Itemwise Missing at Random Modeling for Incomplete Multivariate Data[1]

Mauricio Sadinle

Duke University and NISS

Supported by NSF grant SES-11-31897

[1] Joint work with Jerry Reiter

Incomplete Multivariate Data

Gender	Age	Income	...
F	25	60,000	...
M	?	?	...
?	51	?	...
F	?	150,300	
...	...	...	...

What This Talk is About

- Most common approach to handle missing data: assume the missing data are *missing at random* (MAR)
- We developed an alternative: assume the missing data are *itemwise missing at random* (IMAR)

What This Talk is About

- Most common approach to handle missing data: assume the missing data are *missing at random* (MAR)
- We developed an alternative: assume the missing data are *itemwise missing at random* (IMAR)

Outline

Inference with Missing Data

Itemwise Missing at Random

Take-Home Message

Example

- X: Would you lend me \$1,000?
- Want to estimate $\mathbb{P}(X = \text{yes})$
- $$\begin{aligned}\mathbb{P}(X = \text{yes}) = {} & \mathbb{P}(X = \text{yes}|\text{response})\mathbb{P}(\text{response}) \\ & + \mathbb{P}(X = \text{yes}|\text{non-response})\mathbb{P}(\text{non-response})\end{aligned}$$
- Most likely $\mathbb{P}(X = \text{yes}) << \mathbb{P}(X = \text{yes}|\text{response})$

Inference impossible without extra, usually untestable, assumptions on how missingness arises

Example

- X: Would you lend me \$1,000?
- Want to estimate $\mathbb{P}(X = \text{yes})$
- $$\begin{aligned}\mathbb{P}(X = \text{yes}) = {} & \mathbb{P}(X = \text{yes}|\text{response})\mathbb{P}(\text{response}) \\ & + \mathbb{P}(X = \text{yes}|\text{non-response})\mathbb{P}(\text{non-response})\end{aligned}$$
- Most likely $\mathbb{P}(X = \text{yes}) << \mathbb{P}(X = \text{yes}|\text{response})$

Inference impossible without extra, usually untestable, assumptions on how missingness arises

Example

- X: Would you lend me \$1,000?
- Want to estimate $\mathbb{P}(X = \text{yes})$
- $$\begin{aligned}\mathbb{P}(X = \text{yes}) = \mathbb{P}(X &= \text{yes}|\text{response})\mathbb{P}(\text{response}) \\ &+ \mathbb{P}(X = \text{yes}|\text{non-response})\mathbb{P}(\text{non-response})\end{aligned}$$
- Most likely $\mathbb{P}(X = \text{yes}) << \mathbb{P}(X = \text{yes}|\text{response})$

Inference impossible without extra, usually untestable, assumptions on how missingness arises

Example

- X: Would you lend me \$1,000?
- Want to estimate $\mathbb{P}(X = \text{yes})$
-

$$\begin{aligned}\mathbb{P}(X = \text{yes}) = \mathbb{P}(X = \text{yes}|\text{response})\mathbb{P}(\text{response}) \\ + \mathbb{P}(X = \text{yes}|\text{non-response})\mathbb{P}(\text{non-response})\end{aligned}$$

- Most likely $\mathbb{P}(X = \text{yes}) << \mathbb{P}(X = \text{yes}|\text{response})$

Inference impossible without extra, usually untestable, assumptions on how missingness arises

Notation

Gender	Age	Income	M_{Gender}	M_{Age}	M_{Income}	...
F	25	60,000	0	0	0	...
M	?	?	0	1	1	...
?	51	?	1	0	1	...
F	?	150,300	0	1	0	...
...	...	...	...	...	...	...

In general

- Study variables: $\mathbf{X} = (X_1, \ldots, X_p) \in \mathcal{X}$
- Missingness indicators: $\mathbf{M} = (M_1, \ldots, M_p) \in \{0,1\}^p$
- Missingness mechanism: $\mathbb{P}(\mathbf{M}|\mathbf{X})$

Notation

Gender	Age	Income	M_{Gender}	M_{Age}	M_{Income}	...
F	25	60,000	0	0	0	...
M	?	?	0	1	1	...
?	51	?	1	0	1	...
F	?	150,300	0	1	0	...
...	...	...	...	...	...	...

