STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 13: Missing data

Instructor: Yen-Chi Chen

Please read Section 13.1 on your own; it is from an undergraduate course (STAT 403).

13.1 Introduction: a simple case

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.

There are three common strategies that practitioners are using to handle missing data:

- **Complete-case analysis (ignoring observations with missing entries).** The complete-case analysis removes any observations that contain one or more missing entries. When the proportion of missing is small (and the missingness is irrelevant to any variables, including the one that can be missing), this is an okay procedure. But in general, this would lead to a biased estimate.

  To see this, think about estimating the average income of a city from a social survey. Many rich people would refuse to provide their incomes (it will be easy to identify them), leading to missing entries. In this scenario, if we ignore those individuals whose income is missing, we will get a biased estimate of the average income.

  In an observation study in medical research, sometimes people would perform analysis by adjusting the *inclusion criteria*: the criteria that determines which individual will be included in our analysis. In case that they require individuals to be fully observed, this is essentially a complete-case analysis.

- **Ignorable missingness (missing at random).** Another common approach is to make assumptions and choose a good model so that the missingness is ignorable. Note that in this case, we do NOT remove observations with missing entries–we still use their observed variables to construct our model. This is possible when we assume the missingness is missing at random (MAR; see Section 13.2) and use a proper parametric model.

  However, MAR is just an assumption. It may be violated (which is often called missing not at random-MNAR). When the MAR is violated, it is often hard to obtain an ignorable missingness approach to deal with missing data. Note that sometimes we are still able to construct an ignorable missingness procedure using the selection model and inverse probability weighting estimator (see Section 13.3.1).

- **Imputation.** The imputation is another popular approach that practitioners used in solving missing data problem. The idea is very simple: we impute the missing entries with a proper value that leads to a complete dataset. Then we can treat the problem as if there is no missingness.

  Here is a caveat. If the imputation is done in a deterministic way, i.e., every time a missing entry is imputed, it always be imputed with a fixed number, the imputed data is often problematic because we do not take into account the intrinsic variation of that missing value. This would lead to bias in the later estimation procedure.

  A better approach is to use a stochastic imputation that we impute the missing entries by drawing from a distribution. Later we will show that if the distribution being drawn is the actual distribution that generates the data, the stochastic imputation leads to a dataset without any bias (Section 13.1.2).

A challenge here is that in general, we do not know the actual distribution so how do we perform the stochastic imputation is a problem.

13.1.1 A simple case

Consider a regression problem where we have a binary covariate $X \in \{0, 1\}$ and a continuous response $Y \in \mathbb{R}$. However, in our data, some response variables are missing and only the covariates are observed. So our data can be represented as

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \star), \cdots, (X_{n+m}, \star).$$

The symbol $\star$ denotes a missing value. Namely, we have n observations that are fully observed while the other m observations that we only observe the covariate, not the response. Suppose that the parameter of interest is the marginal median of the response variable m_Y. How should we estimate the median?

We can introduce an additional variable R to denote the missingness such that $R = 0$ means that Y is not observed whereas $R = 1$ means that Y is observed. Note that R itself is another random variable.

Without any assumptions on the missing data, we are not able to accurately estimate the median consistently. There are two common assumptions people made about the missingness:

1. **MCAR:** missing completely at random. This means that the missingness is independent of any variables. Under the above notations, MCAR means that

   $$R \perp X, Y.$$

2. **MAR:** missing at random. Under MAR, the missingness depends only on the observed pattern. In our case.

   $$P(R = 0|X, Y) = P(R = 0|X)$$

   since Y is not observed when $R = 0$.

When the missingness is neither MCAR nor MAR, it is called MNAR–missing completely at random.

Under MCAR, we can completely ignore the data with missing values and just use the sample median as an estimate of m_Y. However, under MAR, we cannot do such thing because the missingness may depends on X and if the distribution of covariate is different under fully observed data ($R = 1$) and partially observed data ($R = 0$), we will obtain a biased estimate.

While there are other ways to estimate the median under MAR, we will focus on the method of imputation.

13.1.2 Imputation

The idea of imputation is to impute a value to the missing entry so that after imputing all missing entries, we obtain a data without any missingness. Then we can simply apply a regular estimator (in the above example, sample median) to estimate the parameter of interest.

However, we cannot impute any number to the missing entry because this would cause bias in the estimation. We need to impute the value in a smart way. Generally, we want to impute the value according to the conditional density

$$p(y|x, R = 0),$$

the conditional density of response variable Y given the covariate X and the missing pattern $R = 0$. Namely, for $n + i$-th observation where only X_{n+i} is observed, we want to draw a random number

$$\tilde{Y}_{n+i} \sim p(y|X_{n+i}, R = 0).$$

If indeed Y_{n+1} is from the above density function, one can show that the sample median

$$\text{median}\{Y_1, \cdots, Y_n, \tilde{Y}_{n+1}, \cdots, \tilde{Y}_{n+m}\}$$

is an unbiased estimator of m_Y.

This idea works regardless of what missing assumption is. However, the problem is that the density function $p(y|x, R = 0)$ cannot be estimated using our data because the only case we observed Y is when $R = 1$.

