A probabilistic introduction to missing data

Yen-Chi Chen
University of Washington
December 6, 2024

This note is modified from the lecture notes of STAT 535-Statistical Machine Learning at the University Washington. The goal of this note is to provide an introduction on missing data methodologies from a probability perspective.

1 Probability model of the missing data

Missing data is a very common problem in every scientific research. In a survey sample, it occurs when there are individuals who refuse to answer some questions. In a medical research, it happens when participants drop out of the study.

To study the missing data problem, we consider a very simple scenario where we have two random variables $X, Y \in \mathbb{R}$ such that X is always observed but Y could be missing. The following tables describe an example of missing data problem.

ID	X	Y
001	13	5
002	7	NA
003	11	3
004	21	NA
005	15	NA
006	9	12

ID	X	Y	R
001	13	5	1
002	7	NA	0
003	11	3	1
004	21	NA	0
005	15	NA	0
006	9	12	1

Table 1: **Left:** A data example with missingness in Y. The notation NA represents the missingness of that entry. **Right:** The same data and we include the response indicator.

Even in this simple scenario, it is not easy to write down a probability model because sometimes we observe Y and sometimes we do not. Here are two frameworks that we can represent the data generating process.

Augmenting the support of Y. We allow $Y \in \mathbb{R} \cup \{\texttt{NA}\}$ such that when $Y \in \texttt{NA}$, it means Y is missing. In this case, our data can be represented as IID random elements

$$(X_1, Y_1), \cdots, (X_n, Y_n) \in \mathbb{R} \times (\mathbb{R} \cup \{\texttt{NA}\})$$

from a probability measure. This is also called *missingness incorporated in attributes*[1].

Response indicator. We introduce a binary random variable $R \in \{0, 1\}$ such that $R = 1$ when Y is observed and $R = 0$ when Y is missing. We can then describe the data generating process as follows. We first generate

[1] See Twala, B. E., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. Pattern Recognition Letters, 29(7), 950-956.

$(X,R) \in \mathbb{R} \times \{0,1\}$. If $R = 0$, this is what we observe. If $R = 1$, then we generate Y given $X, R = 1$. The data will be IID random elements generated from the above two-stage process.

While both framework provide a probability model for missing data problems, it turns out that the response indicator framework is easier to work with because the support augmenting approach makes the underlying probability measure of Y to be a mixture of discrete and continuous. Thus, we will use the response indicator framework in the rest of the note.

2 Simple missing data problem

Under the response indicator framework, our data are IID random elements of either $(X_i, Y_i, R_i = 1)$ or $(X_i, R_i = 0)$. Now we consider a simple problem: estimating the mean of Y. Namely, the parameter of interest is

$$\mu = \mathbb{E}(Y).$$

A naive approach is to use the sample average of the observed Y_i's as an estimator. This method is known as the *complete-case estimator*, which can be expressed as

$$\widehat{\mu}_{\mathsf{CC}} = \frac{\sum_{i=1}^n Y_i R_i}{\sum_{j=1}^n R_j}.$$

Note that $\sum_{i=1}^n R_j$ is the total number of complete-cases (observations without missing values).

Is the complete-case estimator consistent? To analyze this problem, we consider its population version. Clearly, when the sample size is large, $\widehat{\mu}_{\mathsf{CC}}$ is approaching

$$\bar{\mu}_{\mathsf{CC}} = \frac{\mathbb{E}(YR)}{\mathbb{E}(R)}.$$

This quantity is generally NOT $\mu = \mathbb{E}(Y)$ unless R and Y are uncorrelated.

In fact, you can show that without any further assumptions, there is no consistent estimators that can estimate μ. In other words, μ is *unidentifiable* from the observed data. To see this, you can easily decompose

$$\mu = \mathbb{E}(Y) = \mathbb{E}(Y|R=1)P(R=1) + \mathbb{E}(Y|R=0)P(R=0).$$

The second quantity $\mathbb{E}(Y|R=0)$ is unidentifiable from the data because when $R = 0$, we never get to observe Y, so this conditional expectation can be anything without affecting our observed data.

To resolve the identification problem, a common assumption is

$$Y \perp R|X, \tag{1}$$

which is sometimes called the missing-at-random assumption. While it is indeed a special case of the missing-at-random (MAR) assumption, it is also a special case of the complete-case missing value (CCMV) assumption that we will talk about later. Many nice properties of this assumption are in fact results from the complete-case missing value assumption.

2.1 Regression adjustment

Under assumption in equation (1), we have the following interesting property:

$$\mu = \mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) \overset{(1)}{=} \mathbb{E}(\mathbb{E}(Y|X, R=1)) = \mathbb{E}(m_1(X)) \equiv \bar{\mu}_{\mathsf{RA},1},$$

where

$$m_1(x) = \mathbb{E}(Y|X=x, R=1)$$

is the observed-outcome regression model that can be estimated easily by fitting a regression model on the data with observed Y. The quantity m_1 is called the *nuisance* parameter in this case.

Thus, this motivates us to use the following estimator

$$\widehat{\mu}_{\mathsf{RA},1} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_1(X_i), \tag{2}$$

where $\widehat{m}_1(x)$ is an estimated regression model. $\widehat{\mu}_{\mathsf{RA},1}$ is often called the *regression adjustment (RA)* estimator or the g-computation method (from the language in causal inference).

Essentially, $\widehat{\mu}_{\mathsf{RA},1}$ behaves as if we are 'imputing' every Y_i by a prediction $\mathbb{E}(Y_i|X_i)$. This give hints that we may consider a variant of this method that we only 'impute' the missing Y:

$$\widehat{\mu}_{\mathsf{RA},2} = \frac{1}{n}\sum_{i=1}^{n} Y_i R_i + \widehat{m}_1(X_i)(1-R_i). \tag{3}$$

In the above expression, when $R_i = 1$, i.e., Y_i is observed, we use the observed value and when Y_i is missing ($R_i = 0$), we use the predicted value from the regression model.

To see why $\widehat{\mu}_{\mathsf{RA},2}$ is a consistent estimator, we consider it population version:

$$\begin{aligned}
\bar{\mu}_{\mathsf{RA},2} &\equiv \mathbb{E}(YR + m_1(X)(1-R)) \\
&= \mathbb{E}(\mathbb{E}(YR|X) + m_1(X)\mathbb{E}(1-R|X)) \\
&= \mathbb{E}(\mathbb{E}(Y|X, R=1)\underbrace{\mathbb{E}(R|X)}_{=\pi(X)} + m_1(X)(1-\pi(X))) \\
&= \mathbb{E}(m_1(X)) = \mu,
\end{aligned}$$

where

$$\pi(X) = \mathbb{E}(R|X) = P(R=1|X)$$

is the probability of observing Y given X, which is a key quantity related to the propensity score.

2.2 Inverse probability weighting

Under equation (1), the observed Y implies an interesting result:

$$\mathbb{E}(YR) = \mathbb{E}(\mathbb{E}(YR|X)) = \mathbb{E}(\mathbb{E}(Y|X)\mathbb{E}(R|X)) = \mathbb{E}(\mathbb{E}(Y|X)\pi(X)).$$

Thus, this implies that

$$\bar{\mu}_{\mathsf{IPW}} \equiv \mathbb{E}\left(\frac{YR}{\pi(X)}\right) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y) \equiv \mu.$$

The quantity $\pi(x)$ is the nuisance parameter in this case.

With the above result, we can construct an *inverse probability weighting (IPW)* estimator via

$$\widehat{\mu}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{\widehat{\pi}(X_i)}, \tag{4}$$

where $\widehat{\pi}(x)$ is an estimator of $\pi(x) = \mathbb{E}(R|X = x)$.