In general

- Study variables: $\mathbf{X} = (X_1, \ldots, X_p) \in \mathcal{X}$
- Missingness indicators: $\mathbf{M} = (M_1, \ldots, M_p) \in \{0, 1\}^p$
- Missingness mechanism: $\mathbb{P}(\mathbf{M}|\mathbf{X})$

Notation

Gender	Age	Income	M_{Gender}	M_{Age}	M_{Income}	...
F	25	60,000	0	0	0	...
M	?	?	0	1	1	...
?	51	?	1	0	1	...
F	?	150,300	0	1	0	...
...	...	...	...	...	...	...

In general

- Study variables: $\mathbf{X} = (X_1, \ldots, X_p) \in \mathcal{X}$
- Missingness indicators: $\mathbf{M} = (M_1, \ldots, M_p) \in \{0, 1\}^p$
- Missingness mechanism: $\mathbb{P}(\mathbf{M}|\mathbf{X})$

Notation

Gender	Age	Income	M_{Gender}	M_{Age}	M_{Income}	...
F	25	60,000	0	0	0	...
M	?	?	0	1	1	...
?	51	?	1	0	1	...
F	?	150,300	0	1	0	...
...	...	...	...	...	...	...

In general

- Study variables: $\mathbf{X} = (X_1, \ldots, X_p) \in \mathcal{X}$
- Missingness indicators: $\mathbf{M} = (M_1, \ldots, M_p) \in \{0, 1\}^p$
- Missingness mechanism: $\mathbb{P}(\mathbf{M}|\mathbf{X})$

Notation

Given $\mathbf{M} = \mathbf{m}$

- $\mathbf{X}_{\mathbf{m}}$: missing values (often written as $\mathbf{X}_{mis}$)
- $\mathbf{X}_{\bar{\mathbf{m}}}$: observed values (often written as $\mathbf{X}_{obs}$)

Example:

- $\mathbf{X} = (X_1, X_2, X_3)$
- If $\mathbf{m} = (1, 0, 1)$, $\mathbf{X}_{\mathbf{m}} = (X_1, X_3)$, and $\mathbf{X}_{\bar{\mathbf{m}}} = X_2$

Notation

Given $\mathbf{M} = \mathbf{m}$

- $\mathbf{X}_{\mathbf{m}}$: missing values (often written as $\mathbf{X}_{mis}$)
- $\mathbf{X}_{\bar{\mathbf{m}}}$: observed values (often written as $\mathbf{X}_{obs}$)

Example:

- $\mathbf{X} = (X_1, X_2, X_3)$
- If $\mathbf{m} = (1, 0, 1)$, $\mathbf{X}_{\mathbf{m}} = (X_1, X_3)$, and $\mathbf{X}_{\bar{\mathbf{m}}} = X_2$

Notation

Given $\mathbf{M} = \mathbf{m}$

- $\mathbf{X}_{\mathbf{m}}$: missing values (often written as $\mathbf{X}_{mis}$)
- $\mathbf{X}_{\bar{\mathbf{m}}}$: observed values (often written as $\mathbf{X}_{obs}$)

Example:

- $\mathbf{X} = (X_1, X_2, X_3)$
- If $\mathbf{m} = (1, 0, 1)$, $\mathbf{X}_{\mathbf{m}} = (X_1, X_3)$, and $\mathbf{X}_{\bar{\mathbf{m}}} = X_2$

Missing at Random Assumption

After Rubin (1976):

- Missing at random:

$$\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}})$$

- Missing completely at random:

$$\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{M} = \mathbf{m})$$

Missing at Random Assumption

After Rubin (1976):

- Missing at random:

$$\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}})$$

- Missing completely at random:

$$\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{M} = \mathbf{m})$$

Missing at Random Assumption

Under MAR

- $$\begin{aligned}\int \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) f(\mathbf{X} = \mathbf{x}) d\mathbf{x}_{\mathbf{m}} &= \int \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}) f(\mathbf{X} = \mathbf{x}) d\mathbf{x}_{\mathbf{m}} \\ &= \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}) \int f(\mathbf{X} = \mathbf{x}) d\mathbf{x}_{\mathbf{m}}\end{aligned}$$

  $\Rightarrow$ we can "ignore" the missingness mechanism

- $$\mathbb{P}(M_i = 1 | \mathbf{X} = \mathbf{x}) = \mathbb{P}(M_i = 1 | \mathbf{X}_{-i} = \mathbf{x}_{-i})$$

