Comparison of Advanced Methods for Data Imputation in the Context of IRT: A Monte Carlo Simulation Study

Julianne M. Edwards
W. Holmes Finch
Missing Data

- Missing data is a common problem faced by researchers using item response data.

- Examinees might leave items unanswered due to an oversight, because they don’t know the answer, or because they were unable to reach the end of the exam.

- Missing data is typically categorized into one of three types:
  1. Missing Completely at Random (MCAR) – No systematic cause.
  2. Missing at Random (MAR) – Probability of being missing is related to a measured variable.
  3. Missing Not at Random (MNAR) – Probability of being missing is related to the missing value itself.
Methods for Handling Missing Data

- Listwise Deletion (LD) is the common default approach. Completely remove individuals with missing responses from the sample.

- Treat missing item response as fully incorrect, or partially incorrect.

- Replace the missing data with a single value, such as the mean or a prediction from a regression model.

- LD, treating as incorrect, and most common single imputation methods are associated with bias in parameter estimates, and incorrect standard errors (de Ayala et al., 2001; Finch, 2008; Lord, 1974; Ludlow and O'leary, 1999; Rose et al., 2010).
Multiple Imputation

- Multiple Imputation (MI) through data augmentation replaces missing data with model predicted values that are adjusted with a small amount of randomness.

- Several different imputed datasets are created and then subjected to the analysis of interest (e.g. IRT).

- The parameter estimates from the multiple analyses are then combined with one another, as are the standard errors.
Multiple Imputation

- Although MI has been shown to produce unbiased parameter and standard error estimates for normally distributed variables, it assumes a common joint probability distribution (typically multivariate normal) for all of the variables of interest.

- This assumption is not tenable for most item responses, which are typically dichotomous or ordinal in nature.

- MI models for purely categorical data are based upon the multinomial distribution using loglinear models, and are limited in terms of the size of the dataset to which they can be applied.
Prior research on missing data in the context of IRT modeling

- Prior research has shown that missing data has a deleterious impact on IRT parameter estimation (de Ayala et al., 2001; Finch, 2008; Lord, 1974; Wolkowitz & Skorupski, 2013).

- Of the available methods, treating missing values as incorrect appears to yield the most biased estimates of item difficulty and person ability (Lord, 1974).

- The use of MI under the assumption of multivariate normality has been associated with the least amount of IRT estimation bias, though in many cases bias remains (Finch, 2008).
Alternatives to standard MI

- Given the bias associated with commonly used approaches to dealing with missing data in the context of IRT, alternative methods need to be investigated.

- Several alternative methods for data imputation with nonnormal and categorical data have been developed over the past few years.

- These methods have been demonstrated to be effective in a number of situations, but have not yet been explored in the context of IRT parameter estimation.
Alternatives to standard MI

- Random Forest Imputation (RF)
- Multivariate Imputation by Chained Equations (MICE)
- Mice with RF (MICE-RF)
- MICE with recursive partitioning tree imputation (MICE-RPT)
Recursive Partitioning Trees

- RPT, also known as classification and regression trees (CART), develops prediction models using repeated partitioning of the data into ever more homogeneous groups (nodes) based on the dependent variable.

- In the context of data imputation, the outcome variable is the one with missing values (e.g. missing response to an item), and the other variables (e.g. items) are used to develop a prediction model for it.

- A prediction model is first developed, and then for each individual with a missing value the prediction yielded by the tree, given the values of the predictor variables, is imputed for the missing data point.

- RPT makes no assumptions about the distribution of the variables, unlike MI.
Recursive Partitioning Trees

Figure 1: Example Recursive Partitioning Tree
Random Forest

- RF involves growing many (e.g. 100) RPTs using bootstrap samples of the individual observations, as well as the predictors.

- For each sample, a tree is grown and the imputed value for an individual is obtained by averaging imputations across the individual trees.

- For a dichotomous item response, this average is simply the proportion of RPTs for which the respondent is predicted to answer the item correctly.

- If this value exceeds 0.5, then the imputed value is correct (1), otherwise it is incorrect (0).
RF Imputation

1. Obtain an initial estimate for missing values among all of the variables in the dataset using a simple imputation approach such as mean imputation.

2. Sort the variables in the data set in terms of the amount of missing data, from least to most. Imputation is then conducted based on the amount of missing data, with those having the least amount of missing data serving as target variables first.

3. Fit RF for the target item using the other items as the set of predictors.

4. Use the resulting model from step 3 to predict values for target item using other items as predictors.

5. Repeat steps 3-4 for each item that contains missing observations.

6. Check stopping criterion, and repeat steps 3-5 until stopping criterion is met.

7. Stop the algorithm when the difference between the previous and current imputations increases in value.
RF imputation

- RF produces a single imputed value for each missing data point.

- When the stopping criterion has been satisfied, the researcher takes the resulting dataset and conducts the analysis of interest (e.g. IRT).

- For categorical variables such as item responses, the stopping criterion is the proportion of incorrectly classified values for individuals with no missing data on the variable.

