Sequential Imputation of Missing Data in High-Dimensional Data Sets

A Model Selection Approach

Micha Fischer
University of Michigan
Program in Survey Methodology

BigSurv18, Barcelona, Spain

October 27, 2018

Problem

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

$\Rightarrow$ Usual approach: multiple sequential imputation

Why?

Standard procedure needs specified model for each incomplete variable

- Subjectivity: model specification
- Efficiency: limited resources

Research Question

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

How?

Sequential imputation:

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

Within sequential imputation procedure:

- Automated model specification
- Automated model selection
  - Assessing models by an automated version of a visual approach by Bondarenko and Raghunathan (2016)
- Advantages:
  - Many different model types possible
  - Objective procedure

Automated Model Specification

Focus here:

- parametric models (Bayesian LM, Bayesian GLM)
  - Use basis expansion of covariates
  - Perform adaptive LASSO to determine model formula
- nonparametric models (CART)
  - no explicite formula necessary
  - all covariates are used

Visual Approach (Bondarenko and Raghunathan 2016)

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

$$\hat{e}_k = P(R_k = 1|\mathbf{X})$$

$$R_k = \begin{cases} 1 \text{ if } Y_k \text{ observed}, \\ 0 \text{ if } Y_k \text{ missing} \end{cases}$$