Under this case, MAR implies a powerful result:

$$p(y|x, R = 0) = p(y|x, R = 1). \tag{13.1}$$

Namely, the conditional density of Y given X is independent of the missing indicator R. To see how equation (13.1) is derived, note that MAR implies

$$P(R = 1|X, Y) = 1 - P(R = 0|X, Y) = 1 - P(R = 0|X) = P(R = 1|X).$$

Thus, the conditional density

$$\begin{aligned}
p(y|x, R = 0) &= \frac{p(y, x, R = 0)}{P(x, R = 0)} \\
&= \frac{p(x, y)P(R = 0|x, y)}{P(x, R = 0)} \\
&= p(x, y)\frac{P(R = 0|x)}{P(x, R = 0)} \\
&= p(x, y)\frac{1}{p(x)} \\
&= p(x, y)\frac{P(R = 1|x)}{P(x, R = 1)} \\
&= \frac{p(x, y)P(R = 1|x, y)}{P(x, R = 1)} \\
&= \frac{p(y, x, R = 1)}{P(x, R = 1)} \\
&= p(y|x, R = 1).
\end{aligned}$$

Thus, we obtain equation (13.1).

The right-hand side in (13.1), $p(y|x, R = 1)$, can be estimated from the data (in this case, we say it is identifiable). Let $\widehat{p}(y|x, R = 1)$ be an estimator of the conditional PDF. Given X_{n+i}, we then impute $\tilde{Y}_{n+i}$ by sampling from $\widehat{p}(y|X_{n+i}, R = 1)$. Here we demonstrate this with two examples.

Example: Gaussian model. We assume that $p(y|x, R = 1)$ follows a Normal distribution $N(\mu(x), \sigma^2(x))$. We can easily estimate $\mu(x)$ and $\sigma^2(x)$ by the method of moments (or MLE):

$$\widehat{\mu}(x) = \frac{1}{n_x}\sum_{i=1}^{n} Y_i I(X_i = x, R_i = 1), \qquad \widehat{\sigma}^2(x) = \frac{1}{n_x}\sum_{i=1}^{n}(Y_i - \widehat{\mu}(x))^2 I(X_i = x, R_i = 1),$$

where $n_x = \sum_{i=1}^n I(X_i = x, R_i = 1)$ is the number of $X_i = x$ in the completely observed data ($R_i = 1$) and $x \in \{0, 1\}$. For an observation X_{n+i} with a missing response, we impute $\tilde{Y}_{n+i}$ by

$$\tilde{Y}_{n+i} \sim N(\widehat{\mu}(X_{n+i}), \widehat{\sigma}^2(X_{n+i})).$$

Example: kernel density estimator. In this simple case, we may estimate $p(y|x, R = 1)$ via a KDE:

$$\begin{aligned}\widehat{p}(y|x, R = 1) &= \frac{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right) I(X_i = x)}{\frac{1}{n}\sum_{i=1}^n I(X_i = x)} \\ &= \frac{1}{n_x h}\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right) I(X_i = x).\end{aligned}$$

Namely, $\widehat{p}(y|x, R = 1)$ is the KDE applied to the completely observed data with the covariate $X = x$. Given an observation $X_{n+i} = x$, how should we sample $\widehat{Y}_{n+i}$ from $\widehat{p}(y|x, R = 1)$? It is very simple. We first sample the index I such that

$$P(I = i|\mathsf{data}) = \frac{1}{n_x} I(X_i = x).$$

Namely, I is chosen from those fully observed data with the covariate $X_i = x$ with equal probability. Given I we then sample Y_{n+i} from the density function

$$q(y) = \frac{1}{h} K\left(\frac{Y_I - y}{h}\right).$$

Although this may look scary, if the kernel function is Gaussian, $q(y)$ is the normal density with mean Y_I and variance h^2. Namely, when K is a Gaussian,

$$\widehat{Y}_{n+i} \sim N(Y_I, h^2).$$

Remark.

- The equation (13.1) relies on the MAR assumption along with the fact that only one variable is subject to missing. When there are more than one variables that can be missing, we no longer have such a simple equivalence.
- The imputed data can be used for other estimators as well, not limited to estimating the median. You may notice that during our imputation process, we do not use any information about the estimator.
- There are imputation methods that only imputes a fixed, non-random number for each missing entries. This is often called a deterministic imputation. For certain problem, a deterministic imputation works but in general, it may not work. So a rule thumb is to use a random imputation if possible.

13.1.3 Multiple imputation

After doing the imputation for all missing entries, we obtain a complete data

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}_{n+1}).$$

The estimate of m_Y is just the sample median of this impute dataset. However, there will be Monte Carlo errors in this estimator because every time we do the imputation, we will not get the same number (due to

sampling from $p(y|x, R = 0)$). If we just impute the data once (this is often called *single imputation*), we may suffer from the Monte Carlo errors a lot. Thus, a better approach is to perform a *multiple imputation*.