2.3 Augmented and doubly-robust estimator

From both the regression adjustment and inverse probability approaches, we find an interesting result: suppose we have a function $g(X,R,Y)$ that can be evaluated with the observed data and satisfies $\mathbb{E}(g(X,R,Y)) = 0$, we can construct an estimator by using the fact that

$$\mathbb{E}\left(\frac{YR}{\pi(X)} + g(X,R,Y)\right) = \mathbb{E}\left(\frac{YR}{\pi(X)}\right) = \mu.$$

For instance,

$$\mathbb{E}\left(\frac{YR}{\pi(X)} + R(m_1(X) - Y)\right) = \mu$$

because $\mathbb{E}[R(m_1(X) - Y)] = \mathbb{E}[\pi(X)(m_1(X) - m_1(X))] = 0$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{\widehat{\pi}(X_i)} + R_i(\widehat{m}_1(X_i) - Y_i)$$

is expected to be a consistent estimator of μ when both $\widehat{\pi}, \widehat{m}_1$ are consistently estimated.

This idea is called *augmentation* and the above estimator is an augmented IPW estimator.

This prompts to an interesting question

> There are many possible way that we can augment our RA or IPW estimators. Is there a particular augmentation that is better than the others?

Of course the answer may vary depending on what do we mean by 'better'. However, there is a particular augmentation that is useful in practice.

Consider the following quantity

$$\Psi(X,Y,R;\pi,m_1) = \frac{YR}{\pi(X)} + m_1(X)\left(1 - \frac{R}{\pi(X)}\right).$$

Clearly, this is an augmentation because $\mathbb{E}\left(m_1(X)(1-\frac{R}{\pi(X)})\right)=0$. So we have

$$\mu_{\mathsf{DR}} \equiv \mathbb{E}(\Psi(X,Y,R;\pi,m_1) = \mathbb{E}\left(\frac{YR}{\pi(X)} + m_1(X)\left(1-\frac{R}{\pi(X)}\right)\right) = \mu,$$

which leads to the estimator

$$\begin{aligned}\widehat{\mu}_{\mathsf{DR}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i R_i}{\widehat{\pi}(X_i)} + \widehat{m}_1(X_i)\left(1-\frac{R_i}{\widehat{\pi}(X_i)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_1(X_i) + \frac{R_i}{\widehat{\pi}(X_i)}(Y_i-\widehat{m}_1(X_i)).\end{aligned} \tag{5}$$

The estimator $\widehat{\mu}_{\mathsf{DR}}$ is called the *doubly-robust* estimator and it can be viewed as an augmented estimator from either IPW or the RA approaches.

The doubly-robust means that we only need either one of $\widehat{m}_1(x)$ or $\widehat{\pi}(x)$ to be a consistent estimator to ensure the consistency of $\widehat{\mu}_{\mathsf{DR}}$. Namely, even if $\widehat{m}_1(x)$ does not converge to the true $m_1(x)$, as long as $\widehat{\pi}(x)$ is consistent to $\pi(x)$, $\widehat{\mu}_{\mathsf{DR}}$ can consistently estimate μ.

The doubly-robust estimator is also a *semi-parametric efficient estimator*. See http://faculty.washington.edu/yenchic/short_note/note_EIF.pdf for more details on the semi-parametric efficiency theory. I would also recommend the following book for semi-parametric theory and missing data:

> Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer Science & Business Media.

Doubly-robustness. To see why $\widehat{\mu}_{\mathsf{DR}}$ has this doubly robustness, we consider the quantity $\Psi(X,Y,R;\pi,m_1)$. We will show that as long as either π or m_1 is correct, this quantity has the mean being μ. First, we consider $\bar{m}_1 \neq m_1$.

$$\begin{aligned}\mathbb{E}(\Psi(X,Y,R;\pi,\bar{m}_1)) &= \mathbb{E}\left(\frac{YR}{\pi(X)} + \bar{m}_1(X)\left(1-\frac{R}{\pi(X)}\right)\right) \\ &= \mathbb{E}\left(\frac{YR}{\pi(X)}\right) + \mathbb{E}\left(\bar{m}_1(X)\left(1-\frac{R}{\pi(X)}\right)\right) \\ &= \mu + \mathbb{E}\left(\bar{m}_1(X)\underbrace{\mathbb{E}\left(1-\frac{R}{\pi(X)}|X\right)}_{=0}\right) \\ &= \mu.\end{aligned}$$

Next, we consider $\bar{\pi} \neq \pi$.

$$\begin{aligned}
\mathbb{E}(\Psi(X,Y,R;\bar{\pi},m_1)) &= \mathbb{E}\left(\frac{YR}{\bar{\pi}(X)} + m_1(X)\left(1-\frac{R}{\bar{\pi}(X)}\right)\right) \\
&= \mathbb{E}\left(m_1(X) + \frac{R}{\bar{\pi}(X)}(Y-m_1(X))\right) \\
&= \mu + \mathbb{E}\left(\frac{R}{\bar{\pi}(X)}(Y-m_1(X))\right) \\
&= \mu + \mathbb{E}\left(\frac{R}{\bar{\pi}(X)}\underbrace{\mathbb{E}(Y-m_1(X)|X,R=1)}_{=0}\right) \\
&= \mu.
\end{aligned}$$

As a result, we only need one of the two nuisances to be consistently estimated to guarantee the consistency of $\widehat{\mu}_{\mathsf{DR}}$.

2.4 Imputation

Imputation is a very popular approach for handling missing data. A feature of imputation is that after imputation, we can use the imputed data for various statistical analysis. While RA approach implicitly implies an imputation method, this imputation method only works if we are interested in estimating $\mu = \mathbb{E}(Y)$. We have to use another imputation model if we change the parameter of interest.

This leads to the question:

> What should be the correct imputation model that we use?

Under our probability model, we can describe the distribution of the unobserved Y as

$$p(y|x,R=0) \tag{6}$$

and when $R_i = 0$, we should be imputing Y_i by sampling from

$$Y_i \sim p(y|X_i,R_i=0).$$

The distribution $p(y|x,R=0)$ is called *extrapolation distribution/density*.

As is mentioned previously, the imputation model $p(y|x,R=0)$ is not identifiable from the data because we never observe both (X,Y) when $R=0$. However, under the assumption (1), implies that

$$Y \perp R|X \qquad \Longrightarrow \qquad p(y|x,R=0) = p(y|x,R=1),$$

which is identifiable from the observed data!

The quantity $p(y|x,R=1)$ can be estimated by a conditional PDF estimator using the complete cases. Here are two simple examples of how we may estimate the conditional PDF.

Example: Gaussian linear model. We assume that $p(y|x,R=1)$ follows a Normal distribution $N(\alpha+\beta x,\sigma^2)$. The parameters α,β can be estimated by the least-squared method on the complete cases:

$$\widehat{\alpha},\widehat{\beta}=\mathsf{argmin}_{\alpha,\beta}\sum_{i=1}^{n}R_i(Y_i-\alpha-\beta X_i)^2.$$

and we estimate σ^2 using the residuals

$$\widehat{\sigma}^2=\frac{1}{n_1-2}\sum_{i=1}^{n}R_i(Y_i-\widehat{\alpha}-\widehat{\beta}X_i)^2,$$

where $n_1=\sum_{i=1}^{n}R_i$ is the sample size of complete cases ($R_i=1$). For the observations with missing Y, i.e., $R_i=0$, we impute Y_i by

$$\widetilde{Y}_i\sim N(\widehat{\alpha}+\widehat{\beta}X_i,\widehat{\sigma}^2).$$

