TW O-SA G E
M ULTIPL E I M PUTAT I ON

O FER HAREL
D EPARTMENT OF S TATISTICS A ND
T HE M ETHODOLOGY C ENTER
T HE P ENSYLVANIA S TATE U NIVERSITY

F EBRUARY 20, 2003
Outline

1. Motivation
2. Two-stage multiple imputation
3. Modeling the extended missingness
4. Applications
5. Discussion
1. Motivation

The missing-data problem

Most Statistical analysis and estimation procedures were not designed to handle missing values.

- Even small amount of missing data cause great difficulty
- The missing-data aspect is nuisance, not of primary interest
- Ad hoc or unprincipled missing-data methods may do more harm than good (inefficiency, bias, misleading uncertainty measures)
- Principled statistical solutions are computationally messy
Motivation

The goal

To make statistically valid inferences about population parameters from an incomplete dataset.

- Not to estimate, predict, or recover the missing values themselves
- Good to understand reasons for / causes of missingness
- Good to avoid modeling the missing-data mechanism if possible
- Untestable assumptions are inevitable
- Sensitivity analyses are helpful
Motivation

Older missing data methods

• Case deletion
• Case deletion and reweighting
• Averaging the available items
• Single imputation
  – Imputing unconditional means
  – Imputing from unconditional distribution
  – Imputing conditional means
  – Imputing from conditional distribution
Motivation

“New” Missing Data Methods

- Maximum likelihood methods
- Multiple imputation
- Weighting methods
- Selection models
- Pattern-mixture models
Motivation

Multiple imputation (Rubin, 1987): a simulation-based approach to missing data.

Imputation: Create $m$ imputations of the missing data, $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$, under a suitable model.

Analysis: Analyze each of the $m$ completed datasets in the same way.

Combination: Combine the $m$ sets of estimates and SE’s using Rubin’s (1987) rules.
Conventional MI
In general, we need to draw the imputations from a predictive distribution

\[ Y_{mis} \sim P^*(Y_{mis} \mid Y_{obs}, M), \]

where \( M \) is the missingness. This requires a joint model for the complete data \( Y_{com} = (Y_{obs}, Y_{mis}) \) and \( M \),

\[ P(Y_{com}, M) = P(Y_{com}) P(M \mid Y_{com}). \]

**Missing at Random (MAR):** If the distribution of missingness doesn’t depend on the missing data, i.e.

\[ P(M \mid Y_{com}) = P(M \mid Y_{obs}), \]

then \( Y_{mis} \) is MAR and we may ignore the model for \( M \). Thus the imputations are generated from

\[ Y_{mis} \sim P^*(Y_{mis} \mid Y_{obs}) \]

(Rubin, 1976). This is called **ignorability**.
Rubin’s rules

Calculate and store

$$\hat{Q}^{(j)} = \text{estimate of } Q$$
$$U^{(j)} = \text{standard error}^2$$

for \( j = 1, \ldots, m \) and combine:

$$\bar{Q} = m^{-1} \sum_{j=1}^{m} \hat{Q}^{(j)}$$
$$\bar{U} = m^{-1} \sum_{j=1}^{m} U^{(j)}$$

\[
B = (m - 1)^{-1} \sum_{j=1}^{m} \left( \hat{Q}^{(j)} - \bar{Q} \right)^2
\]

\[
T = \bar{U} + (1 + m^{-1})B
\]

An approximate 95\% interval for \( Q \) is

$$\bar{Q} \pm t_\nu \sqrt{T},$$

where the degrees of freedom are

$$\nu = (m - 1) \left[ \frac{(1 + m^{-1})B}{T} \right]^{-2}.$$
The relative increase in variance due to missing data is
\[ r = \frac{(1 + m^{-1})B}{\bar{U}} \]
and the rate of missing information is
\[ \lambda = \frac{r + 2/(\nu + 3)}{1 + r}. \]
This is approximately
\[ \lambda = \frac{r}{1 + r} \]
when \( \nu \) is large.
Motivation for our work

In many situations, missing values may be of two qualitatively different types. Examples include:

- planned missingness versus unplanned
- unit nonresponse versus item nonresponse
- latent variables versus missing manifest items
- dropouts versus subjects who return
- deaths versus other types of dropout
- refusal versus “don’t know”
A good idea

Partition the missing data as $Y_{mis} = (Y^A_{mis}, Y^B_{mis})$ and
draw it in two stages:

**Stage 1:** Draw $m$ imputations of $Y^A_{mis}$.

**Stage 2:** For each imputation of $Y^A_{mis}$, draw $n > 1$
imputations of $Y^B_{mis}$.

This results in $mn$ nested imputation sets.