  Missingness of an item cannot depend on the value of the item

Missing at Random Assumption

Under MAR

- $$\begin{aligned}\int \mathbb{P}(\mathbf{M}=\mathbf{m}|\mathbf{X}=\mathbf{x})f(\mathbf{X}=\mathbf{x})d\mathbf{x}_{\mathbf{m}} &= \int \mathbb{P}(\mathbf{M}=\mathbf{m}|\mathbf{X}_{\bar{\mathbf{m}}}=\mathbf{x}_{\bar{\mathbf{m}}})f(\mathbf{X}=\mathbf{x})d\mathbf{x}_{\mathbf{m}} \\ &= \mathbb{P}(\mathbf{M}=\mathbf{m}|\mathbf{X}_{\bar{\mathbf{m}}}=\mathbf{x}_{\bar{\mathbf{m}}})\int f(\mathbf{X}=\mathbf{x})d\mathbf{x}_{\mathbf{m}}\end{aligned}$$

  $\Rightarrow$ we can "ignore" the missingness mechanism

- $$\mathbb{P}(M_i=1|\mathbf{X}=\mathbf{x})=\mathbb{P}(M_i=1|\mathbf{X}_{-i}=\mathbf{x}_{-i})$$

  Missingness of an item cannot depend on the value of the item

Missing at Random Assumption

Under MAR

- $$\int \mathbb{P}(\mathbf{M}=\mathbf{m}|\mathbf{X}=\mathbf{x})f(\mathbf{X}=\mathbf{x})d\mathbf{x}_{\mathbf{m}} = \int \mathbb{P}(\mathbf{M}=\mathbf{m}|\mathbf{X}_{\bar{\mathbf{m}}}=\mathbf{x}_{\bar{\mathbf{m}}})f(\mathbf{X}=\mathbf{x})d\mathbf{x}_{\mathbf{m}}$$
  $$= \mathbb{P}(\mathbf{M}=\mathbf{m}|\mathbf{X}_{\bar{\mathbf{m}}}=\mathbf{x}_{\bar{\mathbf{m}}})\int f(\mathbf{X}=\mathbf{x})d\mathbf{x}_{\mathbf{m}}$$

  $\Rightarrow$ we can "ignore" the missingness mechanism

- $$\mathbb{P}(M_i=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(M_i=1|\mathbf{X}_{-i}=\mathbf{x}_{-i})$$

  Missingness of an item cannot depend on the value of the item

Missing at Random Example

Under MAR

- $\mathbb{P}(M_1 = 1, M_2 = 1 | X_1 = x_1, X_2 = x_2) = c$
- $\mathbb{P}(M_1 = 0, M_2 = 1 | X_1 = x_1, X_2 = x_2) = u(x_1)$
- $\mathbb{P}(M_1 = 1, M_2 = 0 | X_1 = x_1, X_2 = x_2) = v(x_2)$
- $\mathbb{P}(M_1 = 0, M_2 = 0 | X_1 = x_1, X_2 = x_2) = 1 - c - u(x_1) - v(x_2)$

Missing at Random Example

Under MAR

- $\mathbb{P}(M_1 = 1, M_2 = 1 | X_1 = x_1, X_2 = x_2) = c$
- $\mathbb{P}(M_1 = 0, M_2 = 1 | X_1 = x_1, X_2 = x_2) = u(x_1)$
- $\mathbb{P}(M_1 = 1, M_2 = 0 | X_1 = x_1, X_2 = x_2) = v(x_2)$
- $\mathbb{P}(M_1 = 0, M_2 = 0 | X_1 = x_1, X_2 = x_2) = 1 - c - u(x_1) - v(x_2)$

Missing at Random Example

Under MAR

- $\mathbb{P}(M_1 = 1, M_2 = 1 | X_1 = x_1, X_2 = x_2) = c$
- $\mathbb{P}(M_1 = 0, M_2 = 1 | X_1 = x_1, X_2 = x_2) = u(x_1)$
- $\mathbb{P}(M_1 = 1, M_2 = 0 | X_1 = x_1, X_2 = x_2) = v(x_2)$
- $\mathbb{P}(M_1 = 0, M_2 = 0 | X_1 = x_1, X_2 = x_2) = 1 - c - u(x_1) - v(x_2)$