Example: kernel density estimator. Alternatively, we may use a nonparametric estimator for our imputation model. Here we illustrate the idea using the KDE:

$$\begin{aligned}\widehat{p}(y|x,R=1)&=\frac{\widehat{p}(x,y,R=1)}{\widehat{p}(x,R=1)}\\&=\frac{\frac{1}{nh_xh_y}\sum_{i=1}^{n}R_iK\left(\frac{Y_i-y}{h_y}\right)K\left(\frac{X_i-x}{h_x}\right)}{\frac{1}{nh_x}\sum_{j=1}^{n}R_jK\left(\frac{X_j-x}{h_x}\right)}\\&=\frac{\frac{1}{h_y}\sum_{i=1}^{n}R_iK\left(\frac{Y_i-y}{h_y}\right)K\left(\frac{X_i-x}{h_x}\right)}{\sum_{j=1}^{n}R_jK\left(\frac{X_j-x}{h_x}\right)}\\&=\frac{1}{h_y}\sum_{i=1}^{n}W_i(x)K\left(\frac{Y_i-y}{h_y}\right),\end{aligned}$$

where $W_i(x)=\frac{R_iK\left(\frac{X_i-x}{h_x}\right)}{\sum_{j=1}^{n}R_jK\left(\frac{X_j-x}{h_x}\right)}$ is the weight of the i-th observation. Note that the conditional PDF estimator from the KDE looks like a kernel regression estimator. Sampling from $\widehat{p}(y|x,R=1)$ can be done easily from a two-stage procedure:

1. We sample $L\in\{1,2,\cdots,n\}$ from $P(L=i)=W_i(x)$.
2. Given L, we then generate Y from the PDF $K\left(\frac{Y_L-y}{h_y}\right)/h$. If K is the Gaussian kernel, this means that $Y|L\sim N(Y_L,h_y^2)$.

Multiple imputation. In practice, if we only impute the missing values once, the Monte Carlo errors due to random imputation may be large. Thus, we often repeat the imputation procedure multiple times, leading to multiple imputed data and combine all data to perform downstream analysis. This is called multiple imputation. Here is an interesting connection from multiple imputation to the regression adjustment. Suppose we

have an imputation model $\widehat{p}(y|x,R=1)$ and let $(X_i,R_i=0)$ be the data with missing Y_i. After performing imputation M times, we obtain

$$\widetilde{Y}_i^{(1)},\cdots,\widetilde{Y}_i^{(M)} \sim \widehat{p}(y|X_i,R=1).$$

The average of these imputed values is

$$\frac{1}{M}\sum_{m=1}^{M}\widetilde{Y}_i^{(m)} \approx \int y\widehat{p}(y|X_i,R=1)dy = \widehat{m}_1(X_i).$$

Namely, the average of the imputed values is a Monte Carlo approximation of the implied regression function from the conditional PDF $\widehat{p}(y|x,R=1)$. Thus, the multiple imputation can be viewed as a Monte Carlo approximation of the regression adjustment.

2.5 Estimating other parameters of interest

All the above analysis can be generalized to other parameters of interest as long as it can be expressed as

$$\mu_\omega = \mathbb{E}[\omega(X,Y)]$$

for some ω. Interestingly, if we change the parameter of interest, the RA estimator has to be modified because the outcome regression

$$m_{\omega,1}(X) = \mathbb{E}(\omega(X,Y)|X,R=1)$$

depends on ω. So we have estimate the nuisance parameter $m_{\omega,1}$ again.

However, the IPW estimator remains the same. We are still using the same expression

$$\mu_\omega = \mathbb{E}\left(\frac{\omega(X,Y)R}{\pi(X)}\right)$$

and the nuisance parameter π remains identical.

Similarly, the imputation model $p(y|x,R=1)$ is invariant to the choice of parameter of interest, so once we have created an imputation model, we do not need to make any adjustment if we change the parameter of interest.

In practice, people often find RA estimator numerically more stable than the IPW estimator because the estimated probability $\widehat{\pi}(x)$ can be very close to 0, making the ratio $\frac{\omega(X_i,Y_i)R_i}{\widehat{\pi}(X_i)}$ very large.

3 Missing data: general problems and missing at random

When there are more than one variable that are subject to missing, the problem gets a lot more complex. Consider the case where each individual has d variables $X_1,\cdots,X_d$ and all of them may be missing and we may even have many of them missing at the same time. There are two categories of the missing patterns:

1. **Monotone missingness.** In this case, if X_t is missing, then X_s is also missing for any $s > t$. This occurs a lot in medical research due to *dropout* of the individuals. For instance, let X_t denote the BMI of an individual at year t. If this individual left the study at time point τ, then we only observe $X_1, \cdots, X_\tau$ from this individual. Any information beyond year τ is missing.

2. **Non-monotone missingness.** When the missing pattern is not monotone, it is called non-monotone missingness. The non-monotone missing data is a lot more challenging than monotone missing data because there are many possible missing pattern that can occur in the data. If there are d variables, them monotone missing data has d different missing patterns but the non-monotone case may have up to 2^d different missing patterns!

Let $R \in \{0,1\}^5$ be a multi-index set that denotes the observed pattern and we use the notation $X_R = (X_i : R_i = 1)$. For instance, $R = 11001$ means that we observe variable X_1, X_2, and X_5 and $X_{11001} = (X_1, X_2, X_5)$. Table 2 provides an example of three study variables and we include the corresponding response vector R.

ID	X_1	X_2	X_3	R
001	13	0	2.2	111
002	7	NA	2.7	101
003	NA	NA	2.5	001
004	2	1	1.3	111
005	8	0	NA	110
006	NA	0	NA	010
007	15	1	2.2	111
008	NA	1	1.7	011

Table 2: An example of non-monotone missing data with three study variables X_1, X_2, X_3 and the corresponding response vector R. For ID=001, $X_R = (13, 0, 2.2)$ whereas for ID=002, $X_R = (7, 2.7)$.

Generally speaking, researchers divide the missing data assumptions into three groups:

1. **MCAR: missing completely at random.** This means that $X \perp R$. Namely, the missingness is totally irrelevant to any variables of interest. Note that MCAR can be tested using the data. Using the above notations, MCAR can be written as

$$P(R = r|X) = P(R = r).$$

2. **MAR: missing at random.** The MAR assumption assumes that

$$P(R = r|X) = P(R = r|X_r),$$

   namely, the probability of seeing a pattern $R = r$ only depends on the observed variable.

3. **MNAR: missing not at random.** When the missingness is not MAR nor MCAR, it is called MNAR– missing not at random.

MAR is a very popular assumption although it may not be reasonable in some cases. Why is the MAR still so popular in practice?

There are two reasons for why MAR is so popular. The first reason is that in both monotone and non-monotone case, *MAR makes the likelihood inference a lot easier.* This is due to a property called ignorability. The second reason is that under monotone missing data problem, MAR provides an elegant way to identify the entire distribution function.

3.1 Likelihood inference with MAR

The MAR has a nice property called the *ignorability*, which holds in both monotone and non-monotone missingness. Consider the joint density function $p(x,r)$ of both variable of interest X and the missing pattern R. Recall that $X_R = (X_i : R_i = 1)$ are the observed variables under pattern R. We also denote $X_{\bar{R}} = (X_i : R_i = 0)$ as the missing variables.

We can then factorize it into

$$p(x,r) = P(R = r|X = x)p(x).$$

Suppose we use parametric models separately for both $P(R = r|X = x)$ and $p(x)$, leading to

$$p(x,r;\phi,\theta) = P(R = r|X = x;\phi)p(x;\theta) \overset{(MAR)}{=} P(R = r|X_r = x_r;\phi)p(x;\theta),$$

where θ is the parameter for modeling $p(x)$ and ϕ is the parameter for modeling the missing probability $P(R = r|X_r = x_r)$ (this separability of parameter together with MAR is often called *ignorability*). In our data, what we observe are (x_r, r) so we should integrate over the missing variables $x_{\bar{r}}$:

$$p(x_r, r;\phi,\theta) = \int p(x,r;\phi,\theta)dx_{\bar{r}} = P(R = r|X_r = x_r;\phi)\int p(x;\theta)dx_{\bar{r}}.$$

Thus, the log-likelihood function is

$$\begin{aligned}
\ell(\theta,\phi|x_r,r) &= \log P(R = r|X_r = x_r;\phi) + \log\int p(x;\theta)dx_{\bar{r}} \\
&= \ell(\phi|x_r,r) + \ell(\theta|x_r), \\
\ell(\phi|x_r,r) &= \log P(R = r|X_r = x_r;\phi) \\
\ell(\theta|x_r,r) &= \log\int p(x;\theta)dx_{\bar{r}}.
\end{aligned}$$

The above factorization is very powerful–it decouples the problem of estimating θ and the problem of estimating ϕ!

