is defined under one of the following missing mechanism:

- Missing completely at random (MCAR): if we put
$$\pi_i = p$$
with some known constant p. Then we have
$$f_{R_i|X_i}(r_i|X_i) = p^{r_i}(1-p)^{1-r_i} \cdot I_{\{r_i\in\{0,1\}\}}$$

as the right hand side of the equation does not depend on X, so we have $f_{R_i|X_i}(r_i|X_i) = f_{R_i}(r_i)$, hence that this is MCAR case.

- Missing at random (MAR): if we put

$$\pi_i = I_{\{X_i \geq 1\}} \triangleq C_i$$

as $P(X_i = 0) = 0$ since $X_i \sim \theta Exp(1)$, and $R_i \sim Bernoulli(I_{\{X_i \geq 1\}})$, so

$$\begin{aligned} P(Y_i = 0) &= P(X_i = 0, R_i = 1) + P(R_i = 0) = P(R_i = 0) \\ &= P(Bernoulli(I_{\{X_i \geq 1\}}) = 0) = P(I_{\{X_i \geq 1\}} = 0) = P(X_i < 1) \end{aligned}$$

and notice also that when $X_i \geq 1$, $R_i = 1, a.s.$, hence $Y_i = X_i$ in that case. So overall, we have

$$Y_i = X_i \cdot I_{\{X_i \geq 1\}},\ a.s., \quad \Rightarrow \quad I_{\{X_i \geq 1\}} = I_{\{Y_i \geq 1\}},\ a.s.$$

so, by looking at the missing model, we have

$$f_{R_i|X_i}(r_i|X_i) = (I_{\{Y_i \geq 1\}})^{r_i}(1 - I_{\{Y_i \geq 1\}})^{1-r_i} \cdot I_{\{r_i \in \{0,1\}\}}$$

as Y is our observed data, i.e., Y_{obs}, so $f_{R_i|X_i}(r_i|X_i) = f_{R_i|Y_i}(r_i)$, hence that this is MAR case (But not MCAR case).

- Missing not at random (MNAR): if we put

$$\pi_i = p_1 I_{\{X_i > 1\}} + p_2 I_{\{X_i \leq 1\}} = p_1 C_i + p_2(1 - C_i)$$

with some known constant p_1 and p_2, $0 < p_1 < p_1 + p_2 < 1$ and $p_1 \neq p_2$. For the case X_i is missing, so $Y_i = 0, R_i = 0$, and we have

$$f_{R_i|X_i}(0|X_i) = 1 - p_1 I_{\{X_i \geq 1\}} - p_2 I_{\{X_i < 1\}}$$

depends on the value of X_i, so we have this is MNAR case.

Notice that in each case, by denote $R = (R_1, \cdots, R_n)^T$ and $Y_{obs} = (Y_1, \cdots, Y_n)^T$, we have the complete likelihood is given by

$$L(\theta|R, Y_{obs}) = \prod_{i=1}^{n} \left(\pi_i \theta^{-1} \exp\left(-Y_i/\theta\right)\right)^{R_i} (1 - \pi_i)^{1-R_i} \cdot I_{\{\theta > 0\}}$$

so the posterior is given by

$$\Lambda(\theta|R, Y_{obs}) \propto L(\theta|R_i, Y_i)\Lambda(\theta) \propto \theta^{-\sum_{i=1}^{n} R_i} \exp\left(-\xi\theta - \frac{\sum_{i=1}^{n} R_i Y_i}{\theta}\right) \cdot I_{\{\theta > 0\}}$$

and we can use the posterior mean, i.e.,

$$\hat{\theta}_\Lambda \triangleq E_\Lambda\left[\theta|R, Y_{obs}\right] = \int \theta \cdot \Lambda(\theta|R, Y_{obs}) d\theta$$

as a valid estimator, as a matter of fact, the posterior mean is directly the Bayes estimator corresponding to the quadratic loss function $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$.

6. Data with Ignorability, Multiple Imputation and the MICE Algorithms

In this section, we consider only the case where we have the data with ignorability, i.e., the missing mechanism being the missing at random and the parameter involved in the missing mechanism and the likelihood can be seperated. The observation is given by (Y_{obs}, R) while the missing data $Y_{mis} = (Y_{mis,1}, \cdots, Y_{mis,d})$, i.e., we have exactly d missing terms.

