# 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

JPSM/MPSM Seminar

December 9th, 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

# Outline

- Proposed solution
- Small scale simulation
- Large scale simulation

# Proposed Solution

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

## New:

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

# Used Methods

1. Bayesian (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)

# Automated Model Specification

1. Parametric models: Bayesian (G)LM
   - Perform Elastic Net to determine model formula
   - Fit Bayesian model with determined formula
2. Tree-based methods: (CART, Random Forest , BART)
   - No predefined model formula necessary

# Proposed Solution

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

New:

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

# Automated Method Selection - Criterion 1

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:

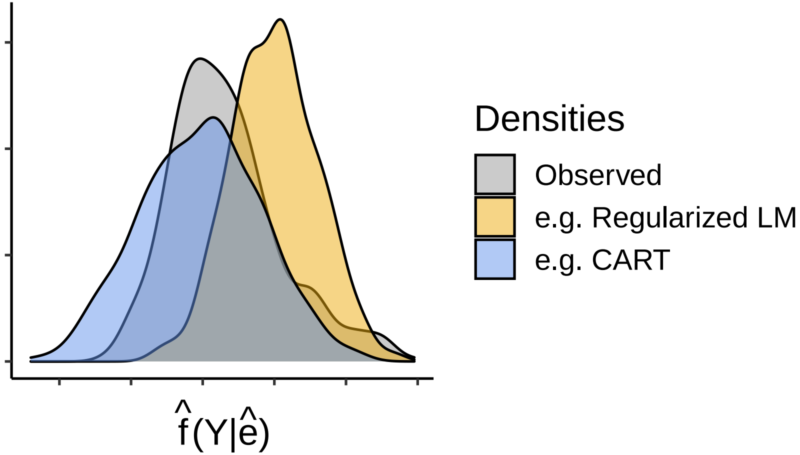
$$\hat{Y}_m|\mathbf{X}, R = 0$$

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

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

# Automated Method Selection - Criterion 1 (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 $H_m$)

# Automated Method Selection - Criterion 2

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

# Proposed Solution

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

New:

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

# 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 (Criterion 1)
        - Conditional mean (Criterion 2)
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

# 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)

# Outline

- Proposed solution ✓
- Small scale simulation
- Large scale simulation

# Small Scale Simulation - Setup

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

# Small Scale Simulation - Data Generation

1. Draw values of $Z$: $Z \sim N(0,1)$
2. Draw values of $X|Z : X \sim N(\alpha_0 + \alpha_1 Z, \sigma_X^2)$
3. Draw values of $Y|Z, X : Y \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma_Y^2)$
4. Generating response indicators $R_Z$ and $R_X$:

   a)
   $$p_X = logit^{-1}(\delta_0^X + \delta_1^X Y),\ p_Z = logit^{-1}(\delta_0^Z + \delta_1^Z X)$$

   b)
   $$R_Z = \begin{cases} 1 \text{ for } p_Z \geq u_Z, \\ 0 \text{ for } p_Z < u_Z \end{cases} \quad R_X = \begin{cases} 1 \text{ for } p_X \geq u_X, \\ 0 \text{ for } p_X < u_X \end{cases}$$

   with $u_Z, u_X \sim Unif(0,1)$.

# Small Scale Simulation - Parameters

$\alpha_0 = 0$, $\alpha_1 = 0.25$, $\sigma_X^2 = 1$

$\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\sigma_Y^2 = 1$

For response indicators $R_Z$ and $R_X$:

$\delta_0^X = \delta_0^Z = 0.7$

$\delta_1^X = -2$, $\delta_1^Z = 0.7$

$\Rightarrow$ Missing at random (MAR) situation

Varying Parameter:

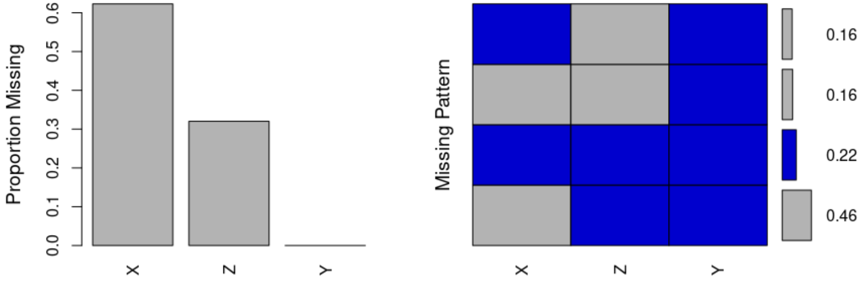
Number of observations: 1.000, 5.000

Other parameters:

Number of iterations: 5

Number of multiply imputed data sets: 5

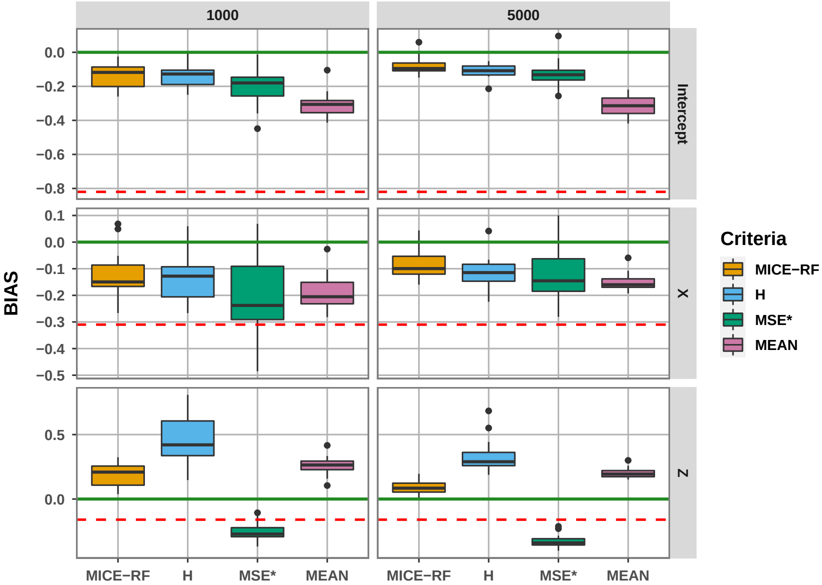
# Small Scale Simulation - Missing Data Pattern



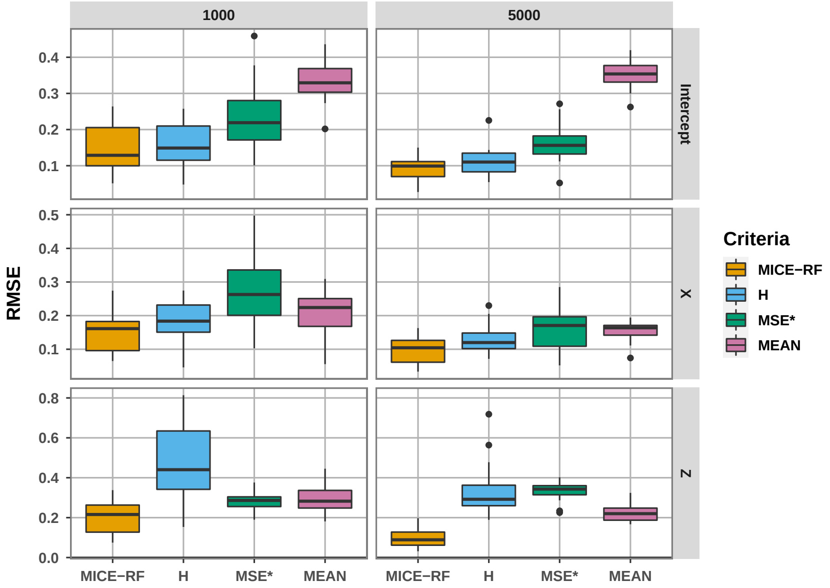

Coefficients of $Y \sim X + Z$:

| | $\beta_0$ | $\beta_X$ | $\beta_Z$ |
|---|---|---|---|
| Original Data | 1 | 1 | 0.5 |
| Complete Cases | 0.18 | 0.69 | 0.34 |