Namely, if we are only interested in the distribution of X, we do not even need to deal with ϕ. We just need to maximize $\ell(\theta|x_r)$. So finding the MLE of θ can be done without estimating the parameter ϕ, leading to a relatively simple problem.

However, computing the MLE may still be a non-trivial problem in practice. To resolve this problem, we often use the following result under MAR:

$$\text{(MAR)} \qquad \Longrightarrow \qquad p(x_{\bar{r}}|x_r,r) = p(x_{\bar{r}}|x_r).$$

Namely, under the (MAR), the correct imputation model is the marginal model $p(x_{\bar{r}}|x_r;\theta)$, this makes it particularly easy to impute the data when we have an pilot estimate $p(x_{\bar{r}}|x_r;\theta^{(t)})$. This insight leads to the following EM algorithm. Notably, the above imputation model property is used in the E-step.

EM algorithm. Estimating θ via maximizing $\ell(\theta|x_r)$ is often done via the EM algorithm. The EM algorithm is an iterative algorithm that finds a stationary point. It consists of two steps, an expectation step (E-step) and a maximization step (M). Given an initial guess of the parameter $\theta^{(0)}$, the EM algorithm iterates the following two steps until convergence ($t = 0, 1, 2, 3, \cdots$):

1. **E-steps.** Compute
$$Q(\theta;\theta^{(t)}|X_r) = \mathbb{E}(\ell(\theta|X);X_r,\theta^{(t-1)}) = \int \ell(\theta|,x_{\bar{r}},X_r)p(x_{\bar{r}}|X_r;\theta^{(t)})dx_{\bar{r}}.$$
2. **M-steps.** Update
$$\theta^{(t+1)} = \mathsf{argmax}_\theta Q(\theta;\theta^{(t)}|X_r).$$

Note that in practice, we have n observations so the Q function will be

$$Q_n(\theta;\theta^{(t)}) = \frac{1}{n}\sum_{i=1}^{n} Q(\theta;\theta^{(t)}|X_{i,R_i})$$

and the M-step will be

$$\theta^{(t+1)} = \mathsf{argmax}_\theta Q_n(\theta;\theta^{(t)}).$$

Clearly, the above EM algorithm does not involve ϕ, the parameter of $P(R = r|X_r)$. So we can totally ignore the model of missingness.

Under good conditions, the EM algorithm has the ascending property, i.e.,

$$\ell(\theta^{(t+1)}|X_r) \geq \ell(\theta^{(t)}|X_r),$$

and will converge to a stationary point. However, the problem is that the stationary point is not guarantee to be the global maximum (MLE). It could be a local mode or even a saddle point.

A good introduction on the EM algorithm and missing data is Section 8 of the following textbook:

> Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

3.2 MAR under monotone case

Under the monotone missing problem, let $T = \sum_{j=1}^{d} R_j$ denotes the index of the last observed variable. This is often referred to as the dropout time (the last time point before individual drops out of the study). Namely, the individual dropouts after time point T. We use the notation $X_{\leq t} = (X_1, \cdots, X_t)$. Table 3 provides an example of monotone missing data.

Then the MAR can be written as

$$P(T = t|X) = P(T = t|X_{\leq t}).$$

The above equation gives us a very powerful result–we can estimate the missing probability $P(T = t|X)$ for every $t = 1, \cdots, d$!

ID	X_1	X_2	X_3	R	T
001	13	0	2.2	111	3
002	7	NA	NA	100	1
003	5	NA	NA	100	1
004	2	1	1.3	111	3
005	8	0	NA	110	2
006	21	0	NA	001	2
007	15	1	2.2	111	3
008	13	1	1.7	111	3

Table 3: An example of monotone missing data of three study variables X_1, X_2, X_3. We include both the response indicator R and the corresponding dropout time T. For ID=001, $X_{\leq T} = (X_1, X_2, X_3) = (13, 0, 2.2)$ and for ID=002, $X_{\leq T} = X_1 = 7$ and for ID=005, $X_{\leq T} = (X_1, X_2) = (8, 0)$.

To see this, consider the case $t = 1$ so MAR implies

$$P(T = 1|X) = P(T = 1|X_1).$$

Note that $P(T > 1|X) = 1 - P(T = 1|X) = P(T \neq 1|X_1) = P(T > 1|X_1)$. Thus, we can estimate $P(T = 1|X_1)$ by comparing pattern $T = 1$ against $T > 1$ given the variable X_1, which is always observed. Thus, $P(T = 1|X)$ is estimatible. For $t = 2$, the MAR implies

$$P(T = 2|X) = P(T = 2|X_1, X_2).$$

Thus,

$$P(T > 2|X) = 1 - P(T = 2|X) - P(T = 1|X) = 1 - P(T = 2|X_1, X_2) - P(T = 1|X_1) = P(T > 2|X_1, X_2).$$

Again, we can compare the pattern $T = 2$ against $T > 2$ and estimate the probability $P(T = 2|X)$. We can keep doing this procedure, and eventually all missing probability $P(T = t|X)$ can be estimated.

For instance, if we are interested in estimating the parameter of interest $\rho = \mathbb{E}(\omega(X_1, \cdots, X_d))$, we can then use the inverse probability weighting (IPW) estimator[2]:

$$\widehat{\rho} = \frac{1}{n\widehat{P}(T = d|X)} \sum_{i=1}^{n} \omega(X_{i,1}, \cdots, X_{i,p}) I(T_i = d),$$

where $\widehat{P}(T = d|X)$ is an estimate of $P(T = d|X)$. $P(T = t|X)$ is called the propensity score.

4 Missing not at random: monotone cases

In MNAR, the missing data problem becomes a lot more complicated. There are two common strategies for handling MNAR–the selection models and the pattern mixture models approaches.

[2]See https://en.wikipedia.org/wiki/Inverse_probability_weighting for more details.

To simplify the problem, we consider monotone missing data problem. Even in this scenario, we will see several identifiability issues so we have to be very careful about our choice of model.

Recall that X denotes the study variable and T is the dropout time. We are interesting in the *full-data density* $p(x,t)$; note that $p(x,t)$ implies the joint PDF of the study variable $p(x)$.

A useful reference: https://content.sph.harvard.edu/fitzmaur/lda/C6587_C018.pdf.

4.1 Selection models

Selection models decompose the full-data density using

$$p(x,t) = P(T=t|x)p(x),$$

where $P(T=t|x)$ is called the missing probability or missing data mechanism.

A common strategy in selection model is to identify $P(T=d|x)$, where d is the end time of the study. There are two reasons for identifying $P(T=d|x)$. First, identifying this quantity is enough for constructing a consistent *inverse probability weighting (IPW)* estimator, similar to the one we saw in the causal inference. The other reason is that we can easily estimate the PDF $p(x,T=d)$ by using the observations without missing entries. If $P(T=d|x)$ is known, then we can identify $p(x)$ using $p(x) = \frac{p(x,T=d)}{P(T=d|x)}$.

The MAR and MCAR conditions are often expressed in a selection model framework. Formally, the MCAR is

$$P(T=t|X) = P(T=t).$$

Namely, the probability of any dropout time is totally independent of the study variable X. The MAR is

$$P(T=t|X) = P(T=t|X_{\leq t}).$$

In other words, the conditional probability of the dropout time only depend on the observed variables.