Why do we want to do this?

- computational convenience (Shen, 2000)
- assess the separate contributions of $Y^A_{mis}$ and
  $Y^B_{mis}$ to overall uncertainty
- provide a framework for applying different
  assumptions to the two types of missing values
  (e.g., $Y^A_{mis}$ is “ignorable” and $Y^B_{mis}$ is
  “nonignorable”)
2. Two-stage MI

Extends Rubin’s (1987) framework to two types of missing values.

**Imputation:** Create \( m \) imputations of \( Y_{mis}^A \) the first level of the missing data, \( Y_{mis}^{A(1)}, \ldots, Y_{mis}^{A(m)} \).

Then, for each \( Y_{mis}^A \), generate \( n \) imputations of \( Y_{mis}^B \) from a conditional predictive distribution given \( Y_{mis}^A \).

**Analysis:** Analyze each of the \( mn \) completed datasets in the same way.

**Combination:** Combine the \( mn \) sets of estimates and SE’s by Shen’s (2000) rules.
Two-stage MI

A A A A B B B B
A A A A B B B B
A A A A B B B B

1 ... n 1 ... n 1 ... n 1 ... n

1 2 3 ... m
**Drawing two-stage imputations**

In general, we need to draw the imputations from the posterior distributions

\[
Y_{mis}^A \sim P^*(Y_{mis}^A \mid Y_{obs}, M^+),
\]
\[
Y_{mis}^B \sim P^*(Y_{mis}^B \mid Y_{obs}, Y_{mis}^A, M^+),
\]

where \( M^+ \) is the **extended missingness**. This requires a joint model for the complete data \( Y_{com} = (Y_{obs}, Y_{mis}^A, Y_{mis}^B) \) and \( M^+ \),

\[
P(Y_{com}, M^+) = P(Y_{com}) P(M^+ \mid Y_{com}).
\]

Under certain MAR-like conditions, we can ignore the missingness, \( M^+ \), partially or fully.
Shen’s rules

From Shen (2000, unpublished Ph.D. thesis)

Calculate and store

\[ \hat{Q}^{(j,k)} = \text{estimate of } Q \]
\[ U^{(j,k)} = \text{standard error}^2 \]

for \( j = 1, \ldots, m \) and \( k = 1, \ldots, n \). Then:

\[ \bar{Q}_{..} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{k=1}^{n} \hat{Q}^{(j,k)} \]
\[ \bar{Q}_{.j} = \frac{1}{n} \sum_{k=1}^{n} \hat{Q}^{(j,k)} \]
\[ \bar{U} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{k=1}^{n} U^{(j,k)} \]
\[ B = \frac{1}{m-1} \sum_{j=1}^{m} \left( \hat{Q}^{(j,.)} - \bar{Q}_{..} \right)^2 \]
\[ W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{k=1}^{n} \left( \hat{Q}^{(j,k)} - \bar{Q}_{.j} \right)^2 \]
\[ T = \bar{U}_{..} + (1 + \frac{1}{m})B + (1 - \frac{1}{n})W \]
An approximate 95% interval for $Q$ is

$$\bar{Q} \pm t_\nu \sqrt{T},$$

where the degrees of freedom are

$$\nu^{-1} = \frac{1}{m(n - 1)} \left( \frac{(1 - \frac{1}{n})W}{T} \right)^{-2} + \frac{1}{(m - 1)} \left( \frac{(1 + \frac{1}{m})B}{T} \right)^{-2}.$$