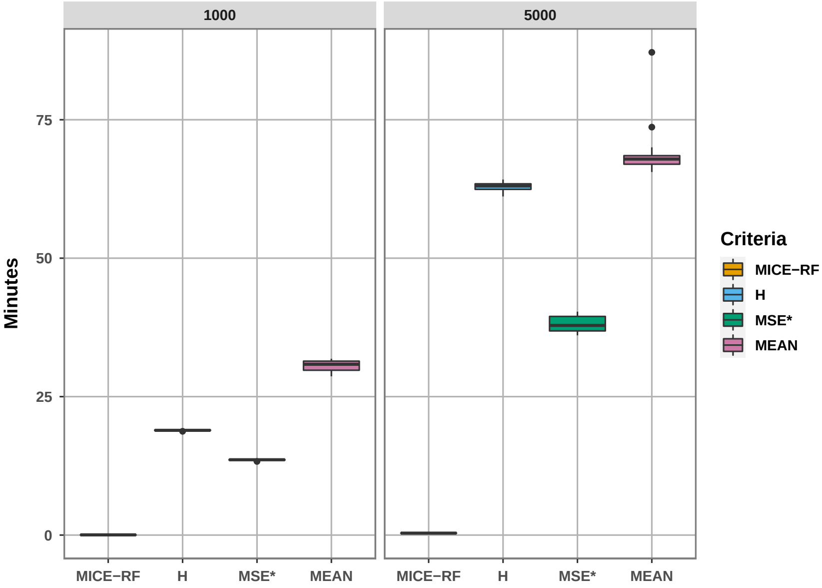
# Results - Bias



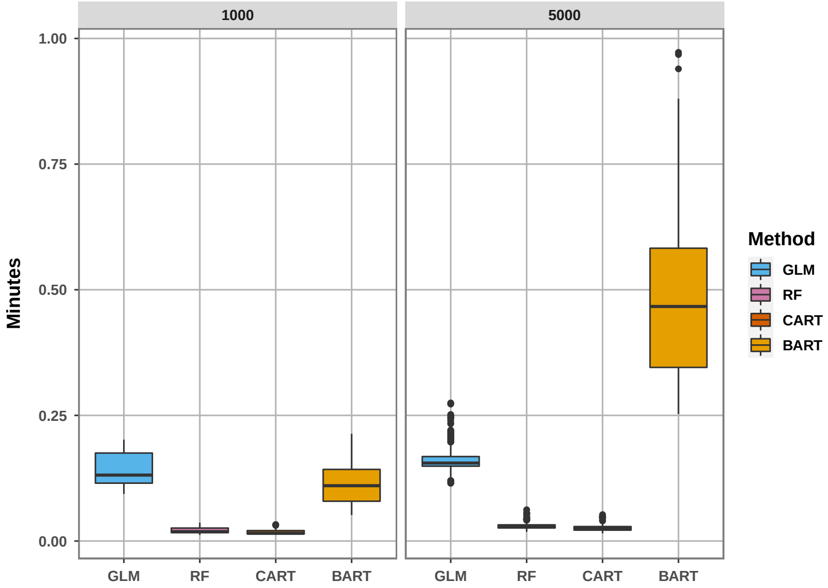
# Results - RMSE



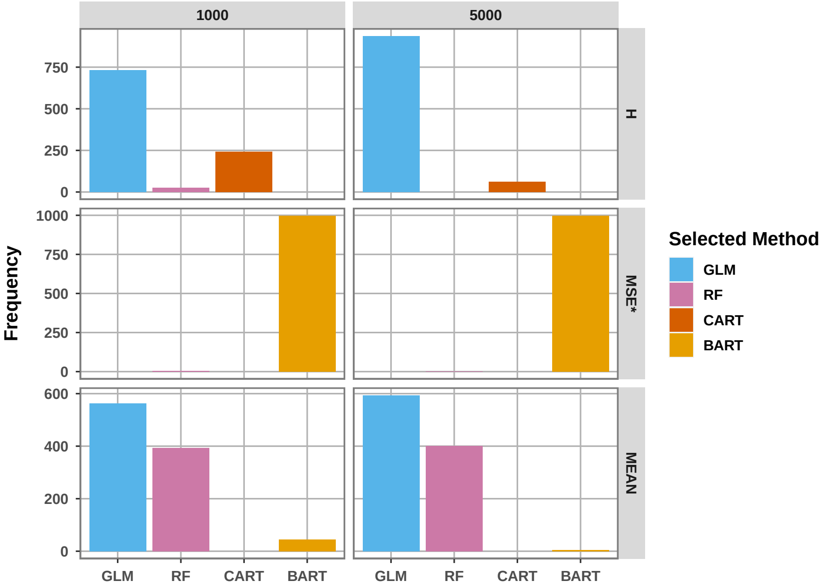
# Results - Runtime



# Results - Runtime (cont.)



# Results - Selected Methods



## Results - Discussion

Bias: reduced but not zero

- More iterations
- Initially imputed values
- Compare implementations in SIIMS and MICE

Runtime: still relatively high

- BART and GLM are bottle necks

# Outline

- Proposed solution ✓
- Small scale simulation ✓
