Sequential Imputation with Integrated Model Selection: A Novel Approach to Missing Value Imputation in High-Dimensional (Survey) Data

Micha Fischer
Advisors: Brady T. West, Roderick J. A. Little
University of Michigan
Program in Survey Methodology

BigSurv20
Continuing to Explore New Statistical Frontiers at the Intersection of Big Data and Survey Science

November 27, 2020

Problem

- Incomplete survey data
  - Item nonresponse
  - Unit nonresponse
  - Failure to link records
  - Panel attrition
- Missing values are most likely not Missing Completely At Random (MCAR)
- High number of variables with any possible distribution in survey data

$\Rightarrow$ Usual approach: multiple sequential imputation

- Iteratively imputing each variable with missing values conditional on all other variables
- Based on Missing At Random (MAR)

Why is it a problem?

Standard procedures (e.g. MICE) need specified model for each incomplete variable

- Subjectivity:
  - Method selection
  - Model specification
- Efficiency: limited resources (time, labor)

Additional, standard procedures can fail in high-dimensional data sets (see e.g. Loh et al. (2018), Razzak and Heumann (2019))

Research Question

How can missing data imputation in high-dimensional (survey) data be automated?

For example:

- Health and Retirement Study: over 6,000 variables
- Panel Study of Income Dynamics: over 5,000 variables

Proposed Solution

- Sequential imputation:
  - Iteratively imputing each variable with missing values conditional on all other variables

New:

- Within sequential imputation procedure:
  - Automated method selection
  - Automated model specification
- Advantages:
  - Many different methods possible
  - Objective procedure

Used Methods

1. Regularized (G)LM (Deng et al. 2016)
2. Classification and regression tree (CART) (Burgette and Reiter 2010)
3. Random Forest (Shah et al. 2014)
4. Bayesian Additive Regression Trees (BART) (Xu, Daniels, and Winterstein 2016)

Sequential Imputation with Integrated Method Selection (SIIMS) - Procedure

For each iteration:

1. For each method *m*:
   - Fit a model using all covariates
   - Estimate criteria assessing:
     - Distribution of imputed values
     - Conditional mean (i.e. the structural form)
2. Combine these criteria to a single method assessment criterion
3. Select method with minimal criterion and update imputed values
4. Repeat 1 - 3 for all variables with missing values

$\Rightarrow$ Repeat procedure to create multiply imputed data sets

Criterion 1: Distribution of Imputed Values

Adapted from Bondarenko and Raghunathan (2016):

1. Estimate response propensity score $\hat{e}$ for incomplete variable Y:

$$\hat{e} = P(R = 1|\mathbf{X}), \;\; R = \begin{cases} 1 \text{ if } Y \text{ observed,} \\ 0 \text{ if } Y \text{ missing} \end{cases}$$

2. Estimate conditional densities for observed values conditional on propensity score:

$$\hat{f}(Y|\hat{e}, R = 1)$$

3. For all m potential methods, fit model and predict sets of missing values:

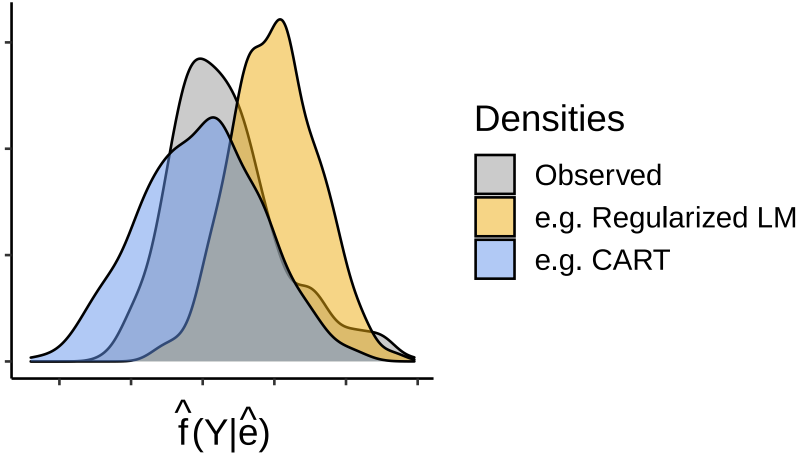
$$\hat{Y}_m|\mathbf{X}, R = 0$$

4. Estimate conditional densities for imputed values conditional on propensity score:

$$\hat{f}(\hat{Y}_m|\hat{e}, R = 0)$$

Criterion 1: Distribution of Imputed Values (cont.)