The overall estimated rate of missing information is

$$\hat{\lambda} = \frac{B + (1 - n^{-1})W}{\bar{U}.. + B + (1 - n^{-1})W}$$

the estimated rate of missing information due to $Y_{mis}^B$ if $Y_{mis}^A$ was known is

$$\hat{\lambda}^{B|A} = \frac{W}{\bar{U}.. + W}$$

and the difference $\hat{\lambda}^A = \hat{\lambda} - \hat{\lambda}^{B|A}$ represent the decrease in the rate of missing information if $Y_{mis}^A$ became known.
3. Models for extended missingness

In many cases, it is natural to factor the distribution of $M^+$ into two sub-models,

$$M^+ = (M^1, M^2)$$

$$P(M^+|Y_{com}) = P(M^1|Y_{com})P(M^2|Y_{com}, M^1)$$

We have identified five different ways to do this:

- Post hoc classification
- Sequential selection
  - Forward
  - Reverse
- Partitioned risk set
  - Forward
  - Reverse
Possible assumptions

Extended ignorability

\( \text{MAR}^+ \): Extended missing at random

\[ P(M^+|Y_{\text{com}}) = P(M^+|Y_{\text{obs}}) \]

CMAR\(^+\): Conditional extended missing at random

\[ P(M^+|Y_{\text{com}}) = P(M^+|Y_{\text{obs}}, Y^A_{mis}) \]

Ignorability conditions for submodels

\( \text{MAR}^{2|1} \): Missing at random for the second sub-model

\[ P(M^{2|1}|Y_{\text{com}}, M^1) = P(M^{2|1}|Y_{\text{obs}}, M^1) \]

\( \text{CMAR}^{2|1} \): Conditional missing at random for the second sub-model

\[ P(M^{2|1}|Y_{\text{com}}, M^1) = P(M^{2|1}|Y_{\text{obs}}, Y^A_{mis}, M^1) \]
Results

Extended ignorability

Result 1: If we have distinctness and MAR\(^+\) holds, we can ignore the information contained in \(M^+\) in both stages of imputation so that

\[
P(Y_{mis}^A | Y_{obs}, M^+) = P(Y_{mis}^A | Y_{obs})
\]

and

\[
P(Y_{mis}^B | Y_{obs}, Y_{mis}^A, M^+) = P(Y_{mis}^B | Y_{obs}, Y_{mis}^A)
\]

Result 2: If we have distinctness and CMAR\(^+\) holds, we can ignore the information contained in \(M^+\) in the second stage of the imputation so that

\[
P(Y_{mis}^B | Y_{obs}, Y_{mis}^A, M^+) = P(Y_{mis}^B | Y_{obs}, Y_{mis}^A)
\]
Results

Ignorability conditions for submodels

Result 3: If we have distinctness and MAR$^{2|1}$ holds, we can ignore the information contained in $M^{2|1}$ in both stages of imputation so that

$$P(Y^A_{mis}|Y_{obs}, M^+) = P(Y^A_{mis}|Y_{obs}, M^1)$$

and

$$P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^+) = P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^1)$$

Result 4: If we have distinctness and CMAR$^{2|1}$ holds, we can ignore the information contained in $M^{2|1}$ in the second stage of the imputation so that

$$P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^+) = P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^1)$$
Post hoc classification

\[ P(M^+ | Y_{com}) = P(M | Y_{com}) P(M^* | Y_{com}, M) \]

where

- \( M \) divides \( Y_{com} \) into \( (Y_{obs}, Y_{mis}) \)
- \( M^* \) divides \( Y_{mis} \) into \( (Y_{mis}^A, Y_{mis}^B) \)

**Example:** A = “don’t know” \quad B = “refusal”

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( M )</th>
<th>( M^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Obs.</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Example

\[ Y_{mis}^A = \text{don’t know} \quad Y_{mis}^B = \text{refusal} \]

\[
\begin{array}{cccc}
X & Y & M & M^* \\
\hline
\text{Obs.} & 0 & 0 & \text{Obs.} \\
\text{Obs.} & 0 & \vdots & \text{Obs.} \\
A & 1 & 0 & \text{A} \\
B & \vdots & 0 & \text{B} \\
& 1 & 1 & \\
\end{array}
\]

- If the regression for \( M^* \) does not involve \( Y \), then \( \text{MAR}^* \) and \( \text{CMAR}^* \) are satisfied; and we do not have to model the process that divides nonrespondents into the two groups.