- This value should be near 0, with a variety of cut off values recommended.
Multivariate Imputation by Chained Equations

- Unlike MI with data augmentation, MICE assumes that each variable can have a unique probability distribution, rather than imposing a common one across the dataset.

- Thus, continuous variables are modeled using ordinary least squares regression, dichotomous variables with logistic regression, and count variables with Poisson regression.
**MICE**

1. Replace each missing value in the dataset with a simple imputation, such as a random draw from the sample.

2. These placeholder imputations are then set back to missing for a target variable, $y$.

3. $y$ for the nonmissing observations is regressed on the other variables in the data (or any other variables the researcher would like to use for imputation), using an appropriate model. For example, binary logistic regression is used for dichotomous variables, polytomous logistic regression for ordinal variables, poisson regression for count variables, and normal theory ordinary least squares regression for continuous variables.

4. Once the appropriate regression model is fit, missing values for $y$ are replaced by random draws, using the Gibbs sampler, from the conditional probability distribution associated with the regression model from step 3. In other words, imputations for the target variable $y_{1+1}^t$ are drawn from the probability distribution defined by the regression model as $P(y_1^t | X_2^t, X_3^t, \ldots, X_k^t)$, where the $X^t$ are the other (nontarget) variables in the dataset used to fit the model, and the imputed value of $y$ is drawn from the appropriate (e.g. normal) probability distribution conditioned on the predictors in the regression model.

5. Steps 2-4 are repeated for each variable in the dataset containing missing data. Completing the initial imputations for all variables in the dataset constitutes a single iteration.

6. Steps 2-5 are then repeated for a given number of iterations (e.g. 40) until the imputations have converged over the iterations.

7. The entire set of steps is repeated $m$ times, where $m$ is the desired number of multiple imputations (e.g. 10). Imputed values for each missing data point are saved from each of the $m$ imputations.
MICE-RF and MICE-RPT

- The MICE algorithm can be used with any modeling approach, including both RF and RPT.

- Thus, rather than using linear models such as OLS regression and logistic regression, either RF or RPT are used, allowing for the inclusion of nonlinear relationships among the variables when developing imputations.

- Imputations are taken as random draws from those in the same terminal node as the individual with missing data.
The goal of this study was to compare the performance of RF, MICE, MICE-RF, and MICE-RPT with one another in the context of missing data and IRT, using a simulation study design.

The primary focus is on recovery of item difficulty and discrimination parameter values.
Method

- Simulation study with 1000 replications per combination of conditions was used to compare the missing data methods.

- Data were generated using the 2-parameter logistic (2PL) model.

- Outcome variables of interest were:
  - Estimation bias for item difficulty and discrimination parameters.
  - Mean square error (MSE) in item parameter estimates.
Method

- Manipulated simulation conditions:
  - Sample size: 1000, 2000
  - Number of items: 20, 30
  - Percent missing data: 5%, 15%, 30%, 50%
  - Type of missing data: MCAR, MAR
  - Missing data method: LD, MI, RF, MICE, MICE-RF, MICE-RCP
Method

- For MICE, MICE-RPT, and MICE-RF 10 imputed datasets with 40 iterations each were generated (van Buuren & Groothuis-Oudshoorn, 2011).

- For MI, 10 imputed datasets were generated, and imputed values were rounded to the nearest integer (0 or 1).

- For RF, the bootstrap sample size was 75% of the full sample, and 50% of the predictors were used in constructing each tree.

- The RF solutions consisted of 100 trees each.

- ANOVA was used to identify the significant main effects and interactions that were related to estimation bias.
Item Difficulty Estimation Bias: MCAR
Item Difficulty MSE: MCAR
Item Discrimination Estimation Bias: MCAR
Item Discrimination MSE: MCAR
Item Difficulty Estimation Bias: MAR
Item Difficulty MSE: MAR
Item Discrimination Estimation Bias: MAR
Item Discrimination MSE: MAR
Discussion

- When data were MCAR, estimation bias for item difficulty increased concomitantly with increasing amounts of missing data for all methods.

- Difficulty estimation bias was low for MCAR data across all methods, and lower for 30 items.

- The highest item difficulty estimation bias under MCAR was for LD and RF.
Discussion

- Item discrimination parameter estimation bias was not related to the percent of missing data under the MCAR condition.

- As with item difficulty, the overall level of discrimination bias was low across methods, where the highest values were associated with the random forest based approaches.
Discussion

- Estimation bias for both item difficulty and discrimination was greater in the MAR condition than for MCAR data.

- Difficulty bias for the MAR condition was highest for LD, followed by MI and MICE-RF.

- Discrimination bias with MAR data was greatest for RF and MICE-RF.
Conclusions

- Across conditions, MICE and MICE-RPT yielded IRT item parameter estimates with relatively low levels of bias, making them optimal general use tools for researchers faced with missing data in the context of IRT.

- These results match earlier findings that MICE-RPT frequently yields the least biased results for other modeling contexts.

- As prior research has shown, using listwise deletion with IRT models is not recommended when the amount of missing data is 15% or greater.

- Multiple imputation was associated with elevated item difficulty estimation bias for MAR data, particularly when compared to MICE and MICE-RPT.
References


