Bayesian models for missing data

Made by
Wu jiajia
Jiang jinfeng
(The 4th group)

Missing data

EM algorithm

Missing data

1. Definition
2. Reasons
3. Types
4. Solutions

- **Definition:** No data value is stored for the variable in an observation**.**

- **Reasons for missing data**
  - ◆inaccessible
  - ◆omitted
  - ◆unavailable attributes
  - ◆high cost

Notations

$$X = (X_{obs} + X_{mis})$$

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} = [X_1, X_2, \ldots X_p]$$

$$M_{ij} = \begin{cases} 1 \; as \; X_{ij} \; is \; observed \\ 0 \; as \; X_{ij} \;\; is \;\; missing \end{cases}$$

Types of missing data

- **MCAR: missing completely at random**

  $p(M|X)=p(M)$

- **MAR: missing at random**

  $p(M|X)=p(M|X_{obs})$

- **MNAR: missing not at random**

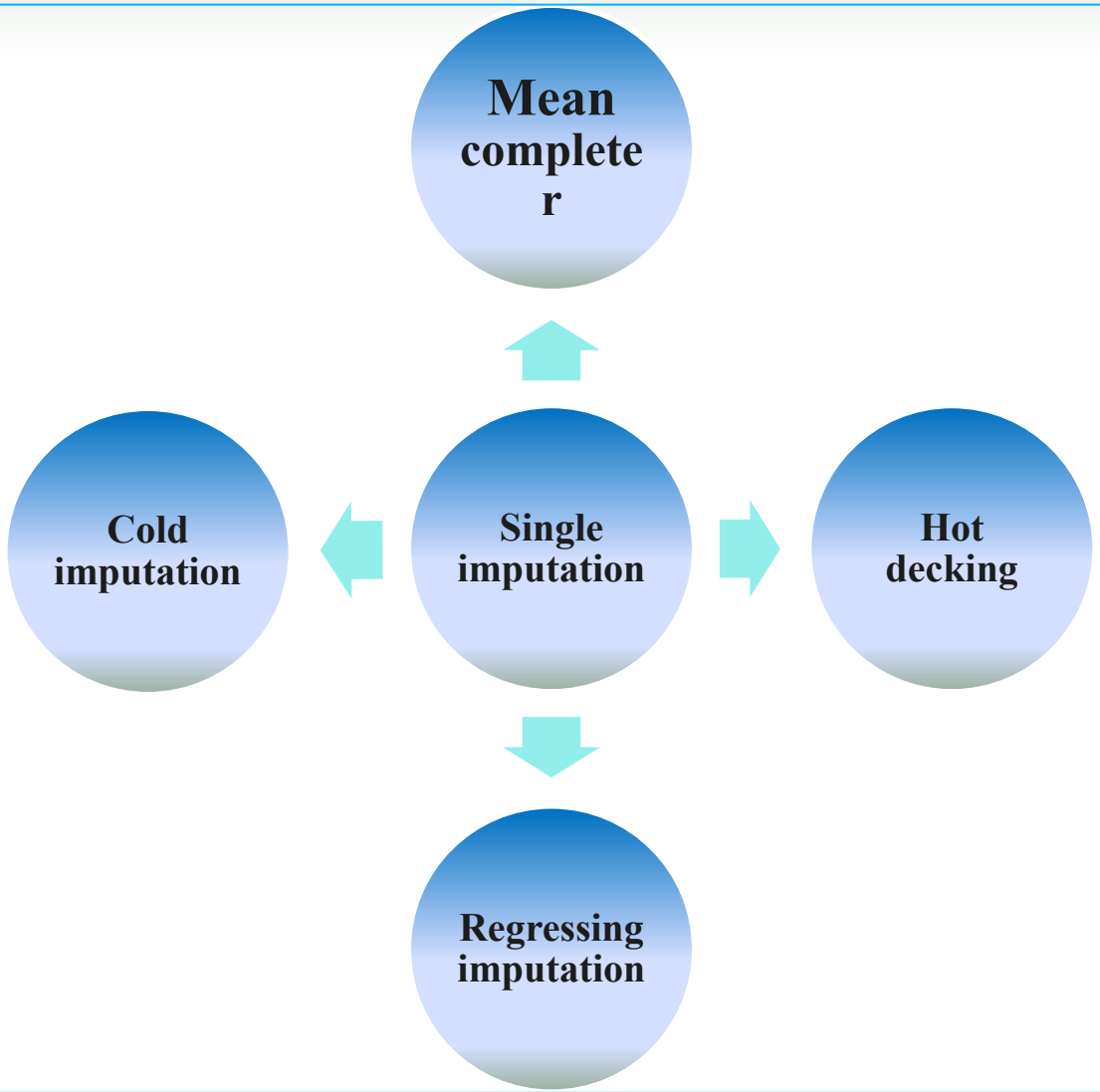
A simple exercise

1. A sudden traffic violation happened, some of those surveyed submitted questionnaires in a hurry with some missing responses. （**MCAR**）
2. People with bad jobs seldom answer their incomes. （**MAR**）
3. People with low incomes are less likely to report their incomes. （**MNAR**）