As we have mentioned, the selection model allows a simple way to construct a consistent estimator of a parameter of interest via the IPW procedure. Here is a simple example. Suppose that the parameter of interest is a linear statistical functional $\theta = \theta(F) = \int \omega(x)dF(x)$, then it can be further written as

$$\theta = \int \omega(x)p(x)dx = \int \omega(x)\frac{p(x,T=d)}{P(T=d|x)}dx = \int \omega(x)\frac{dF(dx,T=d)}{P(T=d|x)}.$$

With an estimator of the selection probability $\widehat{P}(T=d|x)$ (and we only need to estimate the probability of fully-observed case), a simple IPW estimator of θ is

$$\widehat{\theta}_0 = \int \omega(x)\frac{d\widehat{F}(dx,T=d)}{\widehat{P}(T=d|x)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\omega(X_i)I(T_i=d)}{\widehat{P}(T=d|X_i)}. \tag{7}$$

You can show that $\widehat{\theta}_0$ is a consistent estimator (and it has asymptotical normality as well due to the Slutsky theorem). Moreover, the influence function (recall from the bootstrap lecture note) of $\widehat{\theta}_0$ can be easily derived so the variance of $\widehat{\theta}_0$ can be estimated via a plug-in estimate.

Although $\widehat{\theta}_0$ is elegant, it may not be the best estimator in the sense that after estimating the propensity score $P(T=t|x)$, we only rely on the completely observed data (the ones with $T_i = d$) to form the final estimator. Other observations are discarded entirely. Intuitively, this leads to an *inefficient* estimator.

To construct an efficient estimator, consider augmenting $\widehat{\theta}_0$ with an additional term

$$\widehat{\theta}_1 = \widehat{\theta}_0 + \frac{1}{n}\sum_{i=1}^{n}(I(T_i=\tau) - \widehat{P}(T_i=\tau|X_{i,\leq\tau}))g_\tau(X_{i,\leq\tau})I(T_i=\tau),$$

where $\tau < d$ is any time point and g_τ is a function of variable $x_{\leq\tau}$. The augmented term has an asymptotic mean 0 so $\widehat{\theta}_1$ is still a consistent estimator. The insight here is that the function g_τ is something we can choose– namely, we can choose it to minimize the variance of $\widehat{\theta}_1$ and this may leads to a reduction in the total variance compared to the estimator $\widehat{\theta}_0$. The same idea can be applied to every time point $\tau = 1, \cdots, d-1$, leading to an *augmented inverse probability weighting (AIPW)* estimator

$$\widehat{\theta}_{\mathsf{AIPW}} = \widehat{\theta}_0 + \frac{1}{n}\sum_{i=1}^{n}\sum_{\tau=1}^{d-1}(I(T_i=\tau) - \widehat{P}(T_i=\tau|X_{i,\leq\tau}))g_\tau(X_{i,\leq\tau})I(T_i=\tau).$$

With a proper choice of $g_\tau : \tau = 1, \cdots, d-1$, we can construct an estimator with the least variance. This leads to an efficient estimator. How to construct the functions $g_\tau : \tau = 1, \cdots, d-1$ is a central topic of *semi-parametric inference*.

Note that sometimes the AIPW (and IPW) estimators are constructed from solving an estimating equation. This occurs when the parameter of interest $\theta_0 = \theta(F)$ is defined through solving the equation

$$0 = \mathbb{E}(S(X;\theta_0)) = \int S(x;\theta_0)dF(x) = \int S(x;\theta)\frac{dF(dx, T=d)}{P(T=d|x)}.$$

In this case, the IPW estimator will be the solution to

$$0 = \int S(x;\widehat{\theta}_0)\frac{d\widehat{F}(dx, T=d)}{\widehat{P}(T=d|x)} = \frac{1}{n}\sum_{i=1}^{n}\frac{S(X_i;\widehat{\theta}_0)I(T_i=d)}{\widehat{P}(T=d|X_i)}$$

and we can augment it with a set of mean 0 terms to improve the efficiency.

If you are interested in the construction of AIPW, I would recommend the following textbook:

> Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer Science & Business Media.

Note: although we introduce AIPW estimators in the MNAR framework, they are often used in the MAR scenario because the identification of propensity score/selection probability $P(T=t|X)$ is challenging in MNAR. The MAR is a simple case where we can identify the propensity score entirely so AIPW estimators can be constructed easily. Essentially, as long as you can identify the selection probability, you can construct an IPW estimator and attempt to augment it to obtain AIPW estimator to improve the efficiency. So the direction of research is often on how to identify the selection probability.

4.2 Pattern mixture models

Pattern-mixture models (PMMs) use another factorization of the full-data density:

$$p(x,t) = p(x_{>t}|x_{\leq t},t)p(x_{\leq t}|t)P(T=t),$$

where the first term $p(x_{>t}|x_{\leq t},t)$ is called the *extrapolation density* and the later two terms $p(x_{\leq t}|t)P(T=t)$ are called *observed-data density*. The extrapolation density is unobservable and unidentifiable–it describes the distribution of the missing entries. The observed-data density is identifiable since at each dropout time $T=t$, we do observe variables $x_1,\cdots,x_t$.

Here is a nice review on PMMs for MNAR:

> Linero, A. R., & Daniels, M. J. (2018). Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science*, 33(2), 198-213.

The PMMs provide a clean separation about what is identifiable and what is not identifiable. So the strategy for identifying $p(x,t)$ is to make the extrapolation density be identifiable.

In monotone missing problems, the extrapolation density has the following product form:

$$p(x_{>t} \mid x_{\leq t},t) = \prod_{s=t+1}^{d} p(x_s \mid x_{<s}, T=t).$$

Thus, it suffices to identify each term in the product form to identify the extrapolation density. Several identifying restrictions have been proposed in the literature to identify the extrapolation density. The complete case missing value (CCMV) restriction equates that

$$p(x_s \mid x_{<s}, T=t) \overset{CC}{=} p(x_s \mid x_{<s}, T=d),$$

and the available case missing value (ACMV) restriction assumes that

$$p(x_s \mid x_{<s}, T=t) \overset{AC}{=} p(x_s \mid x_{<s}, T\geq s),$$

and the nearest case missing value (NCMV) restriction requires that

$$p(x_s \mid x_{<s}, T=t) \overset{NC}{=} p(x_s \mid x_{<s}, T=s)$$

for $s=t+1,\cdots,d$. In general, one can specify any subset of patterns $\mathcal{A}_{ts} \subset \{s,s+1,\cdots,d\}$ and construct a corresponding identifying restriction

$$p(x_s \mid x_{<s}, T=t) \overset{\mathcal{A}_{ts}}{=} p(x_s \mid x_{<s}, T\in\mathcal{A}_{ts});$$

this is called the donor-baed identifying restriction in the following paper:

> Chen, Y. C., & Sadinle, M. (2019). Nonparametric Pattern-Mixture Models for Inference with Missing Data. arXiv preprint arXiv:1904.11085.

If you make any of these assumptions, the extrapolation density (left-hand-side) equals to a quantity that is identifiable from the data (everything in the right-hand-side can be identified from the data), so you can then estimate the full-data density $p(x,t)$.

