Can Missing Patterns in Covariates Improve Imputation for Missing Data?

BigSurv18, Barcelona, Spain

Micha Fischer, Felicitas Mittereder
University of Michigan
Program in Survey Methodology

October 27, 2018
Background

Most (survey) data sets have missing data:

Treatment:

▶ Imputation of plausible values to receive a data set without missings (e.g., sequential imputation step)

Problems:

▶ Bias due to oversimplified models in the sequential imputation step
▶ Information that “respondent did not answer a question” is lost
▶ If missing data mechanism is Missing Not At Random (MNAR), the item missing pattern can be informative for the imputed values
Basic Idea

- Include “Missing” as own category in imputation model to improve imputation accuracy and therefore estimators from survey data.
- Tree-based methods (e.g., random forest) can incorporate this additional information and account for complex interactions, (Doove, Van Buuren, and Dusseldorp 2014)
- Increasing efficiency by skipping sequential imputation steps

⇒ Easy to implement in current software
Previous Research

- Loh et al. (2018) use missing values in a regression and classification tree (GUIDE) to impute missing values.

- Ding and Simonoff (2010) show that random forest can handle incomplete covariates by coding “missing” as its own category.
Potentially New Approach

- Combine both approaches, “tree-based imputation” and “missing as its own category” with CART and random forest:
  - Categorical covariates: Treat “Missing” as its own category
  - Continuous covariates: Code “Missing” to arbitrary value “(far) away” from the actual data

⇒ One kind of pattern mixture model which partition data by patterns of missing values (Little 1993)
Example: Potential Part of Tree for Imputing Household Income

Value 1

Value 2

Indiv. income

e.g., Indiv. income (categorical)

Missing

Not missing

e.g. Household size

Missing

Not Missing
Assumption on Missing Data - Venn Diagram:

Set of missing mechanisms depending on missing patterns in covariates
Theoretical Considerations

- Many covariates with missing values: potential advantage due to more available data
- Reasonable when missing due to “not applicable” cases
Generating covariables: \( X_i \sim N(0, 1) \) and \( Z_i \sim N(0, 1) \)

Generating response indicators for \( X \) and \( Z \):

\[
R_{Z,i} \sim Ber(p_Z) \quad \text{and} \quad R_{Z,i} = \begin{cases} 
1 \ & \text{for} \ p_Z \geq u_{Z,i}, \\
0 \ & \text{for} \ p_Z < u_{Z,i}
\end{cases}
\]

where \( u_{Z,i} \sim Unif(0, 1) \)
Simulation Study - Set Up (2)

▶ Generating outcome variable $Y$:
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 R_{Z,i} + \beta_4 X_i R_{Z,i} + \epsilon_i$
where $\epsilon_i \sim N(0, 1)$

▶ Generating response indicator for $Y$:
$P(R_{Y,i} = 1) = \logit^{-1}(p_Y + \delta_1 X_i + \delta_2 Z_i + \delta_3 R_{Z,i} + \delta_4 Y_i)$

with $p_Y$ as the baseline response rate,

and $R_{Y,i} = \begin{cases} 1 & \text{for } P(R_{Y,i} = 1) \geq u_{Y,i}, \\ 0 & \text{for } P(R_{Y,i} = 1) < u_{Y,i} \end{cases}$

where $u_{Y,i} \sim \text{Unif}(0, 1)$

$\Rightarrow Y_{\text{obs},i} = \begin{cases} Y_i & \text{if } R_{Y,i} = 1, \\ \text{missing} & \text{if } R_{Y,i} = 0 \end{cases}$
$Z_{\text{obs},i} = \begin{cases} Z_i & \text{if } R_{Z,i} = 1, \\ \text{missing} & \text{if } R_{Z,i} = 0 \end{cases}$
Simulation Study - Data Structure

Table 1: Resulting data structure

<table>
<thead>
<tr>
<th>$Y_{obs}$</th>
<th>$X_{obs}$</th>
<th>$Z_{obs}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>X</td>
<td>Z</td>
</tr>
<tr>
<td>Y</td>
<td>X</td>
<td>miss</td>
</tr>
<tr>
<td>miss</td>
<td>X</td>
<td>Z</td>
</tr>
<tr>
<td>miss</td>
<td>X</td>
<td>miss</td>
</tr>
</tbody>
</table>
Simulation Study - Procedure

- Single imputation with linear models, CART and random forest using $X_{\text{obs},i}$, $Y_{\text{obs},i}$, $Z_{\text{obs},i}$
- Additionally, CART and random forest using “Missing” information as own category
- Assessment on RMSE of regression coefficients after imputation in a substantive model - Here:

$$Y_{\text{imp},i} \sim X_{\text{obs},i} + Z_{\text{imp},i}$$
## Simulation - Parameters

**Table 2: Implemented parameter values**

<table>
<thead>
<tr>
<th>Y</th>
<th>Parameter</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values</td>
<td>1</td>
<td>1</td>
<td>{0; 2}</td>
<td>{0; 2}</td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td>$\delta_1$</td>
<td>$\delta_2$</td>
<td>$\delta_3$</td>
<td>$\delta_4$</td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>1</td>
<td>{0; 1}</td>
<td>{0; 1}</td>
<td>{0; 1}</td>
<td></td>
</tr>
<tr>
<td>Baseline response</td>
<td>Parameter</td>
<td>$p_Y$</td>
<td>$p_Z$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>0.5</td>
<td>{0.3; 0.7}</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\Rightarrow$ MAR situation if $\beta_4 = \delta_4 = 0$

$\Rightarrow$ MNAR situation if $\beta_4 \neq 0 \lor \delta_4 \neq 0$
Simulation Results - MAR

Parameter values: $\beta_3 = 2$, $\beta_4 = 0$ and $\delta_2 = \delta_4 = 0$, $\delta_3 = 1$
Simulation Results - MNAR

Parameter values: $\beta_3 = \beta_4 = 2$ and $\delta_2 = \delta_4 = 0$, $\delta_3 = 1$
Simulation Results - MNAR

Parameter values: $\beta_3 = \beta_4 = 2$ and $\delta_2 = \delta_3 = \delta_4 = 1$

![Graph showing RMSE values for different imputation methods]
Future Research

1. Simulation extension for non-normal distributed variables (e.g., binary variables)
2. Evaluation on survey data linked to administrative records
3. Accounting for imputation uncertainty in variance estimation
Thank you for your attention!

Any questions?

fmitter@umich.edu
michaf@umich.edu


Loh, Wei-Yin, John Eltinge, Moon Jung Cho, and Yuanzhi Li. 2018. “CLASSIFICATION and Regression Trees and Forests for Incomplete Data from Sample Surveys.” *Statistica Sinica*.