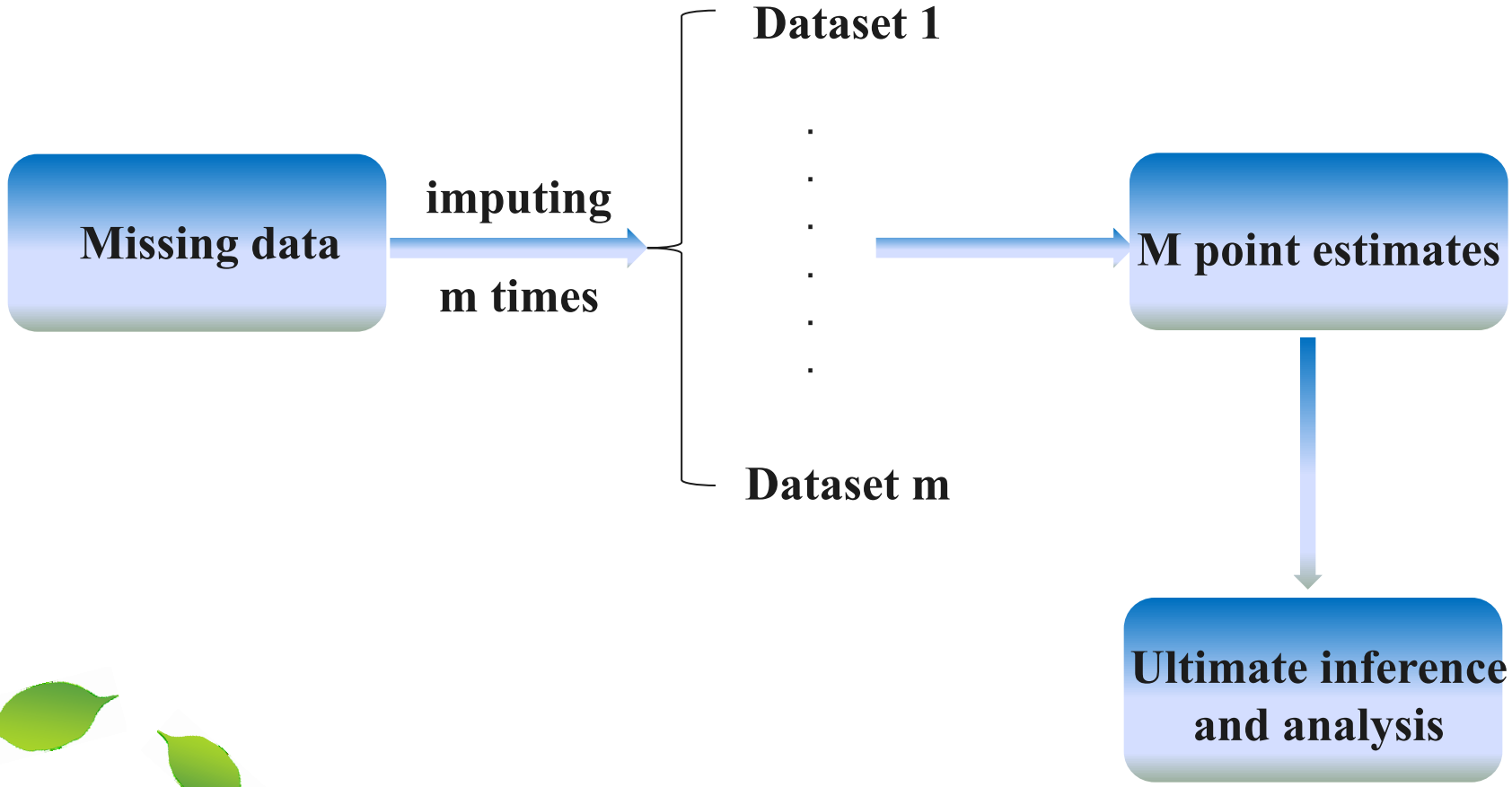
Solutions of missing data

- **Deletion(** listwise and pairwise **)**

	sex	age	income
1	0	24	5,000
2	1	22	3,000
3	1	---	6,000


Mean completer
Cold imputation
Single imputation
Hot decking
Regressing imputation

Multiple imputations



Inference with multiple imputations

$$\overline{\theta_{mi}} = \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m} \qquad \hat{v} = \frac{\sum_{i=1}^{m} v_i}{m}$$

$$\mathrm{B} = \frac{\sum_{i=1}^{m} \left(\hat{\theta}_i - \overline{\theta_{mi}}\right)^2}{m-1}$$

$$\mathrm{T_{mi}} = \hat{\mu} + \left(1 + \frac{1}{\mathrm{m}}\right)\mathrm{B}$$

x_i	4	6	8	9	11	13	16	18	20	25
y_i	4	20	10		14		16	15	18	22

Ratio estimate

$$\widehat{\bar{Y}_R} = \frac{\sum y_i}{\sum x_i}\bar{X} = \hat{R}\bar{X}$$

Step 1

x_i	4	6	8	9	11	13	16	18	20	25
$y_i^{(1)}$	4	20	10	10	14	14	16	15	18	22
$y_i^{(2)}$	4	20	10	16	14	14	16	15	18	22

Step 2

	$\widehat{\bar{Y}_R}$	v_i
1	**13.38**	**2.965**
2	**13.57**	**3.157**

Step 3

$$\widehat{\bar{Y}_R} = \frac{13.38+13.57}{2} = 13.48$$

$$v = \frac{1}{2}(2.965+3.157) = 3.061$$

$$B = \frac{(13.38-13.48)^2 + (13.57-13.48)^2}{2-1} = 0.0181$$

$$T = 3.061+(1+\frac{1}{2})*0.0181 = 3.0882$$

EM algorithm

1. Expectation-Maximization
2. Mixture Model Training
3. A Silly example