Example: ACMV with a simple Gaussian. We now demonstrate how we can use ACMV to create a simple estimator of the extrapolation density. Under ACMV, we have $p(x_s \mid x_{<s}, T=t) = p(x_s \mid x_{<s}, T>s)$ for $s = t+1, \cdots, d$. We model the right-hand sided as

$$N(\beta_{t,s}^T x_{<s}, \sigma_{t,s}^2),$$

where $\beta_{t,s}, \sigma_{t,s}^2$ can be estimated by

$$\widehat{\beta}_{t,s} = \mathsf{argmin}_\beta \sum_{i=1}^n (X_{i,s} - \beta^T X_{i,<s})^2 I(T_i \geq s), \qquad \widehat{\sigma}_{t,s}^2 = \frac{1}{n_s} \sum_{i=1}^n (X_{i,s} - \widehat{\beta}_{t,s}^T X_{i,<s})^2 I(T_i \geq s),$$

where $n_s = \sum_{i=1}^n I(T_i \geq s)$. Note that for any $T_i \geq s$, the vector $X_1, \cdots, X_s$ is observed, so the above estimator can be computed with the observed data. Thus, $\widehat{p}(x_s \mid x_{<s}, T=t)$ is $N(\widehat{\beta}_{t,s}^T x_{<s}, \widehat{\sigma}_{t,s}^2)$.

Remark: equivalence between MAR and ACMV under monotone missing data. MAR under monotone missingness is equivalent to the ACMV assumption. The equivalence between MAR and ACMV is shown in

> Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). *Monotone missing data and pattern-mixture models.* Statistica Neerlandica, 52(2), 153-161.

4.3 Imputation and pattern mixture models

In the previous section, we introduce the idea of imputation when there is only one variable missing. But it can be applied to cases where there are multiple missing entries. Suppose that we have an imputation procedure such that if we observe $X_{\leq T} = (X_1, \cdots, X_T)$ and the dropout time T, the procedure generates random numbers $X_{>T} = (X_{T+1}, \cdots, X_d)$ from a distribution Q.

You can always view this imputation procedure as a PMM such that the PDF corresponds to the imputation distribution Q is the underlying model on the extrapolation density. So any imputation method can be viewed as implicitly handling the problem with a PMM.

Example: ACMV with a simple Gaussian revisited. We now return to our previous example with ACMV and the simple Gaussian model. Recall that our estimated PMM is

$$\widehat{p}(x_s \mid x_{<s}, T=t) \sim N(\widehat{\beta}_{t,s}^T x_{<s}, \widehat{\sigma}_{t,s}^2)$$

for each t and $s = t+1, \cdots, d$. For an observation $T_i = t$ with $X_{i,\leq t}$ observed, here is how we will impute the missing entries $X_{i,>t}$:

1. Set $\widetilde{X}_{i,\leq t} = X_{i,\leq t}$.
2. For $s = t+1, t+2, \cdots, d$ do the following:

(a) Draw $\widetilde{X}_{i,s}$ from $N(\widehat{\beta}_{t,s}^T \widetilde{X}_{i,<s}, \widehat{\sigma}_{t,s}^2)$.

3. Return the imputed vector $\widetilde{X}_i = (\widetilde{X}_1, \cdots, \widetilde{X}_d)$.

In the monotone case, the PMM model leads to a *sequential imputation procedure*: we impute the missing entry one by one according to the time $T = t$.

4.4 Nonparametric Saturation

In MNAR, we need to make identifying restrictions so that the full-data distribution $F(x,t)$ (or $p(x,t)$) is identifiable. However, there is one property that an identifying restriction should have: the implied joint distribution should be compatible/consistent with what we observe. This property is called nonparametric saturation/nonparametric identification/just identification.

The idea is simple: because we can identify $F(x,t)$, we can pretend the implied joint distribution is the true generating distribution and generates a new missing data from it. The generated missing data should be similar to the original data we have.

MAR and any pattern mixture models satisfies this property (when we attempt to estimate the joint distribution via a nonparametric estimator). However, some identifying restrictions, such as the MCAR, does not satisfy this. Whenever you proposed a new MNAR restriction, you should always think about if the implied full-data distribution satisfies this property or not.

4.5 Sensitivity analysis

Sensitivity analysis is a common procedure in handling the missing data problem. In short, sensitivity analysis is to perturb the missing data assumption a bit and see how the conclusion changes. This is often required in handling missing data because as we have shown previously, there is no way to check if a missing data assumption is correct (unless we have additional information) so our conclusion relies heavily on our assumption of missingness. By perturbing the assumption on missingness, we are able to examine if our conclusion is robust to the missing data assumption.

In MAR, one common approach for sensitivity analysis is to introduce the model (called the exponential tilting strategy)

$$\log \frac{P(T=t|X)}{P(T=t|X_{\leq t})} = \gamma^T X,$$

where $\gamma \in \mathbb{R}^d$ is a sensitivity parameter such that if $\gamma = 0$, we have $\frac{P(T=t|X)}{P(T=t|X_{\leq t})} = 1$, which is the MAR condition. We vary γ and examine how the estimator changes as a function of γ and use this as a way to how sensitivity the estimator depends on the MAR assumption.

5 Missing not at random: non-monotone cases

Now we discuss general strategies to deal with nonmonotone missing data. The ideas of selection models and pattern mixture models still work in this case and they lead to different factorization of the problem. Recall that in this case, we use the binary vector $R \in \{0,1\}^d$ to denote the response pattern. To simplify the problem, we consider estimating the marginal mean

$$\theta = \mathbb{E}(\omega(X)).$$

5.1 Selection model and IPW

Let $\pi(x) = P(R = 1_d | X = x)$. This quantity behaves like the propensity score in the single missing data problem. To see this, we have the following equality

$$\begin{aligned}
\mathbb{E}\left(\frac{\omega(X)I(R=1_d)}{\pi(X)}\right) &= \int \frac{\omega(x)}{\pi(x)} p(x,1_d)dx \\
&= \int \frac{\omega(x)}{p(1_d|x)} p(x,1_d)dx \\
&= \int \omega(x)p(x)dx \\
&= \mathbb{E}(\omega(X)) = \theta.
\end{aligned}$$

Therefore, if we know $\pi(x)$, we can construct an IPW estimator

$$\widehat{\theta}_{\mathsf{IPW},0} = \frac{1}{n}\sum_{i=1}^{n} \frac{\omega(X_i)I(R_i = 1_d)}{\pi(X_i)}.$$

While this idea is universally true, it requires the knowledge of $\pi(x)$, which is generally unavailable (except for survey sample that we may know this). What's worst, the quantity $\pi(x)$ is generally non-identifiable so we cannot estimate it without missing data assumptions.

Therefore, a common strategy is make missing data assumption so that $\pi(x)$ becomes identifiable and we then place a model to estimate it.

Example: CCMV. A simple assumption is the CCMV (complete-case missing value) assumption, which requires the following assumption

$$\text{(CCMV-S)} \qquad \frac{P(R=r|X)}{P(R=1_d|X)} = \frac{P(R=r|X_r)}{P(R=1_d|X_r)}. \tag{8}$$

The left-hand-side of equation (8) is an identifiable quantity and we denote

$$O_r(x_r) = \frac{P(R=r|X_r=x_r)}{P(R=1_d|X_r=x_r)}.$$

This quantity is called *observable odds* in

Chen, Y. C. (2022). Pattern graphs: a graphical approach to nonmonotone missing data. The Annals of Statistics, 50(1), 129-146.

and it can be estimated by a binary classification model with two classes $R = 1_d$ versus $R = r$ with variables X_r.

One can easily verify that under equation (8), the propensity score

$$\pi(x) = \frac{1}{\sum_r O_r(x_r)}.$$

Note that if model

$$\log O_r(x_r; \beta_r) = \beta_r^T x_r,$$

this is essentially applying a logistic regression on $R = r$ versus $R = 1_d$ problem. Suppose that we have estimators $\widehat{O}_r(x_r)$ for each r. The IPW estimator will be

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\omega(X_i) I(R_i = 1_d)}{\sum_r \widehat{O}_r(X_{i,r})}.$$

See

Tchetgen, E. J. T., Wang, L., & Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statistica Sinica, 28(4), 2069-2088.

for more discussions on this method and the CCMV assumption.