Multiple Imputation.[1] After obtaining a complete data, we do the same imputation procedure again, which gives us another new complete data. Then we keep repeating the above process, leading to several complete data, which can be represented as

$$\begin{aligned}
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}_{n+1}^{(1)}), \cdots, (X_{n+m}, \widehat{Y}_{n+1}^{(1)}) \\
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}_{n+1}^{(2)}), \cdots, (X_{n+m}, \widehat{Y}_{n+1}^{(2)}) \\
&\cdots \\
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}_{n+1}^{(N)}), \cdots, (X_{n+m}, \widehat{Y}_{n+1}^{(N)}).
\end{aligned}$$

We then combine all these datasets to a huge dataset and compute the estimator of the parameter of interest (in our case, median of the response variable). This estimator has a smaller Monte Carlo error.

13.2 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 $X_1, \cdots, X_5$ 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 $X_{11001} = (X_1, X_2, X_5)$. Under this notation, the MAR assumption can be written as

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

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

MAR is a very popular assumption that people often assumed in practice (although it may not be reasonable in some cases). However, under the non-monotone case, MAR tells us little about the missingness and it is actually not very to work with. 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. The second reason is that under monotone missing data problem, MAR provides an elegant way to identify the entire distribution function.

[1] For more introduction on this topic, see https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

13.2.1 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 decouple 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 simple procedure.

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

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

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.

13.2.2 MAR under monotone case

Under the monotone missing problem, let T denotes the index of the last observed variable. Namely, the individual dropouts after time point T. We use the notation $X_{\leq t} = (X_1, \cdots, X_t)$. 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 $t = 1, \cdots, d$!

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.

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

13.3 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.

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.

13.3.1 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

$$\widehat{\theta}_0 = \int \omega(x)\frac{d\widehat{F}(dx,T=d)}{\widehat{P}(T=d|x)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\omega(X_i)I(T_i=d)}{\widehat{P}(T=d|X_i)}. \tag{13.2}$$

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.

13.3.2 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 $\beta_{t,s}, \sigma^2_{t,s}$ can be estimated by

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

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 $N(\widehat{\beta}^T_{t,s} x_{<s}, \widehat{\sigma}^2_{t,s})$.

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.

13.3.3 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

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

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 $\tilde{X}_{i,\leq t} = X_{i,\leq t}$.
2. For $s = t+1, t+2, \cdots, d$ do the following:
   (a) Draw $\tilde{X}_{i,s}$ from $N(\widehat{\beta}^T_{t,s} \tilde{X}_{i,<s}, \widehat{\sigma}^2_{t,s})$.
3. Return the imputed vector $\tilde{X}_i = (\tilde{X}_1, \cdots, \tilde{X}_d)$.

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 $T = t$.

13.3.4 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.

13.3.5 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.

13.4 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

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

13.4.1 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

$$\text{(CCMV-S)} \qquad \frac{P(R = r|X)}{P(R = 1_d|X)} = \frac{P(R = r|X_r)}{P(R = 1_d|X_r)}. \tag{13.3}$$

The left-hand-side of equation (13.3) 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 (13.3), 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.

13.4.2 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

$$\text{(CCMV-P)} \qquad p(x_{\bar{r}}|x_r, r) = p(x_{\bar{r}}|x_r, 1_d). \tag{13.4}$$

Namely, the extrapolation density is the same conditional density based on the complete case. You can prove that equation (13.4) is equivalent to (13.3). 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

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

13.4.3 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.

13.4.4 Pattern graphs

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

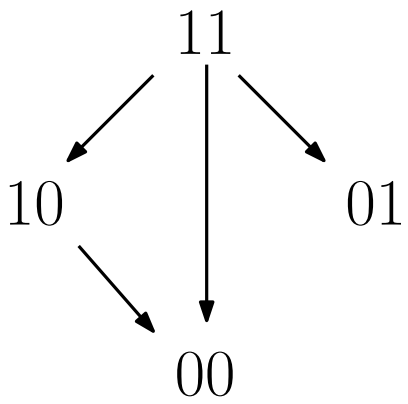

Figure 13.1: An example of a pattern graph of two variables.

[C2022] 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:

- **(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 (13.3) into

$$\text{(PG-S)} \qquad \frac{P(R = r|X)}{P(R \in \mathsf{PA}(r)|X)} = \frac{P(R = r|X_r)}{P(R \in \mathsf{PA}(r)|X_r)}, \tag{13.5}$$

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 (13.5).

For the pattern mixture model, the pattern graph revise the equation (13.4) as

$$\text{(PG-P)} \qquad p(x_{\bar{r}}|x_r, r) = p(x_{\bar{r}}|x_r, \mathsf{PA}(r)). \tag{13.6}$$

Namely, the pattern graph require the imputation model of pattern r is the same set of variables using its parents. Equation (13.6) implies a regression adjustment method as well as an imputation model for every pattern r.

Note that you can show that equations (13.5) and (13.6) are equivalent under a very mild condition.

Example. Suppose we have a pattern graph in Figure 13.1. By equation (13.6), 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 (13.5), 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 13.1):

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

13.4.5 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 $p(x, r)$.

13.4.6 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:

$$p(x_j | x_{-j}; \lambda_j) \tag{13.7}$$

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

Equation (13.7) 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 (13.7). 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 (13.7) 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 (13.7) 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.