Comparing $\hat{f}(Y|\hat{e}, R = 1)$ (observed) and $\hat{f}(\hat{Y}_m|\hat{e}, R = 0)$ (imputed):



$\Rightarrow$ Automation: comparing via measure of similarity (here: Hellinger's distance H_m)

Criterion 2: Conditional mean

Pseudo MSE on observed values $Y|R = 1$:

For a scalar $Y_i|R_i = 1$, we compute a combined measure of prediction accuracy and variability:

$$S_{i,m} = \overbrace{(\bar{Y}_{i,m} - Y_i)^2}^{\text{Bias}^2} + \overbrace{\frac{1}{B-1}\sum_{b=1}^{B}(Y_{i,m}^{(b)} - \bar{Y}_{i,m})^2}^{\text{Variance}}$$

$\Rightarrow$ Averaging over all $S_{i,m}$ leads to the MSE-like measure MSE_m^*

- Measure of how well conditional mean is modeled
- $S_{i,m}$ available on a scalar level

How to combine criteria?

Weighted sum of standardized $H_m (\widetilde{H}_m)$, and $MSE^*_m (\widetilde{MSE^*_m})$:
$\Rightarrow$ single method assessment criterion for method m (MAC_m):

$$MAC_m = w_1 * \widetilde{H}_m + w_2 * \widetilde{MSE^*_m}$$

Weighting:

- H_m: Plausibility of imputed values under MAR
- MSE^*_m: Essential model structure, necessary for unbiased estimates under MAR

$\Rightarrow$ Three different sets of weights:

1. $w_1 = 1$, and $w_2 = 0$
2. $w_1 = 0$, $w_2 = 1$
3. $w_1 = w_2 = 0.5$

Additional features

- Binary variables
- Optional upstream variable selection
- Optional double robust property (Zhang and Little 2009)

Simulation Design

Compared imputation approaches:

- SIIMS
- MICE using Random Forest

Assessment:

- Accuracy of multiple imputed data
- Runtime of the imputation process

$\Rightarrow$ Trade-off between accuracy and process time

Basic Results

- Accuracy (bias of β coefficients):
  - About the same for SIIMS and MICE
- Process time:

	1000 obs.	5000 obs.
SIIMS	28 min	3.1 h
MICE	2.4 sec	23.6 sec

Next Steps and Future Directions

- Increase Speed:
  - Track runtime per method
  - Simplify hyper-parameter tuning
- Simulation on high-dimensional data

Thank you for your attention!

Any questions?

michaf@umich.edu

References

Bondarenko, Irina, and Trivellore Raghunathan. 2016. "Graphical and Numerical Diagnostic Tools to Assess Suitability of Multiple Imputations and Imputation Models." *Statistics in Medicine* 35 (17): 3007–20.

Burgette, Lane F, and Jerome P Reiter. 2010. "Multiple Imputation for Missing Data via Sequential Regression Trees." *American Journal of Epidemiology* 172 (9): 1070–6.

Deng, Yi, Changgee Chang, Moges Seyoum Ido, and Qi Long. 2016. "Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data." *Scientific Reports* 6: 21689.

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*.

Razzak, Humera, and Christian Heumann. 2019. "Hybrid Multiple Imputation in a Large Scale Complex Survey." *STATISTICS* 33.

Shah, Anoop D, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using Mice: A Caliber Study." *American Journal of Epidemiology* 179 (6): 764–74.

Xu, Dandan, Michael J Daniels, and Almut G Winterstein. 2016. "Sequential Bart for Imputation of Missing Covariates." *Biostatistics* 17 (3): 589–602.

Zhang, Guangyu, and Roderick Little. 2009. "Extensions of the Penalized Spline of Propensity Prediction Method of Imputation." *Biometrics* 65 (3): 911–18.