5.2 Pattern mixture model, regression adjustment, and Imputation

Alternatively, we may use the PMM to deal with non-monotone missing data problem. Recall that the PMM uses the following decomposition

$$p(x, r) = p(x_{\bar{r}}|x_r, r)p(x_r, r).$$

Then we have the following decomposition:

$$\begin{aligned}
\theta &= \mathbb{E}(\omega(X)) \\
&= \sum_r \int \omega(x) p(x, r) \\
&= \sum_r \int \omega(x) p(x_{\bar{r}}|x_r, r) p(x_r, r) \\
&= \sum_r \mathbb{E}(\underbrace{\mathbb{E}(\omega(X)|X_r, R = r)}_{m_r(X_r)} I(R = r)) \\
&= \sum_r \mathbb{E}(m_r(X_r) I(R = r)).
\end{aligned}$$

Note that $m_r(X_r)I(R=r)$ is identifiable. As a result, when $m_r(x_r)$ is known, we can construct an estimator

$$\begin{aligned}\widehat{\theta}_{\mathsf{RA},0} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{r} m_r(X_{i,r})I(R_i=r)\\ &= \frac{1}{n}\sum_{i=1}^{n} m_{R_i}(X_{i,R_i}).\end{aligned}$$

This is like the regression adjustment method introduced at the beginning.

However, in general $m_r(x_r)$ is unknown and not identifiable. So we have to make a missing data assumption (and possibly also a model assumption) to identify $m_r(x_r)$.

From the PMM formulation, we know that the major issue comes from the extrapolation density

$$p(x_{\bar{r}}|x_r,r)$$

is not identifiable. So a common strategy is to equate this to something that is identifiable.

Connection to imputation. Here is an interesting fact. Suppose that we have a model of the extrapolation density, say

$$p(x_{\bar{r}}|x_r,r) = q(x_{\bar{r}}|x_r,r)$$

for some function q. Although we want to use

$$m_r(x_r) = \int \omega(x)q(x_{\bar{r}}|x_r,r)dx_{\bar{r}},$$

this integral may not have a closed form, making it hard to use the regression adjustment method. However, suppose that we can *sample* from the distribution q. Then we can perform a Monte Carlo approximation to $m_r(x_r)$. Specifically, for observation X_{i,R_i}, R_i with $R_i = r$, we approximate

$$m_r(x_r) \approx \widetilde{m}_r(x_r) = \frac{1}{N}\sum_{\ell=1}^{N}\omega(X_{i,\bar{r}}^{*(\ell)}, X_{i,r}),$$

where

$$X_{i,\bar{r}}^{*(1)},\cdots,X_{i,\bar{r}}^{*(N)} \sim q(x_{\bar{r}}|X_{i,r},r)$$

are generated from q. By applying this to every observation, you will find that the resulting estimator of θ is essentially the *multiple imputation estimator!*

Therefore, any multiple imputation estimator can be viewed as a PMM with an implicit model on the extrapolation density. The extrapolation density is equivalent to the imputation distribution.

Example: CCMV. A formal statistical approach to use the PMM is via making a statistical assumption to identify the extrapolation density $p(x_{\bar{r}}|x_r,r)$. Here we consider the CCMV assumption. In the PMM, the CCMV means

$$\text{(CCMV-P)} \qquad p(x_{\bar{r}}|x_r,r) = p(x_{\bar{r}}|x_r,1_d). \tag{9}$$

Namely, the extrapolation density is the same conditional density based on the complete case. You can prove that equation (9) is equivalent to (8). In this case, we can easily estimate $p(x_{\bar{r}}|x_r,1_d)$ using the complete

data. For instance, we may assume that $X|R=1_d \sim N(\mu,\Sigma)$ and estimate μ,Σ by the complete data. Then the estimator

$$\widehat{p}(x_{\bar{r}}|x_r,r)=\widehat{p}(x_{\bar{r}}|x_r,1_d)\sim N\left(\widehat{\mu}_r(x_r),\widehat{\Sigma}_r(x_r)\right)$$

is just the implied conditional normal distribution. With an estimator $\widehat{p}(x_{\bar{r}}|x_r,r)$, we obtain an estimator $\widehat{m}_r(x_r)$, which then leads to the final estimate

$$\widehat{\theta}_{\mathsf{RA}}=\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{R_i}(X_{i,R_i}).$$

5.3 Multiply-robust estimator

You may be wondering if we could construct an estimator similar to the doubly-robust estimator as the simple case. It turns out that it is possible to do so but it depends on the missing data assumption we are making.

In the case of CCMV, you can show that the following equality

$$\begin{aligned}\theta&=\mathbb{E}\left(\sum_r[\omega(X)-m_r(X_r)]\frac{I(R=r)}{O_r(X_r)}+m_r(X_r)I(R=r)\right)\\&=\mathbb{E}\left(\sum_r\omega(X)\frac{I(R=r)}{O_r(X_r)}+m_r(X_r)\left[I(R=r)-\frac{1}{O_r(X_r)}I(R=1_d)\right]\right),\end{aligned}$$

which implies a multiply-robust estimator (i.e., for every pair (m_r,O_r), we need one of the two models to be correct). See the following paper for more discussion:

> Tchetgen, E. J. T., Wang, L., & Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statistica Sinica, 28(4), 2069-2088.

While the name multiply-robustness sounds very powerful, it is actually a weaker result than the doubly-robustness because we need *every pair* of models to have at least one model being correct.

5.4 Pattern graphs

Pattern graphs is a special graph-like object introduced in the following paper:

> Chen, Y. C. (2022). Pattern graphs: a graphical approach to nonmonotone missing data. The Annals of Statistics, 50(1), 129-146.

Patten graphs are directed graphs of *response vectors*. Namely, it is a graph of binary vectors $r\in\{0,1\}^d$.

A pattern graph G is called *regular* if it satisfies the following two conditions:

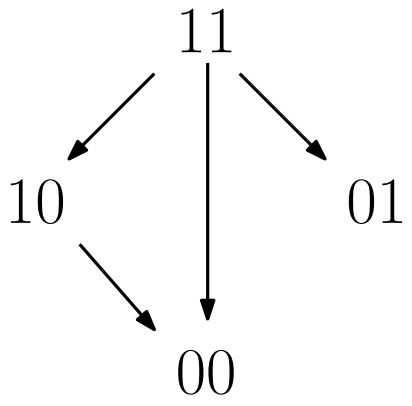

Figure 1: An example of a pattern graph of two variables.

- **(PG1)** If we see an arrow $s \to r$, then $s > r$, i.e., $s_j \geq r_j$ for all $j = 1, \cdots, d$ and at least one strict inequality holds.
- **(PG2)** The only source (vertex without parents) is the pattern $1_d = (1, 1, 1, \cdots, 1)$.

Clearly, *pattern graphs are NOT conventional graphical models*.

Pattern graph can be viewed as a generalization of the CCMV in the sense that we change equation (8) into

$$\text{(PG-S)} \qquad \frac{P(R = r|X)}{P(R \in \mathsf{PA}(r)|X)} = \frac{P(R = r|X_r)}{P(R \in \mathsf{PA}(r)|X_r)}, \tag{10}$$

where $\mathsf{PA}(r)$ is the parents of r. Similar to the CCMV, we can estimate the odds in the right-hand-side by a binary classification method. You can derive an IPW estimator based on equation (10).

For the pattern mixture model, the pattern graph revise the equation (9) as

$$\text{(PG-P)} \qquad p(x_{\bar{r}}|x_r, r) = p(x_{\bar{r}}|x_r, \mathsf{PA}(r)). \tag{11}$$

Namely, the pattern graph require the imputation model of pattern r is the same set of variables using its parents. Equation (11) implies a regression adjustment method as well as an imputation model for every pattern r.

Note that you can show that equations (10) and (11) are equivalent under a very mild condition.