- **The EM algorithm formalizes an intuitive idea for obtaining parameter estimates when some of the data are missing:**

    1. replace missing values by estimated values,
    2. estimate parameters.
    3. Repeat

    Step (1) use estimated parameter values as true values, and

    step (2) use estimated values as “observed” values, iterating until convergence.

- **EM is a method to find θ_{ML} where**

$$\theta_{ML} = \arg\max_{\theta \in \Omega} L(\Theta)$$

$$= \arg\max_{\theta \in \Omega} \log P(X \mid \Theta) \Rightarrow \arg\max_{\theta \in \Omega} \log P(Z \mid \Theta)$$

- **Z = (X, Y)**
  - ◆Z: complete data ("augmented data")
  - ◆X: observed data ("incomplete" data)
  - ◆Y: hidden data ("missing" data)
  - ◆Θ: a parameter vector.

The inner loop of the EM algorithm

E-step: Compute expectation of
(θ,θ^t : old, new distribution parameters)

$$Q(\theta, \theta^t) = E_\theta\{\log p(z;\theta) \mid x\}$$

M-step: Find θ that maximizes Q

$$\theta^{t+1} = \arg\max Q(\theta, \theta^t), \text{for all } \theta$$

- In particular, when the distribution of the complete-data vector belongs to the exponential family,and the log-likelihood is linear in the sufficient statistic for θ,the E-step reduces to computing the expectation of the complete-data sufficient statistic givien the observed data.

