TREE-BASED MULTIPLE IMPUTATION METHODS

Michael Dellermann, Anatol Sluchych, and Jonah Wermter

Freie Universität Berlin

1. Motivation

- Parametric MICE methods: conditional models to be specified for *all* variables with missing data
- Still may fail to capture interactive and nonlinear relations among variables as well as non-standard distributions
- Tree-based methods *automatically* capture interactions, nonlinear relations, and complex distributions with no parametric assumptions or data transformations needed (Burgette & Reiter 2010)
- Implementation in R: *mice* and *miceRanger* packages

2. Tree-based methods

Classification and regression trees (CART):

- seek to approximate conditional distribution of univariate outcome from multiple predictors
- segment predictor space into non-overlapping regions with relatively homogeneous outcomes
- segments found by recursive binary splits of predictors
- prediction for observations that fall into the same region is mean (or mode) of response values for training observations in region
- may be very non-robust and have lower predictive accuracy

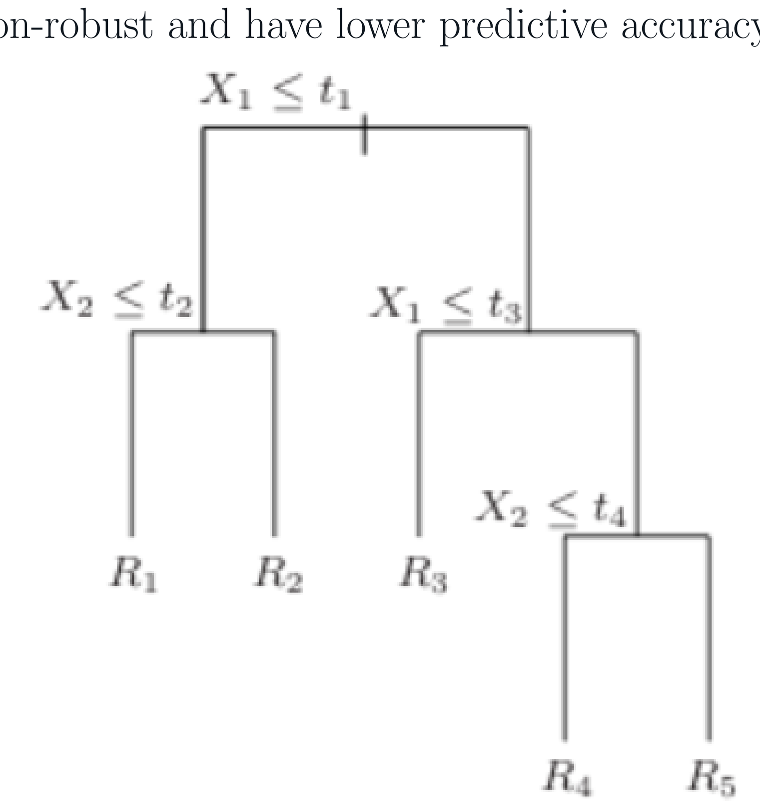

Fig. 1: Example of tree structure. Source: Hastie et al. (2009)

Random forest:

- *ensemble* method that addresses non-robustness and low predictive accuracy
- average predictions from B non-pruned trees constructed using B bootstrapped training sets
- *decorrelates* trees by performing each split on *randomly* chosen subset of predictors

3. Imputation algorithm

4-steps algorithm:

1. Initial values for the missing values filled in as follows:
   (a) Define a matrix Z equal to Y_c
   (b) Impute missing values in Y_i, where $i = 1, \ldots p_1$, using tree-based method on Z and append the completed version of Y_i to Z prior to incrementing i
2. Replace the originally missing values of Y_i, where $i = 1, \ldots p_1$, with tree-based methods on Y_{-i}
3. Repeat l times step 2
4. Repeat steps 1–3 m times and obtain m imputed sets.