Example. Suppose we have a pattern graph in Figure 1. By equation (11), this pattern graph implies the following imputation models:

$$\begin{aligned}
&p(x_1|x_2, R = 01) = p(x_1|x_2, R = 11) \\
&p(x_2|x_1, R = 10) = p(x_2|x_1, R = 10) \\
&p(x_1, x_2|R = 00) = p(x_1, x_2|R \in \{11, 10\}) = \frac{p(x_1, x_2, R = 11) + p(x_1, x_2, R = 10)}{P(R \in \{11, 10\})} \\
&\qquad = p(x_1, x_2|R = 00) \underbrace{\frac{P(R = 11)}{P(R \in \{11, 10\})}}_{=\rho} + p(x_2|x_1, R = 11) p(x_1|R = 10) \underbrace{\frac{P(R = 10)}{P(R \in \{11, 10\})}}_{=1-\rho}.
\end{aligned}$$

Namely, for $R = 01$, we impute X_1 by sampling from $p(x_1|x_2, R = 11)$, a model based on the complete-case. For $R = 10$, we impute X_2 by sampling from $p(x_2|x_1, R = 10)$. The case of $R = 00$ is more interesting. With a probability ρ, we sample both (X_1, X_2) from the complete-case distribution $p(x_1, x_2|R = 11)$. With a probability $1 - \rho$, we first sample X_1 from $p(x_1|R = 10)$, the marginal distribution of X_1 when only X_1 is observed, and then sample X_2 given on the previously imputed X_1 and the distribution $p(x_2|x_1, R = 11)$.

If we use the selection model in equation (10), we have a total of three observable odds:

$$\begin{aligned} O_{10}(x_1) &= \frac{P(R = 10|x_1)}{P(R = 11|x_1)} \equiv \frac{P(R = 10|x_1, x_2)}{P(R = 11|x_1, x_2)}, \\ O_{01}(x_2) &= \frac{P(R = 01|x_2)}{P(R = 11|x_2)} \equiv \frac{P(R = 01|x_1, x_2)}{P(R = 11|x_1, x_2)}, \\ O_{00} &= \frac{P(R = 00)}{P(R \in \{10, 11\})} \equiv \frac{P(R = 00|x_1, x_2)}{P(R \in \{10, 11\}|x_1, x_2)}. \end{aligned}$$

Recall that the goal is to identify the propensity score $P(R = 11|x_1, x_2) = \pi(x_1, x_2)$ so that we can implement the IPW estimator. Thus,

$$\begin{aligned} &P(R = 10|x_1, x_2) = O_{10}(x_1)\pi(x_1, x_2), \\ &P(R = 01|x_1, x_2) = O_{01}(x_2)\pi(x_1, x_2), \\ &P(R = 00|x_1, x_2) = O_{00} \cdot (\pi(x_1, x_2) + P(R = 10|x_1, x_2)) = O_{00} \cdot (1 + O_{10}(x_1))\pi(x_1, x_2). \end{aligned}$$

Using the fact that $\sum_r P(R = r|x_1, x_2) = 1$, we obtain the following equation

$$\begin{aligned} 1 &= [1 + O_{10}(x_1) + O_{01}(x_2) + O_{00}(1 + O_{01}(x_2))]\pi(x_1, x_2), \\ \pi(x_1, x_2) &= \frac{1}{1 + O_{10}(x_1) + O_{01}(x_2) + O_{00} + O_{00} \cdot O_{10}(x_1)}. \end{aligned}$$

Note that each term in the above quantity can be interpreted as a *path-specified probability* in the pattern graph (Figure 1):

$$\begin{aligned} 11 \to 10: &\quad O_{10}(x_1) \\ 11 \to 01: &\quad O_{01}(x_2) \\ 11 \to 00: &\quad O_{00} \\ 11 \to 10 \to 00: &\quad O_{00} \cdot O_{10}(x_1) \end{aligned}$$

5.5 Graphical model approach

An alternative approach to handling the non-monotone missing data is based on graphical model. In particular, the graphical model for the random vector $(X, R) \in \mathbb{R}^d \times \{0, 1\}^d$ is a directed acyclic graph (DAG) of the random vector where the arrows indicate the decomposition of the joint distribution and may be interpreted as a causal relation. Because of the DAG, the graphical model approach can be interpreted easily (via causal interpretation).

In the graphical model approach, we often need to make the following assumption:

- **(NSC)** For each variable j, no arrow between X_j and R_j.

The assumption (NSC) is called *no self-censoring* assumption in

> 1. Nabi, R., Bhattacharya, R., & Shpitser, I. (2020, November). Full law identification in graphical models of missing data: Completeness results. In International conference on machine learning (pp. 7153-7163). PMLR.
> 2. Malinsky, D., Shpitser, I., & Tchetgen Tchetgen, E. J. (2022). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. Journal of the American Statistical Association, 117(539), 1415-1423.

The (NSC) is equivalent to the follow condition

$$\text{(ICIN) } X_j \perp R_j | X_{-j}, R_{-j},$$

which is known as ICIN (itemwise conditionally independent nonresponse) in

> Sadinle, M., & Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. Biometrika, 104(1), 207-220.

Under either (NSC) or (ICIN), we can identify the full data distribution $p(x,r)$.

5.6 MICE: multiple imputation by chained equations

The MICE (multiple imputation by chained equations) is a popular approach that practitioners love to use. However, it has some problems (we will discuss them soon), you have to use it with caution.

The idea of MICE is very simple. We specify a leave-one-out conditional model every variable:

$$p(x_j | x_{-j}; \lambda_j) \tag{12}$$

for every $j = 1, \cdots, d$, where x_{-j} is all variables except for x_j and λ_j is the underlying parameter.

Equation (12) indicates an imputation model for variable X_j given all other variables. Starting with an initial guess for every missing values and the model's parameters, the MICE algorithm then update each imputed value sequentially by equation (12). After updating all missing values, we then update the parameters λ (can be either via MLE or a Bayesian approach) and repeat the whole process again. The above procedure forms a Markov chain. After running the above procedure many times, we can take the last M batches of data and use them as our multiple imputation data.

The model in equation (12) is called the fully conditional specification–we fully specify every conditional model. These models are often easy to specify since they are density of a single variable and can be easily interpreted (how other variables contribute to a single variable). Because of these features an a well-developed algorithm, the MICE algorithm is very popular.

THE MICE ALGORITHM.

1. Input: (X_{i,R_i}, R_i) for $i = 1, \cdots, n$.
2. Initialize: $X_{i,\bar{R}_i}^{(0)}$ and $\widehat{\lambda}^{(0)}$.
3. For each iteration $t = 1, 2, \cdots$, do the following:
   (a) (Imputation) For $i = 1, \cdots, n$ do the following:
      i. Let $X_i^{(t-1)} = (X_{i,R_i}, X_{i,\bar{R}_i}^{(t-1)})$ be the imputed value from previous iteration.
      ii. For $j = 1, \cdots, p$, do the following:
         A. If $R_{ij} = 0$, sample $X_{i,j}^{(t)}$ from $p(x_j | X_{i,-j}^{(t-1)}; \widehat{\lambda}_j^{(t-1)})$.
         B. If $R_{ij} = 1$, skip.
   (b) (Parameter update) Update $\widehat{\lambda}^{(t)}$ by either using the MLE of the imputed data $X_1^{(t)}, \cdots, X_n^{(t)}$ or sample from $p(\lambda | X_1^{(t)}, \cdots, X_n^{(t)})$ (if using a Bayesian approach).

However, the MICE algorithm has a severe problem–*incompatibility*. The conditional distributions in equation (12) may not be compatible with each other. To see this, we consider $d = 3$. The MICE will specify models

$$p(x_1 | x_2, x_3; \lambda_1), p(x_2 | x_1, x_3; \lambda_2), p(x_3 | x_1, x_2; \lambda_3).$$

There may be no joint distribution $p(x_1, x_2, x_3)$ whose conditional distributions agree with all three of them! This problem gets even more severe when the number of variables is large.