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

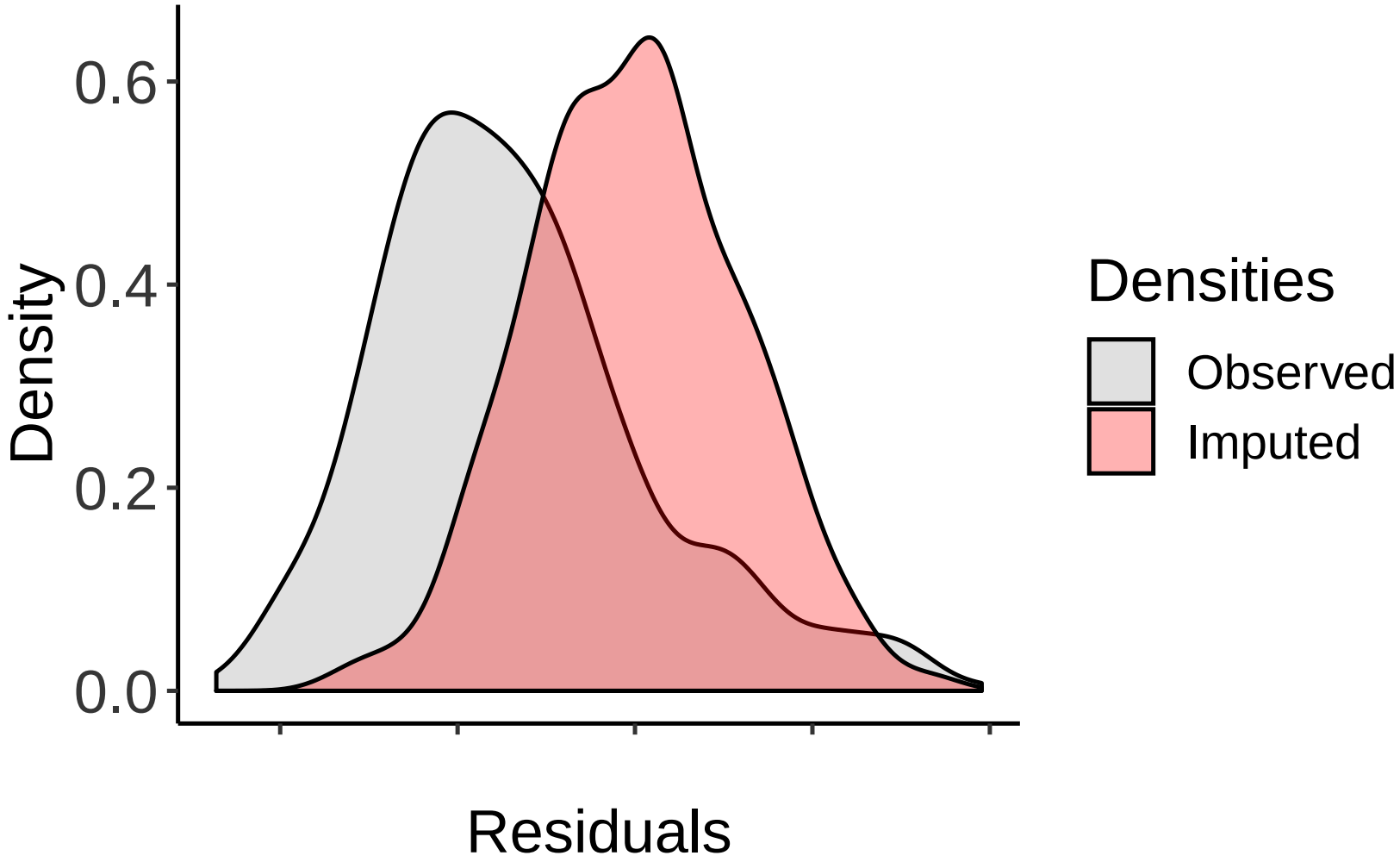
$$\hat{f}(Y_k|\hat{e}_k, R_k = 1)$$

3. Fit imputation model and predict missing values $\hat{Y}_k|\mathbf{X}, R_k = 0$
4. Estimate residual density for imputed values conditional on propensity score:

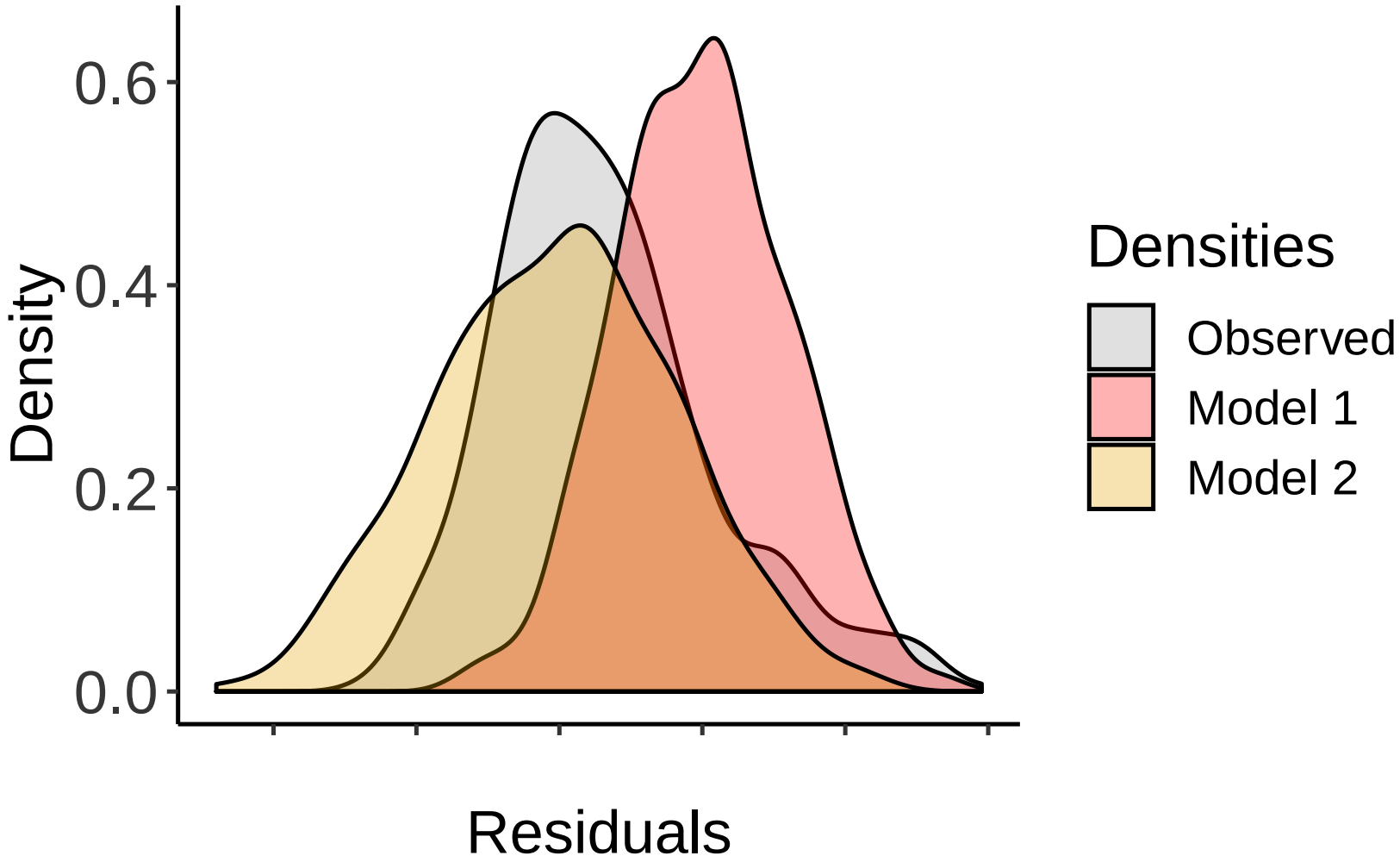
$$\hat{f}(\hat{Y}_k|\hat{e}_k, R_k = 0)$$