- Large scale simulation

# Large Scale Simulation - Based on Real Data Set

Why real data?

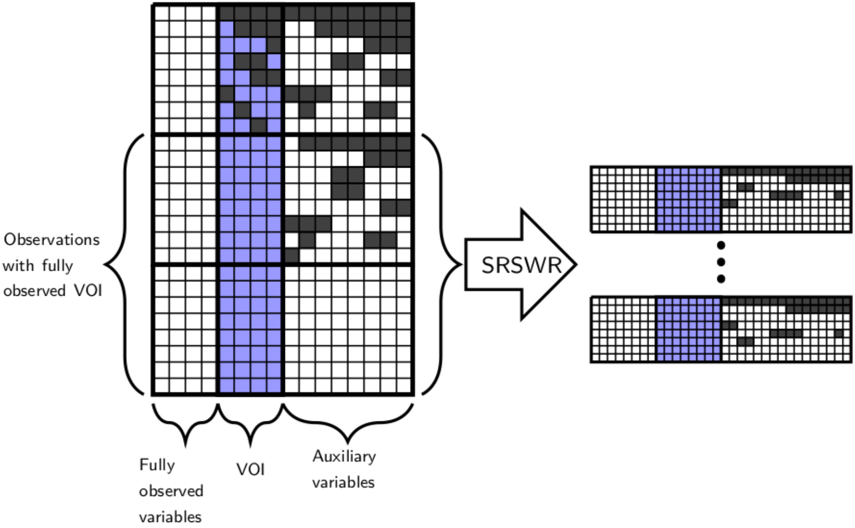
- Imputation procedures sensitive to data generating process

What data set?

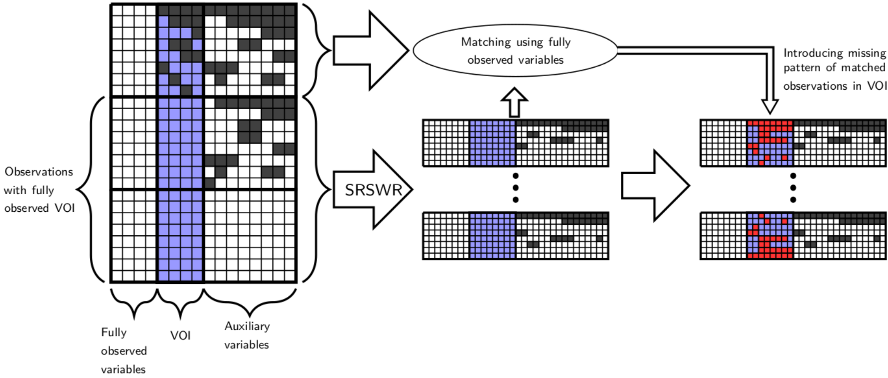
National Health and Nutrition Examination Survey (NHANES) data

- 5 waves collected 1999 - 2016
- Variables: questionnaire data, dietary data (diary), physical examination data (mobile examination center)
- Missing values: blockwise + item nonresponse