6.1. *Parametric Multiple Imputation by Chained Equations*

Assume the complete data likelihood and the missing mechanism are $L(\theta|Y_{obs}, Y_{mis})$ and $f(R|Y, \xi)$ respectively. Since we consider only the MAR case, so $f(R|Y,\xi) = f(R|Y_{obs}, \xi)$. Further

$$
\begin{aligned}
f(Y_{mis}|Y_{obs}, R, \theta, \xi) &= \frac{f(Y_{obs}, Y_{mis}, R|\theta, \xi)}{f(Y_{obs}, R|\theta, \xi)} \\
&= \frac{f(R|Y_{obs}, Y_{mis}, \theta, \xi) \cdot f(Y_{obs}, Y_{mis}|\theta, \xi)}{f(R|Y_{obs}, \theta, \xi) f(Y_{obs}|\theta, \xi)} \\
&= \frac{f(R|Y_{obs}, \xi) \cdot f(Y_{obs}, Y_{mis}|\theta)}{f(R|Y_{obs}, \xi) f(Y_{obs}|\theta)} \\
&= f(Y_{mis}|Y_{obs}, \theta).
\end{aligned}
$$

Thus, in order to impute missing values Y_{mis}, we only need to correctly specify the likelihood function $L(\theta|Y_{obs}, Y_{mis})$, or in other words, $f(Y_{mis}|Y_{obs}, \theta)$.

However, when d is large, i.e., the number of terms that are missing is large, the distribution assumption $f(Y_{mis}|Y_{obs}, \theta)$ is still complicated in calculation. As an alternative, we try to utilize the conditional distribution of one single missing value conditional on all other variables, which leads to the chained equations. Specifically, the parametric MICE algorithm is implemented as follows:

i) Create initial imputations $Y^{(0)}_{mis} = (Y^{(0)}_{mis,1}, \cdots, Y^{(0)}_{mis,d})$ of the missing values by some approximated procedure, e.g., the mean imputation.

ii) Given current imputed values after k iterations, $Y^{(k)}_{mis} = (Y^{(k)}_{mis,1}, \cdots, Y^{(k)}_{mis,d})$, generate updated imputed values for each variable as draws from the following sequence of predictive distribution:

$$
\begin{aligned}
Y^{(k+1)}_{mis,1} &\sim f\left(Y_{mis,1}|Y_{obs}, Y^{(k)}_{mis,2}, \cdots, Y^{(k)}_{mis,d}\right) \\
Y^{(k+1)}_{mis,2} &\sim f\left(Y_{mis,1}|Y_{obs}, Y^{(k+1)}_{mis,1}, Y^{(k)}_{mis,3} \cdots, Y^{(k)}_{mis,d}\right) \\
Y^{(k+1)}_{mis,3} &\sim f\left(Y_{mis,1}|Y_{obs}, Y^{(k+1)}_{mis,1}, Y^{(k+1)}_{mis,2}, Y^{(k)}_{mis,4} \cdots, Y^{(k)}_{mis,d}\right) \\
\vdots &\sim \quad \vdots
\end{aligned}
$$

$$Y^{(k+1)}_{mis,d} \;\sim\; f\left(Y_{mis,1}|Y_{obs}, Y^{(k+1)}_{mis,1}, \cdots, Y^{(k+1)}_{mis,(d-1)}\right)$$

Thus based on the above chained equations, we obtained the imputed values after $(k+1)$ iterations, $Y^{(k+1)}_{mis} = (Y^{(k+1)}_{mis,1}, \cdots, Y^{(k+1)}_{mis,d})$.

iii) In the $k-th$ iteration, denote our estimator of the parameter θ as $\hat{\theta}^{(k)}$, set the burning out iterations k_1 and k_2. We will end after the k_2 iteration, and the overall estimator for θ would be

$$\hat{\theta} = \frac{1}{k_2 - k_1} \sum_{i=k_1+1}^{k_2} \hat{\theta}^{(i)}$$

When the set of conditional distributions correspond to a coherent joint distribution, this algorithm is a Gibbs' sampler, and the sequence converges to draws from the correct posterior distribution. In other settings, convergence to a stationary distribution is not guaranteed; however, the procedure appears to produce useful imputations, provided the conditional distributions yield good fits to the observed data. The increased flexibility in modeling these conditional distributions may out weigh the lack of clear theoretical justification of the method.

6.2. Non-parametric Multiple Imputation by Chained Equations

The non-parametric MICE algorithm follows a similar idea. But instead of assign a distribution model $f(Y_{mis}|Y_{obs}, \theta)$, it impose a regression model. For instance, assume the complete data Y forms a panel

$$\begin{pmatrix}
Y_{11} & \cdots & Y_{1(j(s)-1)} & Y_{1j(s)} & Y_{1(j(s)+1)} & \cdots & Y_{1m} \\
Y_{21} & \cdots & Y_{2(j(s)-1)} & Y_{2j(s)} & Y_{2(j(s)+1)} & \cdots & Y_{2m} \\
\vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
Y_{i(s)1} & \cdots & Y_{i(s)(j(s)-1)} & Y_{mis,s} & Y_{i(s)(j(s)+1)} & \cdots & Y_{i(s)m} \\
\vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
Y_{n1} & \cdots & Y_{n(j(s)-1)} & Y_{nj(s)} & Y_{n(j(s)+1)} & \cdots & Y_{nm}
\end{pmatrix}$$