4. Comparison mice/miceRanger packages

- both implement van Buuren's multivariate imputation by chained equations
- *mice* supports variety of imputation methods, *miceRanger* only random forest
- *mice* uses common R packages *rpart* and *randomForest* to implement tree based imputation methods (van Buuren 2023)
- *miceRanger* uses the *ranger* package instead, which claims to be faster and more efficient with medium and large data sets (Wilson 2022)
  - ⇒ core functions written in C++ (faster than R, compiled vs. interpreted code) (Wright & Ziegler 2017)
  - ⇒ lacks pooling function

5. Empirical simulation study

Empirical data set:

- RAND's Health Insurance Experiment: $n = 20185$, $k = 46$

Missing data mechanisms:

- p=25% and 50%
- MAR with $\rho = 0, \tau = 0$: $P(mdvis\_miss \mid xage < 25) = p$, $P(mdvis\_miss \mid mhi > 74) = p$
- MCAR: $P(income\_miss) = p$, $P(educdec\_miss) = p$

Monte Carlo simulation: R = 1000, M = 5, n = 2000, niter = 10, nrtree = 10

6. Results

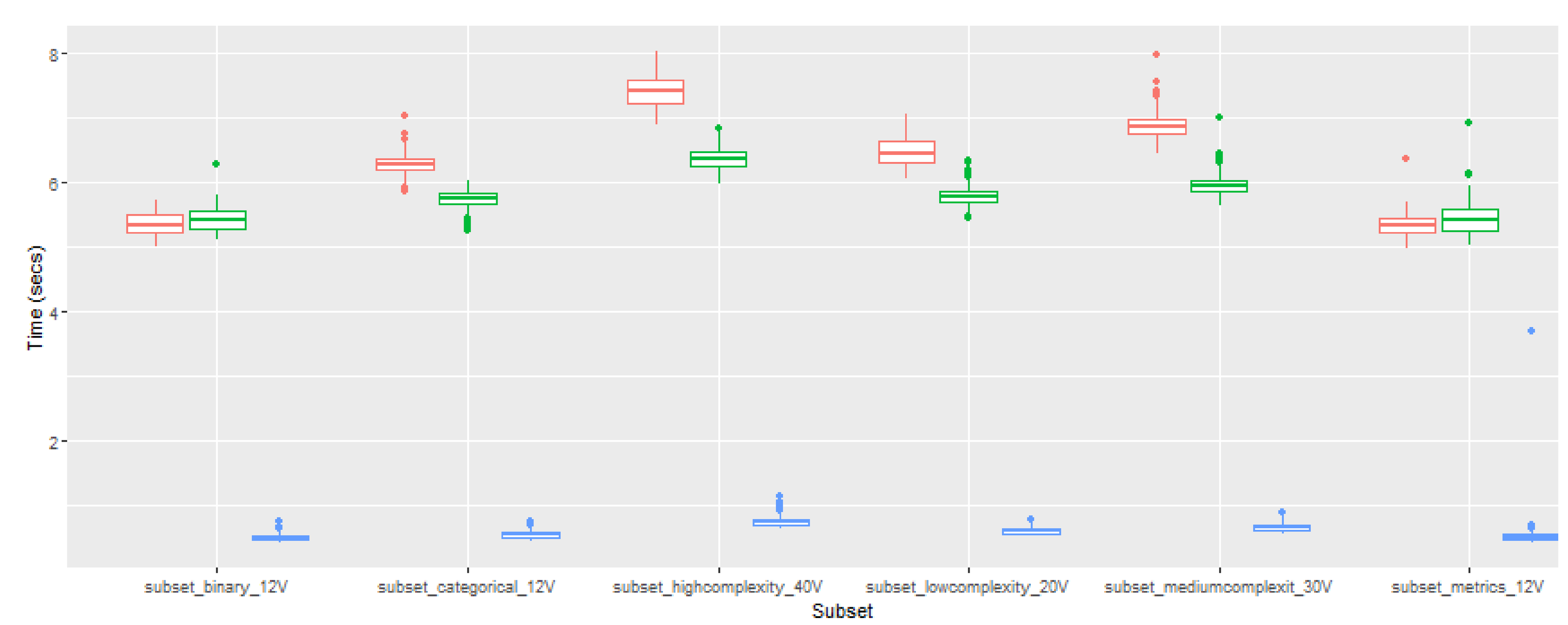

Fig. 2: Imputation Time per Subset per Method

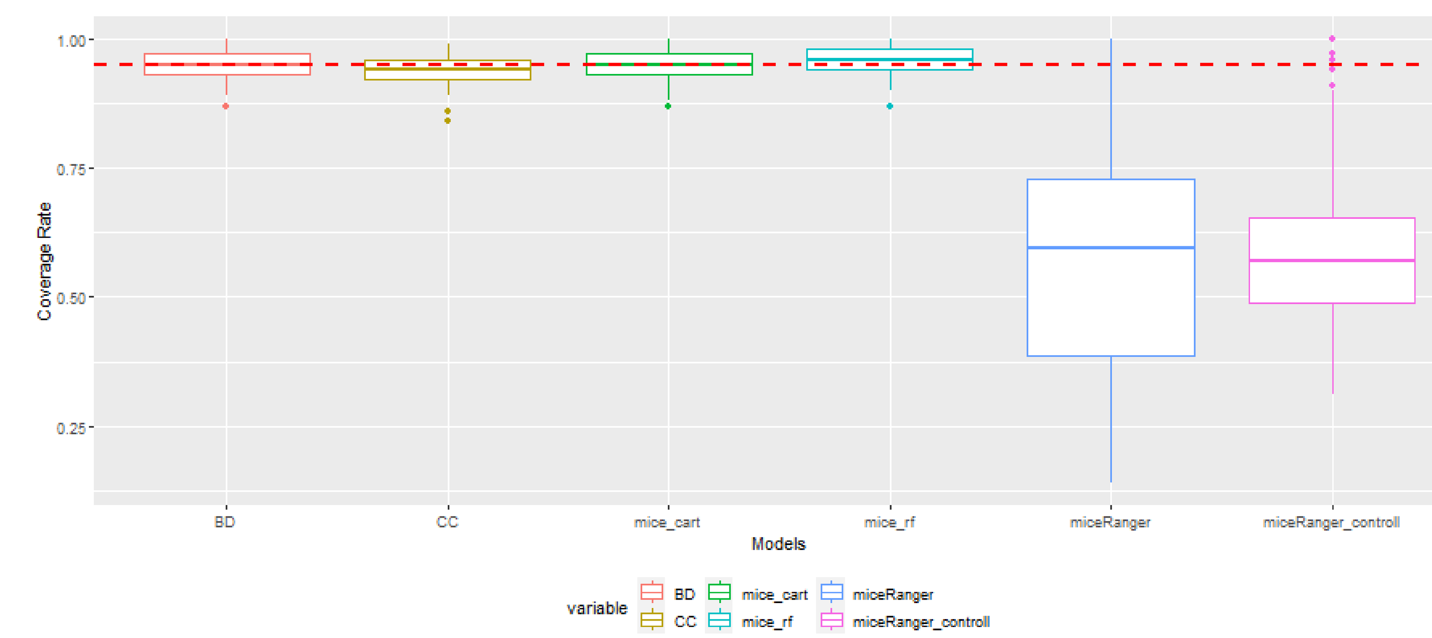

Fig. 3: Boxplot of Coverage Rates by Model

7. Conclusion

- *miceRanger* outperforms other random forest imputation methods, working on average approximately ...% faster per simulation cycle
- With changing the variability of data types, *miceRanger* works on average ...% faster per simulation cycle.
- With changing size of data sets, *miceRanger* works on average ...% fast per simulation cycle

References

Burgette, Lane F, and Jerome P Reiter. "Multiple imputation for missing data via sequential regression trees". *American journal of epidemiology* 172, no. 9 (2010): 1070–1076.

Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* Vol. 2. Springer, 2009.

Marvin, N. Wright, and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". *Journal of Statistical Software* 77, no. 1 (2017). `https://doi.org/10.18637/jss.v077.i01`.

van Buuren, Stef. *Package 'mice'.* Last updated on 05.06.2023, last checked on 28.12.2023. `https://cran.r-project.org/web/packages/mice/mice.pdf`.

Wilson, Sam. *Package 'miceRanger'.* Last updated on 13.10.2022, last checked on 28.12.2023. `https://cran.r-project.org/web/packages/miceRanger/miceRanger.pdf`.

Wright, Marvin N. *Package 'ranger'.* Last updated on 12.11.2023, last checked on 23.12.2023. `https://cran.r-project.org/web/packages/ranger/ranger.pdf`.