- If, in addition, the regression for \( M \) does not involve \( Y \), we have \( \text{MAR}^+ \) and \( \text{CMAR}^+ \), and we do not need to model overall nonresponse.
**Sequential selection: forward**

\[ P(M^+ | Y_{com}) = P(M^A | Y_{com}) P(M^{B|A} | Y_{com}, M^A) \]

where

- \( M^A \) divides \( Y_{com} \) into \( Y^A_{mis} \) and \( (Y_{obs}, Y^B_{mis}) \).
- \( M^{B|A} \) divides \( (Y_{obs}, Y^B_{mis}) \) into \( Y_{obs} \) and \( Y^B_{mis} \).

**Example:** two waves with dropout

| \( M^A \) | \( M^{B|A} \) |
|---|---|
| 0 | 0 |
| 0 | 1 |
| 1 | 1 |

**Table:**

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y_1 )</th>
<th>( Y_2 )</th>
<th>( M^A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs.</td>
<td>Obs.</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>Obs.</td>
<td>B</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
If the regression for $M^{B|A}$ does not involve $Y_2$, then MAR$^{B|A}$ and CMAR$^{B|A}$ are satisfied. Thus we do not have to model the process of dropout at time 2.

If, in addition, the regression for $M^A$ does not involve $Y_1$ or $Y_2$, we have MAR$^+$ and CMAR$^+$. We do not need to model the dropout at time 1 either.
Sequential selection: reverse

Like sequential forward, but labels A,B are reversed.

\[ P(M^+|Y_{com}) = P(M^B|Y_{com}) P(M^{A|B}|Y_{com}, M^B) \]

where

- \( M^B \) divides \( Y_{com} \) into \( Y^{B}_{mis} \) and \( (Y_{obs}, Y^{A}_{mis}) \).
- \( M^{A|B} \) divides \( (Y_{obs}, Y^{A}_{mis}) \) into \( Y_{obs} \) and \( Y^{A}_{mis} \).

\[
\begin{array}{cccc}
X & Y_1 & Y_2 & M^B & M^{A|B} \\
\hline
\text{Obs.} & & & 0 & 0 \\
\text{Obs.} & & & 0 & \vdots \\
\text{Obs.} & & & 0 & 0 \\
\text{A} & & 0 & 1 & \vdots \\
\text{Obs.} & & & 0 & 1 \\
\text{B} & & & 1 & \vdots \\
\text{B} & & & 1 & 1 \\
\end{array}
\]

Results are analogous to previous ones.
Partitioned risk set

Appropriate when \( Y_{com} = (Y^A_{com}, Y^B_{com}) \), where

\[
\begin{align*}
Y^A_{com} & : \text{at risk for } A \text{ type} \\
Y^B_{com} & : \text{at risk for } B \text{ type}
\end{align*}
\]

**Forward:**

\[
P(M^+|Y_{com}) = P(M^A|Y_{com})P(M^{BA}|Y_{com}, M^A)
\]

- \( M^A \) divides \( Y^A_{com} \) into \( Y^A_{obs} \) and \( Y^A_{mis} \).
- \( M^{BA} \) divides \( Y^B_{com} \) into \( Y^B_{obs} \) and \( Y^B_{mis} \).