Visual Approach II

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



Visual Approach III



$\Rightarrow$ Automation: comparing via measure of similarity

Measure of Similarity

Here: Hellinger's distance (e.g. Van der Vaart 1988, 211–12) for each model m

$$H_m(\hat{f}(Y_k|\hat{e}_k, R_k = 1), \hat{f}(\hat{Y}_{k,m}|\hat{e}_k, R_k = 0)) = \sqrt{1 - \int \sqrt{\hat{f}(Y_k|\hat{e}_k, R_k = 1)\hat{f}(\hat{Y}_{k,m}|\hat{e}_k, R_k = 0)}\,\mathrm{d}Y_k}.$$

$H_m(.,.) \in [0, 1]$

Other distance measures could be used as well.

Model Selection within Sequential Imputation

For each iteration:

1. Estimate response propensity score based on all other variables
2. Estimate density of observed values conditional on propensity score
3. For each model specification m:
   - Fit model using all covariates
   - Predict plausible values for the missing values using the model
   - Estimate density of plausible values conditional on propensity score
   - Estimate Hellinger distance between densities
4. Select model specification with minimal Hellinger distance and update imputed values
5. Repeat 1 - 4 of all variables with missing values

Preliminary Simulation

Comparing different techniques for multiple sequential imputation:

1. Bayesian linear regression models
2. Random forest
3. Model selection approach with Bayesian linear regression model, Bayesian generalized linear regression model with log link (for skewed outcome distributions), CART

Data Generation

1. Draw values of X:

$$X \sim N(0,1)$$

2. Draw values of outcome variables:

$$Y_1|X : log(Y_1) \sim N(\alpha_0 + \alpha_1 X + \alpha_2 X^2, \sigma^2_{Y_1})$$

$$Y_2|X, Y_1 : log(Y_2) \sim N(\beta_0 + \beta_1 X + \beta_2 log(Y_1) + \beta_3 X Y_1, \sigma^2_{Y_2})$$

3. Generating response indicators R_1 and R_2:

3.1

$$p_1 = logit^{-1}(\delta^1_0 + \delta^1_1 X)$$
$$p_2 = logit^{-1}(\delta^2_0 + \delta^2_1 X)$$

3.2

$$R_1 = \begin{cases} 1 \text{ for } p_1 \geq u_1, \\ 0 \text{ for } p_1 < u_1 \end{cases}$$

$$R_2 = \begin{cases} 1 \text{ for } p_2 \geq u_2, \\ 0 \text{ for } p_2 < u_2 \end{cases}$$

with $u_1, u_2 \sim Unif(0,1)$.