Missing at Random Example

Under MAR

- $\mathbb{P}(M_1 = 1, M_2 = 1 | X_1 = x_1, X_2 = x_2) = c$
- $\mathbb{P}(M_1 = 0, M_2 = 1 | X_1 = x_1, X_2 = x_2) = u(x_1)$
- $\mathbb{P}(M_1 = 1, M_2 = 0 | X_1 = x_1, X_2 = x_2) = v(x_2)$
- $\mathbb{P}(M_1 = 0, M_2 = 0 | X_1 = x_1, X_2 = x_2) = 1 - c - u(x_1) - v(x_2)$

Non-Ignorable Missingness Mechanisms

- Missing not at random:

$$\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) \neq \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}})$$

- General interest in developing non-ignorable approaches to handle missing data

Non-Ignorable Missingness Mechanisms

- Missing not at random:

$$\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X} = \mathbf{x}) \neq \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}})$$

- General interest in developing non-ignorable approaches to handle missing data

Identifiability Issues

- Generally speaking, inferences should be based on the full data distribution

$$f(\mathbf{X}, \mathbf{M})$$

- This distribution is not identifiable
- Examples of probabilities that cannot be estimated from the data alone, without extra assumptions
  - $\mathbb{P}(X_1 = x_1, M_1 = 1)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 1, M_2 = 0)$

Identifiability Issues

- Generally speaking, inferences should be based on the full data distribution

$$f(\mathbf{X}, \mathbf{M})$$

- This distribution is not identifiable
- Examples of probabilities that cannot be estimated from the data alone, without extra assumptions
  - $\mathbb{P}(X_1 = x_1, M_1 = 1)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 1, M_2 = 0)$

Identifiability Issues

- Generally speaking, inferences should be based on the full data distribution

$$f(\mathbf{X}, \mathbf{M})$$

- This distribution is not identifiable

- Examples of probabilities that cannot be estimated from the data alone, without extra assumptions

  - $\mathbb{P}(X_1 = x_1, M_1 = 1)$

  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 1, M_2 = 0)$

Identifiability Issues

- Generally speaking, inferences should be based on the full data distribution

$$f(\mathbf{X}, \mathbf{M})$$

- This distribution is not identifiable

- Examples of probabilities that cannot be estimated from the data alone, without extra assumptions

  - $\mathbb{P}(X_1 = x_1, M_1 = 1)$

  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 1, M_2 = 0)$

Observed Data Distribution

- The observed data distribution is all we can identify from samples
- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_{\mathbf{m}}} f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}}$
- For example with two categorical variables:
  - $\mathbb{P}(M_1 = 1, M_2 = 1)$
  - $\mathbb{P}(X_1 = x_1, M_1 = 0, M_2 = 1)$
  - $\mathbb{P}(X_2 = x_2, M_1 = 1, M_2 = 0)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 0, M_2 = 0)$

Observed Data Distribution

- The observed data distribution is all we can identify from samples
- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_{\mathbf{m}}} f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}}$
- For example with two categorical variables:
  - $\mathbb{P}(M_1 = 1, M_2 = 1)$
  - $\mathbb{P}(X_1 = x_1, M_1 = 0, M_2 = 1)$
  - $\mathbb{P}(X_2 = x_2, M_1 = 1, M_2 = 0)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 0, M_2 = 0)$

Observed Data Distribution

- The observed data distribution is all we can identify from samples
- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_{\mathbf{m}}} f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}}$
- For example with two categorical variables:
  - $\mathbb{P}(M_1 = 1, M_2 = 1)$
  - $\mathbb{P}(X_1 = x_1, M_1 = 0, M_2 = 1)$
  - $\mathbb{P}(X_2 = x_2, M_1 = 1, M_2 = 0)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 0, M_2 = 0)$

Observed Data Distribution

- The observed data distribution is all we can identify from samples
- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_{\mathbf{m}}} f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}}$
- For example with two categorical variables:
  - $\mathbb{P}(M_1 = 1, M_2 = 1)$
  - $\mathbb{P}(X_1 = x_1, M_1 = 0, M_2 = 1)$
  - $\mathbb{P}(X_2 = x_2, M_1 = 1, M_2 = 0)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 0, M_2 = 0)$

