Tree-based Multiple Imputation Methods

Demonstration that MissForest compared to MI FCS variants does not comply with the goal of statistical inference using a proper simulation study

Tahomina Akter Ety, Tipu Sultan and Lara Krell

University of Bamberg, University of Trier

1. Motivation

- Multiple imputations (MI) are statistical techniques that generate multiple imputed datasets to account for the uncertainty introduced by missing data.
- Modern data acquisition based on advanced technologies is often confronted with the problem of missing data, and commonly used algorithms often rely on a complete data set. Therefore, knowing suitable imputation methods for these cases is essential.
- This poster aims to compare MissForest and MI FCS variants in terms of the Bias, RMSE, and Coverage and come to a reasonable decision.

2. Fundamental Theory of MI FCS

The Fully Conditional Specification (FCS) is a specific method within MI [6], [3], [4].

- FCS uses a series of regression models to impute missing values, modeling each variable with its own imputation model.
- The regression models used in FCS are usually specified as linear, logistic or ordinal regression models.

The FCS algorithm includes the following steps [5], [6]:

1. Define the missing data structure (e.g., missing completely at random, missing at random, etc.).
2. Specify the imputation models for each variable with missing data.
3. Impute missing values using the imputation models.
4. Repeat steps 2 and 3 several times to generate multiple imputed datasets.
5. Analyze each imputed dataset and combine results using appropriate methods (e.g., Rubins rules).
6. Evaluate the performance of the imputation models and adjust if necessary.

3. Fundamental Theory of MissForest

- First proposed by Stekhoven and Bühlmann (2012) [5]. It is an iterative imputation method based on the Breiman's Random Forest algorithm [2].
- MissForest is applicable for mixed-type data (missings in numeric and categorical variables) with nonlinearity and interaction effects.
- MissForest considers multiple imputation by averaging numerous unpruned classification or regression trees [1].
- Unlike MICE, it makes no assumptions about the relationship between the features.

The MissForest algorithm includes the following steps [7], [5]:

1. Make an initial guess for all missing values (e.g., mean, mode), sorted in ascending order.
2. The observations in the dataset are divided into two parts: the training set (observed observations) and the prediction set (missing observations).
3. MissForest fits a random forest on the observed part and then predicts the missing part for each variable.
4. These steps are repeated until the stopping requirement is met or the maximum number of iterations is reached.

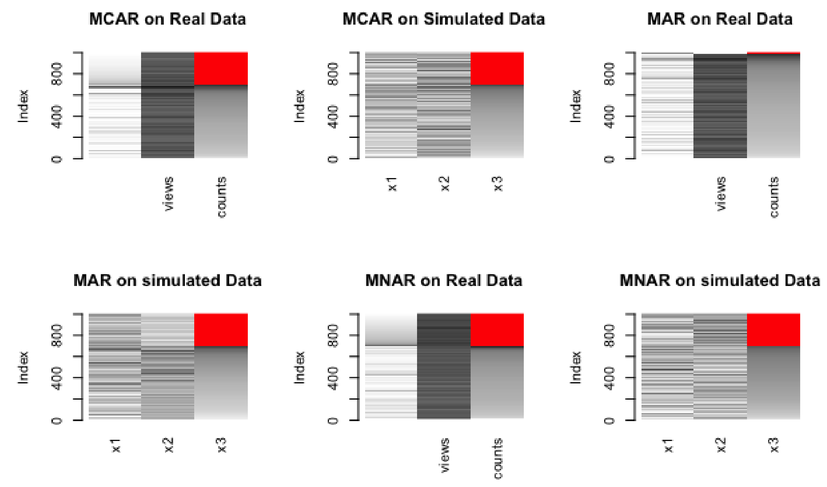
4. Simulation Study and Data

Real Data:

- Data of the top 1000 YouTubers based on their subscribers, video views, and counts. The values are so massive, which is why log transformation is applied.

Simulated Data

- An artificial dataset has been generated involving 1000 observations and x1, x2, and x3 variables, respectively.



Both datasets are normalised and centralised by using the R built-in *scale* function to ease the calculation.

MCAR, MAR (by using the deterministic model), and MNAR (by using the probabilistic model) missing values have been created in both of the datasets.

5. Simulation Results

- The algorithm MICE and MissForest are implemented to impute the six datasets containing the missing values.
- R packages MICE and MissForest have been used to implement the MICE and MissForest algorithms. In the case of MICE, the predictive mean matching method (pmm) is selected as the datasets are continuous.
- To compare the algorithms, Bias, RMSE, and Coverage are measured in the imputed datasets regarding their true values.

MICE	Bias	RMSE	Coverage
Real Data (MCAR)	-0.0101218	0.6899317	0.493
Simulated Data (MCAR)	0.005214564	0.6610747	0.087
Real Data (MAR)	-0.006527207	0.08909458	0.07
Simulated Data (MAR)	-0.03593683	0.6677199	0.079
Real Data (MNAR)	0.1581902	0.8630873	0.065
Simulated Data (MNAR)	0.3595537	0.7902291	0.068

MissForest	Bias	RMSE	Coverage
Real Data (MCAR)	-0.001469021	0.4990769	0.5
Simulated Data (MCAR)	-0.007963796	0.4957609	0.074
Real Data (MAR)	-0.01671218	0.1816667	0.068
Simulated Data (MAR)	0.08492853	0.5516818	0.074
Real Data (MNAR)	0.1380952	0.7169724	0.089
Simulated Data (MNAR)	0.4072629	0.8177913	0.078

6. Interpretation

- In terms of low Bias and high coverage, MICE outperforms MissForest.
- In terms of low RMSE, MissForest outperforms MICE except for MAR missing data type.
- In the case of MAR missing data type, MICE completely dominates MissForest.
- According to the other studies, MAR is the most suitable selection based on multiple imputations because MAR complies with all the goals of the proper simulation study [8].

7. Conclusion

- MissForest might not comply with the goal of statistical inference using a proper simulation study; however, it is still an asset in the data imputation department.
- According to researchers MissForest works more precisely in the **Complete case (CC)** [8].
- To conclude, one can say that unlike MI FCS variants MissForest failed to comply with the goal of statistical inference using a proper simulation study.

References

[1] Heejin Jin, Surin Jung, and Sungho Won. "MissForest with feature selection using binary particle swarm optimization improves the imputation accuracy of continuous data". In: *Genes Genomics* 44.6 (2022), pp. 651–658.

[2] Małgorzata Misztal et al. "Comparison of Selected Multiple Imputation Methods for Continuous Variables–Preliminary Simulation Study Results". In: *Acta Universitatis Lodziensis. Folia Oeconomica* 6.339 (2018), pp. 73–98.

[3] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.

[4] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[5] Daniel J Stekhoven and Peter Bühlmann. "MissForest-non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1 (2012), pp. 112–118.

[6] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[7] Shengkai Zhang et al. "Imputation of gps coordinate time series using missforest". In: *Remote Sensing* 13.12 (2021), p. 2312.

[8] Xiaoping Zhu et al. "Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study". In: *Open Journal of Statistics* 4.11 (2014), p. 933.

Universität Trier

Tahomina Akter Ety, Tipu Sultan and Lara Krell

s4taetyy@uni-trier.de, s4tisult@uni-trier.de, s4lakrel@uni-trier.de