Mixture Model Training

Let the complete-data vector y=$(y_1, \cdots y_n)^T$ be a random sample from N(u,σ^2).Suppose $y_i, i = 1, \ldots, m$ are observed and $i = m + 1, \ldots, n$ are missing.

$$f(y;u,\sigma^2)=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-u)^2}{\sigma^2}\right\}$$

$(\sum y_i, \sum y_i^2)$ are sufficient statistics for θ = (μ,σ^2)

At the t^{th} iteration

For the E-step,compute

- $E_\theta(\sum y_i|y_{obs})=\sum_{i=1}^{m} y_i+(n-m)u^{(t)}$
- $E_\theta(\sum {y_i}^2|y_{obs})=\sum_{i=1}^{m} {y_i}^2+(n-m)({u^{(t)}}^2-{\sigma^{(t)}}^2)$

For the M-step

- $\hat{u}=\frac{1}{n}\sum_{i=1}^{n} y_i$
- $\hat{\sigma}^2=(\frac{1}{n}\sum_{i=1}^{n} y_i)^2-\frac{1}{n}\sum_{i=1}^{n} {y_i}^2$

An example of EM

- **Let events be "grade in a class"**

Get's an A	P(A)=1/2	a=number 0f A's
Get's a B	P(B)=u	b=number 0f B's
Get's a C	P(C)=2u	c=number 0f C's
Get's a D	P(D)=1/2-3u	d=number 0f D's

Suppose we know the

number 0f (A's+B's) =h

number 0f C's=c

number 0f D's=d

Same Problem with Hidden Information

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

REMEMBER

P(A) = ½

P(B) = μ

P(C) = 2μ

P(D) = ½-3μ

What is the max. like estimate of μ now?

We can answer this question circularly:

EXPECTATION

If we know the value of μ we could compute the expected value of a and b

Since the ratio a:b should be the same as the ratio ½ : μ

$$a = \frac{1/2}{1/2 + \mu} h \qquad b = \frac{\mu}{1/2 + \mu} h$$

MAXIMIZATION

If we know the expected values of a and b we could compute the maximum likelihood value of μ

$$\mu = \frac{b + c}{6(b + c + d)}$$

E.M. for our Trivial Problem

REMEMBER
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

We begin with a guess for μ

We iterate between EXPECTATION and MAXIMALIZATION to improve our estimates of μ and *a* and *b*.

Define μ(t) the estimate of μ on the t'th iteration
b(t) the estimate of *b* on t'th iteration

$$\mu(0) = \text{initial guess}$$

$$b(t) = \frac{\mu(t)h}{1/2 + \mu(t)} = \mathrm{E}[b \mid \mu(t)]$$

E-step

$$\mu(t+1) = \frac{b(t)+c}{6(b(t)+c+d)}$$

$$= \max \text{ like est of } \mu \text{ given } b(t)$$

M-step

Continue iterating until converged.
Good news: Converging to local optimum is assured.
Bad news: I said "local" optimum.

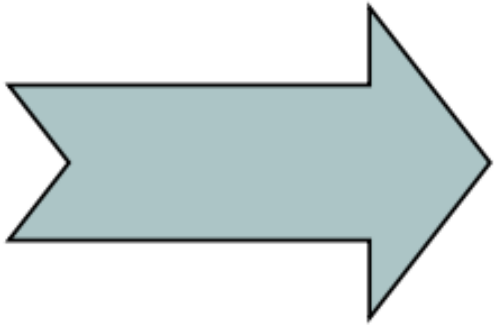
E.M. Convergence

- Convergence proof based on fact that Prob(data | μ) must increase or remain same between each iteration [NOT OBVIOUS]
- But it can never exceed 1 [OBVIOUS]

So it must therefore converge [OBVIOUS]

In our example, suppose we had

h = 20
c = 10
d = 10
μ(0) = 0



t	μ(t)	b(t)
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

Convergence is generally linear: error decreases by a constant factor each time step.

Thank you !