Simulation Parameters

Fixed parameters:

For $log(Y_1)$:

$\alpha_0 = 0$, $\alpha_1 = 0.25$, $\alpha_2 = 0.25$, $\sigma^2_{Y_1} = 1$

For $log(Y_2)$:

$\beta_0 = -1$, $\beta_1 = 0.25$, $\beta_2 = 0.25$, $\sigma^2_{Y_2} = 1$

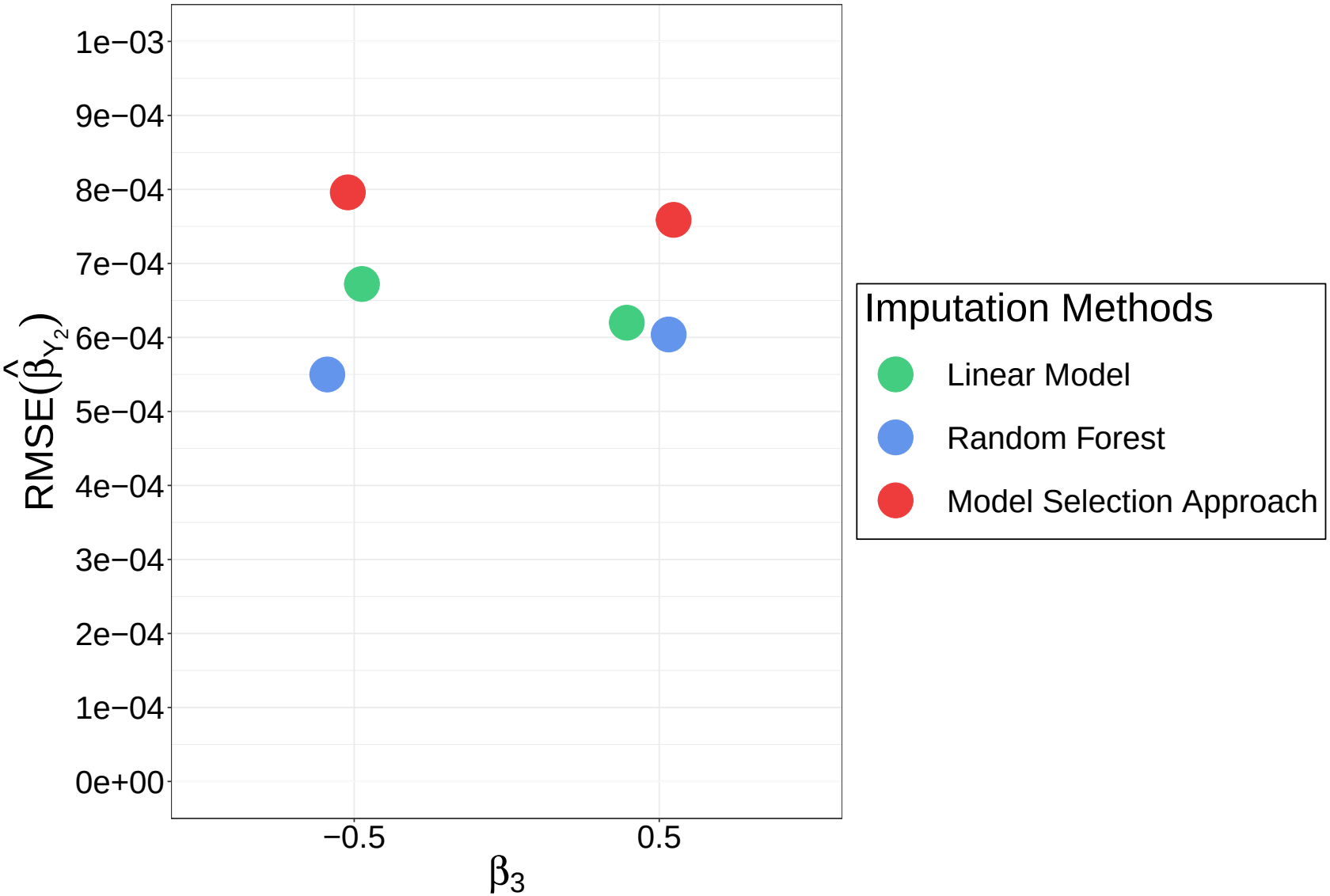
For response indicators R_1 and R_2:

$\delta^1_0 = \delta^2_0 = 0$
$\delta^1_1 = \delta^2_1 = 1$

Varying Parameter:
$\beta_3 \in \{-0.5, 0.5\}$

$\Rightarrow$ MAR situation

Preliminary Results



Limitations & Extensions

1. Simulation on higher dimensional data sets
2. Evaluation on survey data linked to administrative records
3. Currently only for incomplete continuous variables
   $\Rightarrow$ Bondarenko and Raghunathan (2016) provide also tools for binary variables
4. Based on MAR assumption
   $\Rightarrow$ Sensitivity analysis can provide more insights

Thank you for your attention!

Any questions?

michaf@umich.edu

Appendix - notation I

- $\mathbf{X}$ be a set of fully observed variables
- $\mathbf{Y} = Y_1, \ldots, Y_K$ be a set of continuous variables containing missing values
- $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$ is data set with $i = 1, \ldots, n$ observations
- R_k denote the vector of response indicators for variable Y_k,
- $Y_k|R_k = 1$ be the subset of observed values and $Y_k|R_k = 0$ be the subset of missing values for variable Y_k
- Y_k^j denote the variable Y_k at iteration j
- $\mathbf{Y}_{-k}^j$ $(k \in \{1, \ldots, K\})$ denote the set of variables $Y_1^j, \ldots, Y_{k-1}^j, Y_{k+1}^{j-1}, \ldots, Y_K^{j-1}$ where variable Y_k is excluded
- $m \in \{1, \ldots, M\}$ be an imputation model in a pool of models of size M
- $Y_{k,m}^j|R_k = 0$ be the values replacing $Y_k^{j-1}|R_k = 0$, predicted by model m in iteration j

Appendix - notation II

- $f(Y_k|R_k = 1)$ and $f(Y_k|R_k = 0)$ denote the densities of observed and missing values for variable Y_k
- $e_k = P(R_k = 1|\mathbf{X}, \mathbf{Y}^j_{-k})$ be the propensity score of all n values for a response of variable k based on all other variables $\mathbf{X}$ and $\mathbf{Y}^j_{-k}$
- $f(Y_k|e_k, R_k = 1)$ and $f(Y_k|e_k, R_k = 0)$ define the densities of residuals for Y_k regressed on e_k for the observed $(R = 1)$ and unobserved $(R = 0)$ values
- $H_m(f(Y_k|R_k = 1), f(Y_k|R_k = 0))$ defines Hellinger's distance - quantifying the similarity of $f(Y_k|R_k = 1)$ and $f(Y_{k,m}|R_k = 0))$
- All estimates based on data be denoted by "$\hat{}$" on top of the estimated quantities.

Appendix - algorithm for sequential imputation I

For an iteration $j > 1$ the following steps will be performed:

1. Repeat for all $k \in \{1, \ldots, K\}$ variables containing missing values:
   - 1.1 Estimate $\hat{e}_k = P(R_k = 1 | \mathbf{X}, \mathbf{Y}^j_{-k})$ for all n values in Y_k
   - 1.2 Estimate $\hat{f}(Y_k | e_k, R_k = 1)$ using kernel density estimation
   - 1.3 Repeat for all $m \in \{1, \ldots, M\}$ potential imputation models:
     - Fit model m with Y_k^{j-1} as the dependent variable and $\mathbf{X}$ and $\mathbf{Y}^j_{-k}$ as the independent variables
     - Predict plausible values $Y^j_{k,m} | R_k = 0$ for $Y_k | R_k = 0$ using model m
     - Estimate $\hat{f}(Y^j_{k,m} | e_k, R_k = 0)$ using kernel density estimation
     - Estimate Hellinger distance $\hat{H}_m = H(\hat{f}(Y_k | e_k, R_k = 1), \hat{f}(Y^j_{k,m} | e_k, R_k = 0))$
   - 1.4 Select model

$$m_{opt} = \min_m \hat{H}_m$$

and use $Y^j_{k,m_{opt}} | R_k = 0$ to update $Y_k^{j-1} | R_k = 0$

Appendix - algorithm for sequential smputation II

2. Repeat step 1) J times or until convergence, i.e. $|(Y_{k,i}^{j}|R_{k,i}=0)-(Y_{k,i}^{j-1}|R_{k,i}=0)|<c_k, \forall_k, \forall_i$ with $c_k>0$, and use imputed values from the last iteration to receive one imputed data set.
3. Repeat steps 1)-2) ℓ times to receive ℓ multiply imputed data sets.