Thus,

i) Create initial imputations $Y^{(0)}_{mis} = (Y^{(0)}_{mis,1}, \cdots, Y^{(0)}_{mis,d})$ of the missing values by some approximated procedure, e.g., the mean imputation.

ii) Given current imputed values after k iterations, $Y^{(k)}_{mis} = (Y^{(k)}_{mis,1}, \cdots, Y^{(k)}_{mis,d})$, generate updated imputed values for each variable as draws from the following sequence of predictive regression model, i.e., for $Y_{mis,s}$, suppose it is $(i(s), j(s))$-th term in the panel, denote three terms, $Y_{i(s),-j(s)}$ being the $i(s)$-th row in the panel with the $j(s)$-th item being deleted, $Y_{-i(s),j(s)}$ being the $j(s)$-th column in the panel with the $i(s)$-th item being deleted, and $Y_{-i(s),-j(s)}$ are the matrix in the panel with the $i(s)$-th row and the $j(s)$-th column being deleted,

$$Y_{i(s),-j(s)} = (Y_{i(s),1}, \cdots, Y_{i(s),(j(s)-1)}, Y_{i(s),(j(s)+1)}, \cdots, Y_{i(s),m})$$

$$Y_{-i(s),j(s)} = (Y_{1,j(s)}, \cdots, Y_{(i(s)-1),j(s)}, Y_{(i(s)+1),j(1)}, \cdots, Y_{n,j(s)})^T$$

$$Y_{-i(s),-j(s)} = \begin{pmatrix} Y_{11} & \cdots & Y_{1(j(s)-1)} & Y_{1(j(s)+1)} & \cdots & Y_{1m} \\ Y_{21} & \cdots & Y_{2(j(s)-1)} & Y_{2(j(s)+1)} & \cdots & Y_{2m} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ Y_{(i(s)-1)1} & \cdots & Y_{(i(s)-1)(j(s)-1)} & Y_{(i(s)-1)(j(s)+1)} & \cdots & Y_{(i(s)-1)m} \\ Y_{(i(s)+1)1} & \cdots & Y_{(i(s)+1)(j(s)-1)} & Y_{(i(s)+1)(j(s)+1)} & \cdots & Y_{(i(s)+1)m} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ Y_{n1} & \cdots & Y_{n(j(s)-1)} & Y_{n(j(s)+1)} & \cdots & Y_{nm} \end{pmatrix}$$

where the missing values inside $Y_{-i(s),-j(s)}$, $Y_{i(s),-j(s)}$, $Y_{-i(s),j(s)}$ are replaced by what we have imputed after $k-th$ iterations and part of the imputations in $(k+1)-th$ iteration, i.e., the missing values are replaced by $(Y_{mis,1}^{(k+1)}, \cdots, Y_{mis,(s-1)}^{(k+1)}, Y_{mis,(s+1)}^{(k)}, \cdots, Y_{mis,d}^{(k)})$, with the only $Y_{mis,s}$ remain missing and being a "place holder".

Now, we regress $Y_{-i(s),j(s)}$ on $Y_{-i(s),-j(s)}$ to obtain the regression coefficients β_s, i.e.,

$$Y_{-i(s),j(s)} = Y_{-i(s),-j(s)}\beta_s + \epsilon_s$$

and we obtain the regression coefficients β_s through least square estimator, i.e.,

$$\hat{\beta}_s = \arg\min_{\beta_s} \left\| Y_{-i(s),j(s)} - Y_{-i(s),-j(s)}\beta_s \right\|_2,$$

and obtain the imputation for $Y_{mis,s}$ in the $(k+1)$-th iteration as

$$Y_{mis,s}^{(k+1)} \triangleq Y_{i(s),-j(s)}\hat{\beta}_s.$$

The procedure continues from $s = 1, 2, \cdots, d$ and completes the $(k+1)$-th iteration.

iii) In the $k-th$ iteration, denote our estimator of the parameter θ as $\hat{\theta}^{(k)}$, set the burning out iterations k_1 and k_2. We will end after the k_2 iteration, and the overall estimator for θ would be

$$\hat{\theta} = \frac{1}{k_2 - k_1} \sum_{i=k_1+1}^{k_2} \hat{\theta}^{(i)}$$

Two things worth mentioning are, first, the regression model are not restricted to linear regression. If the outcome $Y_{mis,s}$ is indeed to be binary, we may fit logistic regression instead of linear regressions. Therefore, the regression to fit within the model need to be adjusted with respect to the outcome; second, the regression model does not need to utilize the whole $Y_{-i(s),-j(s)}$ but only part of the columns of $Y_{-i(s),-j(s)}$.

Notice the above procedure based also on chained equations, and thus lead to the name MICE algorithm.

References

Dahiya, R. C., & Korwar, R. M. (1980). Maximum likelihood estimates for a bivariate normal distribution with missing data. The Annals of Statistics, 8(3), 687-692.