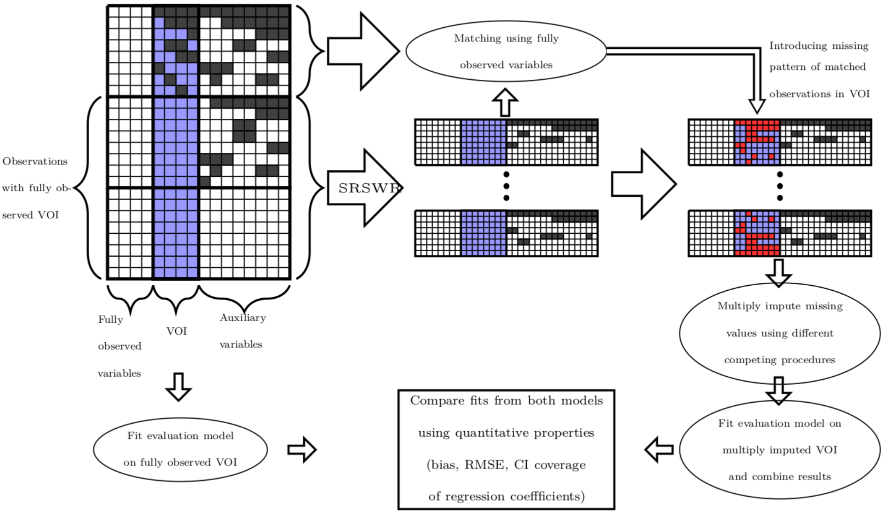
# Large Scale Simulation - Assessment Process (adapted from Ezzati-Rice et al. 1995)



# Large Scale Simulation - Assessment Process - Step 2



# Large Scale Simulation - Assessment Process - Step 3




**Legend:**

SRSWR: simple random sample with replacement; Observed values: □ ; Naturally missing values: ■; Variables of interest (VOI): ■ ; Introduced missing values: ■

# Large Scale Simulation - Variables of Interest (VOI)

Selection criteria

1. Relationship: approximately linear, i.e. a linear model can be fit
2. Missing values: mostly incomplete, to introduce missing data patterns (following Ezzati-Rice et al. 1995).
3. Data collection: different modes of data collection (different missing data patterns)
4. Population: not target a sub-population (e.g. smokers), to avoid "not applicable" cases.
5. Wave: measured in NHANES wave 2015/16
6. Missing values should rather be in predictors than in outcomes for improved $\hat{\beta}$-coefficients after MI (Little 1992).

Problem: most papers use variables with missing values as outcomes and control for (almost) completely observed variables (like social-demographics).

# Large Scale Simulation - Identified VOIs

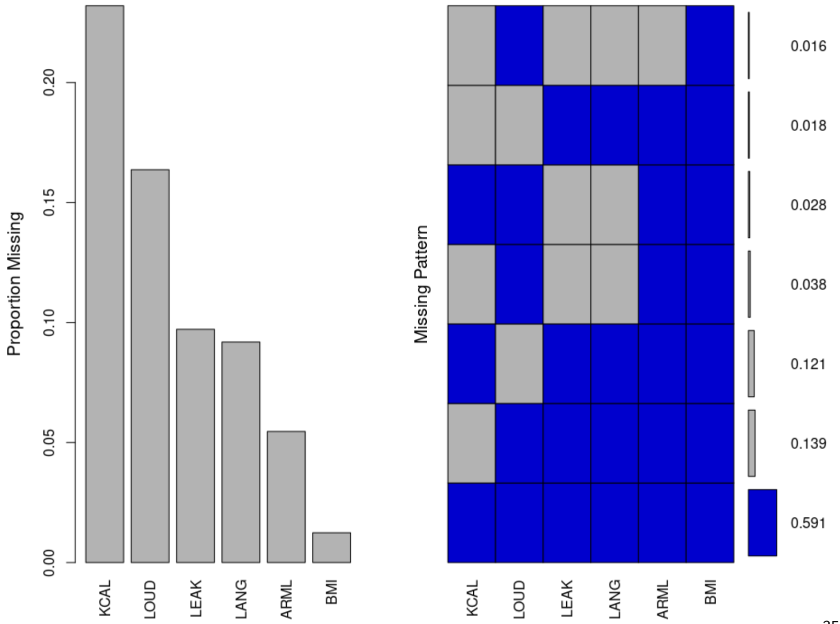
Outcome:

BMI (continuous, log-transformed) - mode: physical examination

Covariates:

- Kilo-calories intake (KCAL) (continuous) - mode: nutrition diary
- Language of the physical examination interview (LANG) (binary, English vs not English) - mode: physical examination
- Leak urine during physical activities (LEAK) (binary, yes, no) - mode: physical examination
- Upper arm length (ARML, continuous) - mode: physical examination
- Loud noise exposure (LOUD) (binary, yes, no) - mode: questionnaire

# Large Scale Simulation - Missing Data Patterns



# Large Scale Simulation - Expectations

Results:

- Binary vs. continuous variables
- Upstream variable selection on quantitative properties and runtime
- Double robust property on quantitative properties

# Outline

- Proposed solution ✓
- Small scale simulation ✓
- Large scale simulation ✓

# Next Steps

1. Increase Speed
2. Simulation on high-dimensional data
3. Compare procedures in SIIMS and MICE

Thank you for your attention!

Any questions?

michaf@umich.edu

# Appendix

# Used Methods - details

Bayesian (G)LM (glmnet, rstan):

- Parameters tuned: elastic net mixing parameter (5-fold cross-validation)
- Parameters specified: default of R package “glmnet”
- Imputed data: draws from posterior predictive distribution

CART (rpart):

- Parameters tuned: none
- Parameters specified: min. number of observations in terminal node = 5 (MICE default)
- Imputed data: draws within terminal nodes

# Used Methods - details (cont.)

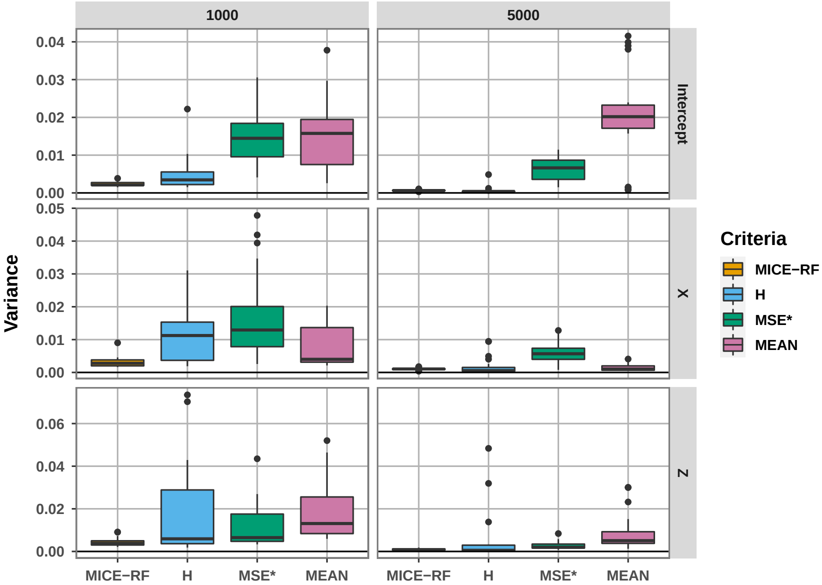
Random Forest (randomForest):

- Parameters tuned: none
- Parameters specified: number of trees = 20, min. number of observations in terminal node = 5 (MICE default)
- Imputed data: draws from normal distribution, mean and standard deviation estimated from predictions of single trees

BART (bartMachine):

- Parameters tuned: none
- Parameters specified: number of trees = 50 (following Kapelner and Bleich (2013))
- Imputed data: draws from posterior predictive distribution

# Results - Variance



# 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.

Kapelner, Adam, and Justin Bleich. 2013. "BartMachine: Machine Learning with Bayesian Additive Regression Trees." *arXiv Preprint arXiv:1312.2171*.

Little, Roderick. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87 (420): 1227–37.

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.