Appendix - pool of imputation models

- Bayesian linear models
- Bayesian generalized linear models
- Regression trees based on bootstrap samples

Appendix - modification 1 - rejection of samples

Only one model is used, values can be rejected or accepted, based on a threshold H_0.

1. Repeat for all $k \in \{1, \ldots, K\}$ variables containing missing values:
   - 1.1 Estimate response propensity scores $\hat{e}_k = P(R_k = 1 | \mathbf{X}, \mathbf{Y}^j_{-k})$ for all n values in Y_k
   - 1.2 Estimate $\hat{f}(Y_k | e_k, R_k = 1)$ (the density of residuals for Y_k regressed on $\hat{e}_k$ for observed values) using kernel density estimation.
   - 1.3 Repeat until $\hat{H} < H_0$:
     - Fit new model with Y_k^{j-1} as the dependent variable and $\mathbf{X}$ and $\mathbf{Y}^j_{-k}$ as the independent variables
     - Predict plausible values $Y_k^j | R_k = 0$ for $Y_k | R_k = 0$ using new model
     - Estimate density of residuals for Y_k regressed on $\hat{e}_k$ for imputed values ($\hat{f}(Y_k^j | e_k, R_k = 0)$) using kernel density estimation
     - Estimate Hellinger distance $\hat{H} = H(\hat{f}(Y_k | e_k, R_k = 1), \hat{f}(Y_k^j | e_k, R_k = 0))$
     - Compare $\hat{H}$ with H_0

Appendix - modification 2 - editing of sampled values

The modified values can be computed by:

$$(Y_k^{j*}|R_k = 0) = \frac{(Y_k^j|R_k = 0) - \hat{d}_k^j}{\hat{R}_k^j}$$

with

$$\hat{d}_k^j = \hat{\mu}_{0,k}^j - \hat{\mu}_{1,k}$$

denoting the distance between means ($\hat{\mu}_{0,k}^j$ and $\hat{\mu}_{1,k}$) of $\hat{f}(Y_k^j|e_k, R_k = 0)$ and $\hat{f}(Y_k|e_k, R_k = 1)$ and

$$\hat{R}_k^j = \frac{\hat{S}_{0,k}^j}{\hat{S}_{1,k}}$$

denoting the ratio of estimated standard deviations ($\hat{S}_{0,k}^j$ and $\hat{S}_{1,k}$) of $\hat{f}(Y_k^j|e_k, R_k = 0)$ and $\hat{f}(Y_k|e_k, R_k = 1)$.

Appendix - splines of principal components and propensity score as covariates - for continuous variables

Instead of using all covariates $\mathbf{X}, \mathbf{Y}^{j}_{-k}$ directly

1. Estimate residuals of covariates regressed on the propensity score:

$$\mathbf{X} \sim \hat{e_k} \Rightarrow \mathbf{X}^*$$
$$\mathbf{Y}^{j}_{-k} \sim \hat{e_k} \Rightarrow \mathbf{Y}^{j*}_{-k}$$

2. Estimate principal components $\widehat{\mathbf{P}}^*$ of $\mathbf{X}^*, \mathbf{Y}^{j*}_{-k}$
3. Use spline function of propensity score $s(\hat{e}_k)$ and most important principal components $s(\widehat{\mathbf{P}}^*)$ as covariates in imputation model:

$$Y_k \sim s(\hat{e}_k) + s(\widehat{\mathbf{P}}^*)$$

$\Rightarrow$ No colinearity in covariates, only main effects necessary, reduced dimensions, highly flexible

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). Wiley Online Library: 3007–20.

Van der Vaart, Aad W. 1988. "Asymptotic Statistics." Cambridge University Press.