Lecture Notes VII – Missing Data and Imputation

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

The missing data problem

Imputation for MAR

Imputation for MAR with Monothone Missing data

Missing data: the problem

- Data sampled i.i.d. $\sim F$, but some (parts of each sample) are missing
- Here we denote data point $= (X^i, Y^i)$ where X^i is always observed, Y^i may be missing or not.

- A new random variable R, $R^i = 0$ when Y^i not observed, $R^i = 1$ when observed

- **What is the distribution of R?**

MCAR Missing Completely At Random $R \perp X, Y$

MAR Missing At Random $R \perp Y \mid X$

What is the distribution of R?

- Let data $\mathcal{D} = \{(x^1, y^1), \ldots (x^n, y^n)\} \cup \{x_{n+1}, \ldots x_{n+m}\}$
  - complete data $\{(x^1, y^1), \ldots (x^n, y^n)\}$ with $R^{1:n} = 1$
  - data with missing values $\{x_{n+1}, \ldots x_{n+m}\}$ with $R^{n+1:n+m} = 0$

MCAR Missing Completely At Random $R \perp (X, Y)$

- no bias if only observed Y values $\{y^1, \ldots y^n)\}$ used
- Can also use imputation as in MAR case

MAR Missing At Random $R \perp Y \,|\, X$

- **bias possible** if only observed Y values $\{y^1, \ldots y^n)\}$ are used because $R \not\perp Y$
- $\{(x^1, y^1, R^1), \ldots (x^n, y^n, R^n)\}$ together with $\{(x^{n+1}, R^{n+1}), \ldots (x^{n+m}, R^{n+m})\}$ has information about $\{y^{n+1}, \ldots y^{n+m}\}$
- Probability distribution of missing values $Y^{n+1:n+m}$ is

$$Y \sim p(Y \,|\, X, R = 0) \qquad \text{and} \qquad \boxed{p(Y \,|\, X, R = 0) = p(Y \,|\, X, R = 1)} \tag{1}$$

Imputation for MAR

$$\boxed{p(Y \mid X, R = 0) \;=\; p(Y \mid X, R = 1)}$$

- **Idea**
  1. Estimate $p(Y \mid X, R = 1)$ the distribution of $Y|X$ from the observed data
  2. "Guess" (Sample) $\tilde{y}^{n+1:n+m} \sim p(Y \mid X, R = 1)$
  3. Use $y^{1:n} \cup \tilde{y}^{n+1:n+m}$ to estimate $\hat{\theta}$

Estimate $p(Y \mid X, R = 1)$ from $\mathcal{D}^1 = \{(x^1, y^1, R^1), \ldots (x^n, y^n, R^n)\}$

Example $X \in \{1, \ldots K\}$ discrete variable

- For each $k = 1, \ldots K$, estimate $p_k(Y) \equiv p(Y \mid X = k)$

Estimation methods

- Kernel Density Estimation (KDE)

$$p_k(y) \;=\; \frac{1}{n_k h} \sum_{i : x^i = k} K_h(y^i - y) \tag{2}$$

n_k = number of $x^i = k$, K = kernel function, $K_h(z) = K(z/h)$, h = kernel width (can depend on k)

- Parametric distributions (e.g. Normal(μ_k, σ_k^2))

Sampling $\tilde{y}^i$ from KDE

Given $p_1, \ldots p_k$, $x^i = k$, $i > n$

1. Sample $i' \sim 1 : n_k$ uniformly
2. Sample $\tilde{y}^i \;\sim\; N(x^{i'}, h^2)$

Multiple Imputation

Given $p_1, \ldots p_k$, $x^i = k$, $i = n+1 : n+m$, $B \geq 1$ an integer

for $i = n+1 : n+m$, $b = 1 : B$

1. Sample $i' \sim 1 : n_k$ uniformly
2. Sample $\tilde{y}^{i,b} \sim N(x^{i'}, h^2)$

3. Use data $\underbrace{\{y^{1:n}\}}_{\times B} \cup \{\tilde{y}^{i,b}, \text{ for } i = n+1 : n+m,\ b = 1 : B\}$

Imputation for MAR with Monothone Missing data

- Example Clinical trials with drop-out Each patient i is observed for a $t = 1, 2, \ldots T^i$ with $T^i \leq p$. Hence $y^i = (y^i_1, \ldots y^i_{T^i})$.
- T can depend on the previous values $y^{1:T}$ but not on the future values $y^{T=1:p}$.

- T^i is the missingness variable
- $x^i_{1:t}$ is the "always observed" data
- $x^i_{t+1:p}$ is the missing data

Estimating $p(T|X)$

- Assume for simplicity $X_{1:p} \in \{1, \dots K\}$
- Dependence of missingness T on observed $X_{1:T}$ (MAR)
- (Recursively for $t = 1, 2, \dots p-1$)
- Start with $t = 1$

$$P(T = 1|X) = P(T = 1|X_1) \tag{3}$$

$$\text{for each } k \qquad P(T = 1|X = k) = \frac{\#\{x^i \text{ with } t^i = 1, x_1^i = k\}}{\#\{x^i \text{ with } t^i \geq 1, x_1^i = k\}}$$

- For $t = 2$

$$P(T = 2|X) = P(T = 2|X_1, X_2) \tag{4}$$

$$\text{for each } k_1, k_2 \qquad P(T = 2|X_{1,2} = k_{1,2}) = \frac{\#\{x^i \text{ with } t^i = 2, x_{1,2}^i = k_{1,2}\}}{\#\{x^i \text{ with } t^i \geq 2, x_{1,2}^i = k_{1,2}\}}$$

...

- For t

$$P(T = t|X) = P(T = t|X_{1:t}) \tag{5}$$

$$\text{for each } k_{1:t} \qquad P(T = 2|X_{1:t} = k_{1:t}) = \frac{\#\{x^i \text{ with } t^i = t, x_{1:t}^i = k_{1:t}\}}{\#\{x^i \text{ with } t^i \geq t, x_{1:t}^i = k_{1:t}\}}$$