Observed Data Distribution

- The observed data distribution is all we can identify from samples
- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = \int_{\mathcal{X}_{\mathbf{m}}} f(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}}$
- For example with two categorical variables:
  - $\mathbb{P}(M_1 = 1, M_2 = 1)$
  - $\mathbb{P}(X_1 = x_1, M_1 = 0, M_2 = 1)$
  - $\mathbb{P}(X_2 = x_2, M_1 = 1, M_2 = 0)$
  - $\mathbb{P}(X_1 = x_1, X_2 = x_2, M_1 = 0, M_2 = 0)$

General Strategy

$$\begin{tikzcd}
\boxed{f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})} \arrow[d, "\text{Identifying assumption}"] \\
\boxed{\tilde{f}(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m})} \arrow[d, "\text{Sum over } \mathbf{m}"] \\
\boxed{\tilde{f}(\mathbf{X} = \mathbf{x})}
\end{tikzcd}$$

Non-Parametric Saturated Modeling

If

$$\int \tilde{f}(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}} = f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$$

- Modeling assumption does not impose restrictions on observed data distribution
- Robins (1997) refers to such modeling approach as being *non-parametric saturated*
- Important property for sensitivity analysis

Non-Parametric Saturated Modeling

If

$$\int \tilde{f}(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}} = f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$$

- Modeling assumption does not impose restrictions on observed data distribution
- Robins (1997) refers to such modeling approach as being *non-parametric saturated*
- Important property for sensitivity analysis

Non-Parametric Saturated Modeling

If

$$\int \tilde{f}(\mathbf{X} = \mathbf{x}, \mathbf{M} = \mathbf{m}) d\mathbf{x}_{\mathbf{m}} = f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m})$$

- Modeling assumption does not impose restrictions on observed data distribution
- Robins (1997) refers to such modeling approach as being *non-parametric saturated*
- Important property for sensitivity analysis

Outline

Inference with Missing Data

Itemwise Missing at Random

Take-Home Message

Itemwise Missing at Random

DEFINITION. *The missing data are itemwise missing at random (IMAR) if*

$$X_j \perp\!\!\!\perp M_j \mid \mathbf{X}_{-j}, \mathbf{M}_{-j}, \text{ for all } j = 1, \ldots, p.$$

REMARK. *X_j and M_j are conditionally independent but not necessarily marginally independent.*

Itemwise Missing at Random

DEFINITION. *The missing data are itemwise missing at random (IMAR) if*

$$X_j \perp\!\!\!\perp M_j \mid \mathbf{X}_{-j}, \mathbf{M}_{-j}, \text{ for all } j = 1, \ldots, p.$$

REMARK. *X_j and M_j are conditionally independent but not necessarily marginally independent.*

IMARㅤDistribution

THEOREM 1. *For each missingness pattern* $\mathbf{m} \in \mathcal{M} \subseteq \{0,1\}^p$, *given* $f(\mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{m}) > 0$, *let the function* $\eta_{\mathbf{m}} : \mathcal{X}_{\bar{\mathbf{m}}} \mapsto \mathbb{R}$ *be defined recursively as*

$$\eta_{\mathbf{m}}(\mathbf{x}_{\bar{\mathbf{m}}}) = \log f(\mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{m}) - \log \int_{\mathcal{X}_{\mathbf{m}}} \exp\left\{ \sum_{\mathbf{m}' \prec \mathbf{m}} \eta_{\mathbf{m}'}(\mathbf{x}_{\bar{\mathbf{m}}'}) I(\mathbf{m}' \in \mathcal{M}) \right\} \mu(d\mathbf{x}_{\mathbf{m}}).$$

Then

$$\tilde{f}(\mathbf{x}, \mathbf{m}) = \exp\left\{ \sum_{\mathbf{m}' \preceq \mathbf{m}} \eta_{\mathbf{m}'}(\mathbf{x}_{\bar{\mathbf{m}}'}) I(\mathbf{m}' \in \mathcal{M}) \right\}$$

satisfies

$$\int_{\mathcal{X}_{\mathbf{m}}} \tilde{f}(\mathbf{x}, \mathbf{m}) \mu(d\mathbf{x}_{\mathbf{m}}) = f(\mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{m}),$$

for all $(\mathbf{x}, \mathbf{m}) \in \mathcal{X} \times \mathcal{M}$.

