STAT 403

5/7/25

Lecture 15

Missing data &
imputation

Lecture Notes VII – Missing Data and Imputation

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

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 $R \perp Y \mid X$

MNAR Missing Not At Random ← nothing to do (OR no general recipe)

What is the distribution of R?

$R = 1$ if y obs
0 if y missing

- 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 $\boxed{R \perp (X, Y)}$ $\Rightarrow$ $R \perp Y \mid X$

- 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 $\boxed{R \perp Y \mid X}$

- **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

$$Y \sim p(Y \mid X, R = 0) \quad \text{and} \quad \boxed{p(Y \mid X, R = 0) = \underbrace{p(Y \mid X, R = 1)}_{\text{estimate from observed } (x^{1:n}, y^{1:n})}} \tag{1}$$

X = always observed $\rightarrow$ "helper"
Y = may be missing $\rightarrow$ $\theta(Y)$ of interest $\nearrow$ Bias $\searrow$ (Var)

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 $\hat{\theta}$

treat them as observed

typical $q(y|x)$

- $x, y \in \mathbb{R}$: regression
  $x \in \{0,1\}, y \in \mathbb{R}$
- density estimation
  $p(y \mid x=0)$
  $p(y \mid x=1)$
- $y \in \{0,1\}$ or discrete
  classification

2'. sample $\tilde{y}^{i,1:B} \sim \hat{p}(Y \mid X = x^i)$

B values for each $i = n+1 : n+m$

(Multiple imputation)

3'. use $\{\underbrace{y^{1:n}}_{\times B \text{ times}}\} \cup \{\tilde{y}^{n+1:n+m,\,1:B}\}$

	1.2	1.5	1.5
obs :	1.2,	1.5,	1.1
imputed :		1.6	1.1
		1.7	1.2

$\hat{\theta}$

$y^{1:n}$

$\tilde{y}^{4,1}$ $\tilde{y}^{5,1}$

$\tilde{y}^{4,2}$ $\tilde{y}^{5,2}$

Ex : $n=3, m=2, B=2$

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 , $Y \in \mathbb{R}$

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

1. Estimation methods

- Kernel Density Estimation (KDE) ← can estimate any $p(\,)$ ← is non-parametric

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

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

step.2 Sampling

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

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

KDE

$$\hat{p}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y^i}{h}\right)$$

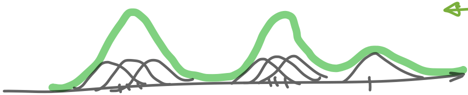
data

place a ∼ on each y^i

"$N(0,1)$"

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \text{kernel}$$

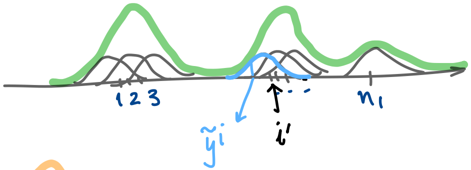
h = parameter (kernel width)



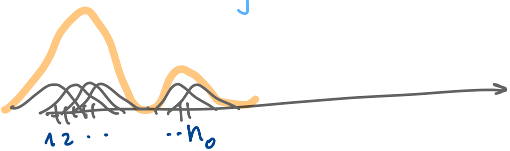
adapts to shape of data

for n large

$p(y|x=1)$



$p(y|x=0)$



Imputation for y^i

1. $x^i = ?$ $\quad x^i = 1 \Rightarrow$ sample from $p(y|x=1)$
2. sample $i' \sim \text{unif}(1, \dots n_1) \Rightarrow i' = 4$
3. sample $\tilde{y}^i \sim N(x^{i'}, h^2) \Rightarrow \tilde{y}^i$

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