**Reverse:**

\[
P(M^+|Y_{com}) = P(M^B|Y_{com})P(M^{AB}|Y_{com}, M^B)
\]

- \( M^B \) divides \( Y^B_{com} \) into \( Y^B_{obs} \) and \( Y^B_{mis} \).
- \( M^{AB} \) divides \( Y^A_{com} \) into \( Y^A_{obs} \) and \( Y^A_{mis} \).
Partitioned risk set

Example:

\[ L = \text{Latent variable} \]
\[ (Y_1, Y_2, \ldots, Y_p) = \text{manifest variables} \]

\[
\begin{array}{c|c|c|c|c}
L & Y_1 & Y_2 & \cdots & Y_p \\
\hline
A & B & & & \\
A & B & & & \\
A & B & & & \\
A & B & & & \\
A & B & & & \\
A & B & & & \\
A & B & & & \\
\vdots & B & & & \\
A & B & & & \\
\end{array}
\]

(Forward)

\[
\begin{array}{c|c|c|c|c}
L & Y_1 & Y_2 & \cdots & Y_p \\
\hline
B & A & & & \\
B & A & & & \\
B & A & & & \\
B & A & & & \\
B & A & & & \\
B & A & & & \\
B & A & & & \\
\vdots & A & & & \\
B & A & & & \\
\end{array}
\]

(Reverse)

Results are analogous to previous ones.
Application #1: LCA

Attitudes toward abortion.

- Data from General Social Survey (GSS) 1974-1994 (not all the years), sample of 32,380 adults.

- Respondents were asked whether they approved or disapproved of abortions in three scenarios involving ‘ethical/medical’ reasons (1, 2, 3) and three scenarios involving ‘social’ reasons (4, 5, 6).

- A Latent Class model with 3 classes.
  - disapprove in all circumstances.
  - approve for ethical/medical reasons only.
  - approve in all circumstances.

- Missing values on each item about 20%.
LCA Example (continued)

$Y_{mis}^A$ = Latent variable  $Y_{mis}^B$ = manifest variables

- Allows us to “see” what may happen if measurement error is eliminated.

$Y_{mis}^A$ = manifest variables  $Y_{mis}^B$ = latent variable

- Allows us to “see” what may happen if item NR is eliminated.

Note: In both analyses, we will assume \(\text{MAR}^+\) so that we can use WinLTA.
LCA Example (continued)

- Follow the following steps:
  1. Impute $Y_{mis} = (Y^A_{mis}, Y^B_{mis})$ $m$ times using WinLTA.

    For each of the $m$ imputations:
    (a) **Forward**: throw away the imputed values for $Y^B_{mis} = \text{item NR}$, and re-impute them $n - 1$ times.
    (b) **Reverse**: throw away the imputed values for $Y^B_{mis} = \text{latent class}$, and re-impute them $n - 1$ times.

**Note**: In (a), we have to modify the model to keep latent classes fixed at these imputed values.
**LCA Example - Results**

\[
\begin{align*}
\gamma_1 &= P(\text{always disapprove}) \\
\gamma_2 &= P(\text{approve for ethical/medical reasons only}) \\
\gamma_3 &= P(\text{always approve})
\end{align*}
\]

**Forward:** \( A = \text{Latent variable} \quad B = \text{item NR} \)

<table>
<thead>
<tr>
<th></th>
<th>est</th>
<th>SE</th>
<th>%mis</th>
<th>%misA</th>
<th>%misB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_1 )</td>
<td>.45</td>
<td>.01</td>
<td>28</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>.45</td>
<td>.01</td>
<td>28</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>( \gamma_3 )</td>
<td>.10</td>
<td>.04</td>
<td>80</td>
<td>69</td>
<td>11</td>
</tr>
</tbody>
</table>

**Reverse:** \( A = \text{item NR} \quad B = \text{latent variable} \)

<table>
<thead>
<tr>
<th></th>
<th>est</th>
<th>SE</th>
<th>%mis</th>
<th>%misA</th>
<th>%misB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_1 )</td>
<td>.45</td>
<td>.01</td>
<td>21</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>.45</td>
<td>.01</td>
<td>21</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>( \gamma_3 )</td>
<td>.10</td>
<td>.04</td>
<td>74</td>
<td>11</td>
<td>63</td>
</tr>
</tbody>
</table>
Discussion

- One should consider modeling $M^+$.  
- To say “$Y_{mis}^A$” is ignorable and “$Y_{mis}^B$” is nonignorable has many different meanings; it depends on how you factor $M^+$. 
- Especially useful for research design.  
- In principle, one could do ML in two stages, but two-stage MI is easier.