THEOREM 2. *The distribution induced by* $\tilde{f}$ *encodes the IMAR assumption.*

IMAR Distribution

THEOREM 1. *For each missingness pattern* $\mathbf{m} \in \mathcal{M} \subseteq \{0,1\}^p$*, given* $f(\mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{m}) > 0$*, let the function* $\eta_{\mathbf{m}} : \mathcal{X}_{\bar{\mathbf{m}}} \mapsto \mathbb{R}$ *be defined recursively as*

$$\eta_{\mathbf{m}}(\mathbf{x}_{\bar{\mathbf{m}}}) = \log f(\mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{m}) - \log \int_{\mathcal{X}_{\mathbf{m}}} \exp\left\{\sum_{\mathbf{m}' \prec \mathbf{m}} \eta_{\mathbf{m}'}(\mathbf{x}_{\bar{\mathbf{m}}'}) I(\mathbf{m}' \in \mathcal{M})\right\} \mu(d\mathbf{x}_{\mathbf{m}}).$$

Then

$$\tilde{f}(\mathbf{x}, \mathbf{m}) = \exp\left\{\sum_{\mathbf{m}' \preceq \mathbf{m}} \eta_{\mathbf{m}'}(\mathbf{x}_{\bar{\mathbf{m}}'}) I(\mathbf{m}' \in \mathcal{M})\right\}$$

satisfies

$$\int_{\mathcal{X}_{\mathbf{m}}} \tilde{f}(\mathbf{x}, \mathbf{m}) \mu(d\mathbf{x}_{\mathbf{m}}) = f(\mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{m}),$$

for all $(\mathbf{x}, \mathbf{m}) \in \mathcal{X} \times \mathcal{M}$*.*

THEOREM 2. *The distribution induced by* $\tilde{f}$ *encodes the IMAR assumption.*

IMAR Distribution for Categorical Variables

- Log-linear model without interactions involving jointly X_j and M_j
- In the case of two variables:

$$\log \mathbb{P}(x_1, x_2, m_1, m_1) = \eta^{X_1 X_2}_{x_1 x_2} + \eta^{X_1 M_2}_{x_1 m_2} + \eta^{X_2 M_1}_{x_2 m_1} + \eta^{M_1 M_2}_{m_1 m_2} + \eta^{X_1}_{x_1} + \eta^{X_2}_{x_2} + \eta^{M_1}_{m_1} + \eta^{M_2}_{m_2} + \eta$$

IMAR Distribution for Categorical Variables

- Log-linear model without interactions involving jointly X_j and M_j
- In the case of two variables:

$$\begin{aligned}\log \mathbb{P}(x_1, x_2, m_1, m_1) = {} & \eta^{X_1 X_2}_{x_1 x_2} + \eta^{X_1 M_2}_{x_1 m_2} + \eta^{X_2 M_1}_{x_2 m_1} + \eta^{M_1 M_2}_{m_1 m_2} \\ & + \eta^{X_1}_{x_1} + \eta^{X_2}_{x_2} + \eta^{M_1}_{m_1} + \eta^{M_2}_{m_2} + \eta\end{aligned}$$

The Slovenian Plebiscite Data Revisited

- Slovenians voted for independence from Yugoslavia in a plebiscite in 1991
- The Slovenian public opinion survey included these questions:
  1. Independence: Are you in favor of Slovenian independence?
  2. Secession: Are you in favor of Slovenia's secession from Yugoslavia?
  3. Attendance: Will you attend the plebiscite?
- Rubin, Stern and Vehovar (1995) analyzed these three questions under MAR
- Plebiscite results give us the proportions of non-attendants and attendants in favor of independence

The Slovenian Plebiscite Data Revisited

- Slovenians voted for independence from Yugoslavia in a plebiscite in 1991

- The Slovenian public opinion survey included these questions:

  1. Independence: Are you in favor of Slovenian independence?
  2. Secession: Are you in favor of Slovenia's secession from Yugoslavia?
  3. Attendance: Will you attend the plebiscite?

- Rubin, Stern and Vehovar (1995) analyzed these three questions under MAR

- Plebiscite results give us the proportions of non-attendants and attendants in favor of independence

The Slovenian Plebiscite Data Revisited

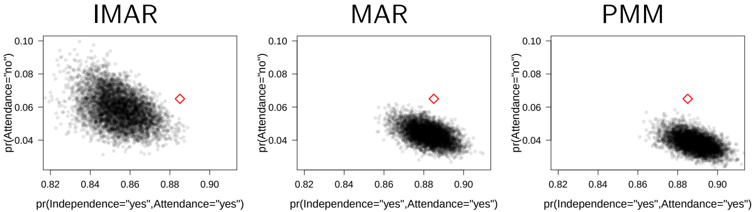

Figure: Samples from joint posterior distributions of pr(Independence = Yes, Attendance = Yes) and pr(Attendance = No). Pattern mixture model (PMM) under the complete-case missing-variable restriction.

IMAR Distribution for Continuous Variables

With continuous variables

$$f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}} | \mathbf{M} = \mathbf{m}) \mathbb{P}(\mathbf{M} = \mathbf{m})$$

- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}} | \mathbf{M} = \mathbf{m})$ can be estimated parametrically or non-parametrically
- IMAR distribution can be obtained following Theorem 1

IMAR Distribution for Continuous Variables

With continuous variables

$$f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}}, \mathbf{M} = \mathbf{m}) = f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}} | \mathbf{M} = \mathbf{m}) \mathbb{P}(\mathbf{M} = \mathbf{m})$$

- $f(\mathbf{X}_{\bar{\mathbf{m}}} = \mathbf{x}_{\bar{\mathbf{m}}} | \mathbf{M} = \mathbf{m})$ can be estimated parametrically or non-parametrically
- IMAR distribution can be obtained following Theorem 1

Self-Reporting Bias in Height Measurements

From the National Health and Nutrition Examination Survey — NHANES (1999–2000 and 2001–2002 cycles):

- X_1: self-reported height
- X_2: actual height measured by survey staff

Informally, the IMAR assumptions are:

- *the association between self-reported height and the reporting of this value is explained away by the true height and whether this measurement is taken*
- *the association between the true height and whether this measurement is taken is explained away by the height that would be self-reported and whether this value is reported*

Self-Reporting Bias in Height Measurements

From the National Health and Nutrition Examination Survey — NHANES (1999–2000 and 2001–2002 cycles):

- X_1: self-reported height
- X_2: actual height measured by survey staff

Informally, the IMAR assumptions are:

- *the association between self-reported height and the reporting of this value is explained away by the true height and whether this measurement is taken*
- *the association between the true height and whether this measurement is taken is explained away by the height that would be self-reported and whether this value is reported*

Self-Reporting Bias in Height Measurements

From the National Health and Nutrition Examination Survey — NHANES (1999–2000 and 2001–2002 cycles):

- X_1: self-reported height
- X_2: actual height measured by survey staff

Informally, the IMAR assumptions are:

- *the association between self-reported height and the reporting of this value is explained away by the true height and whether this measurement is taken*
- *the association between the true height and whether this measurement is taken is explained away by the height that would be self-reported and whether this value is reported*

Self-Reporting Bias in Height Measurements

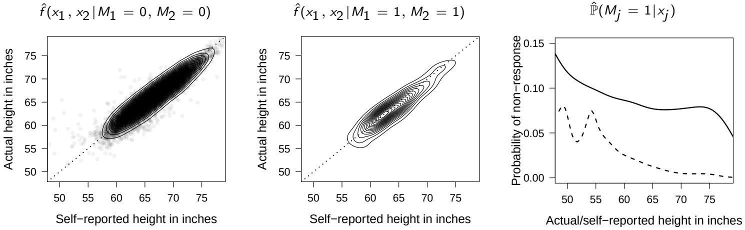

Figure: Estimated IMAR densities. Left: $\hat{\mathbb{P}}(M_j = 1 | x_j)$ for actual height (solid line), and for self-reported height (dashed line).

$f(x_1, x_2 | M_1 = 0, M_2 = 0)$, $f(x_2 | M_1 = 1, M_2 = 0)$, and $f(x_1 | M_1 = 0, M_2 = 1)$ estimated using survey-weighted kernel density estimators

Further Uses of IMAR Assumption

- Monotone missingness patterns (dropout/attrition)
- Sensitivity analysis to departures from IMAR assumption
- Use marginal information (e.g. from the Census) to parameterize departures from IMAR

Outline

Inference with Missing Data

Itemwise Missing at Random

Take-Home Message

Take-Home Message

- Itemwise missing at random assumption provides an alternative to MAR assumption
- Allows M_j to depend on X_j marginally
- Can be used with arbitrary missingness patterns and types of variables

Questions?

msadinle@stat.duke.edu