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 (van Buuren & Groothuis-Oudshoorn, 2011)
- 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 relatively low predictive accuracy

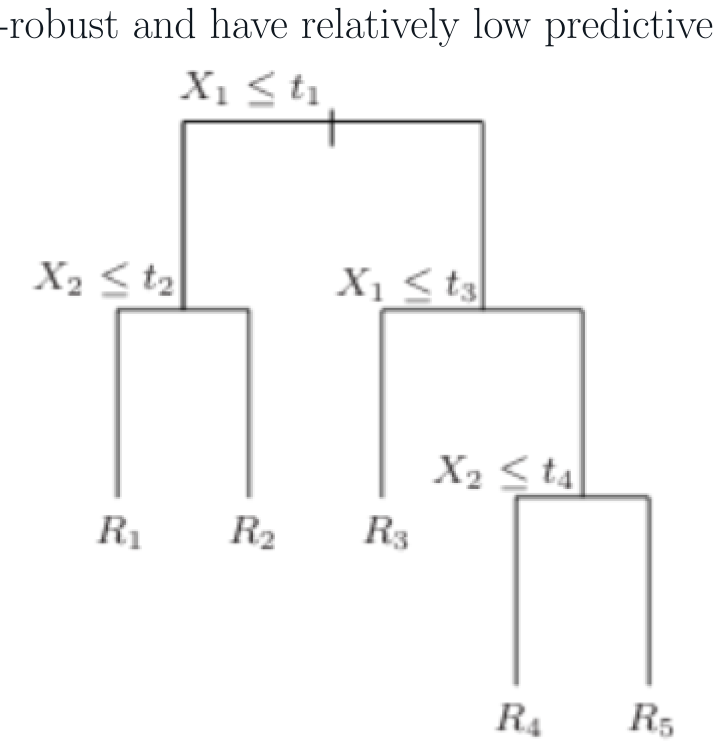

Fig. 1: Example of tree structure. Source: Hastie, Tibshirani, & Friedman (2009)

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
- accurate model to impute missing values (Stekhoven & Bühlmann, 2011)

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 $i = 1, ...p_1$, 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 $i = 1, ...p_1$, 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

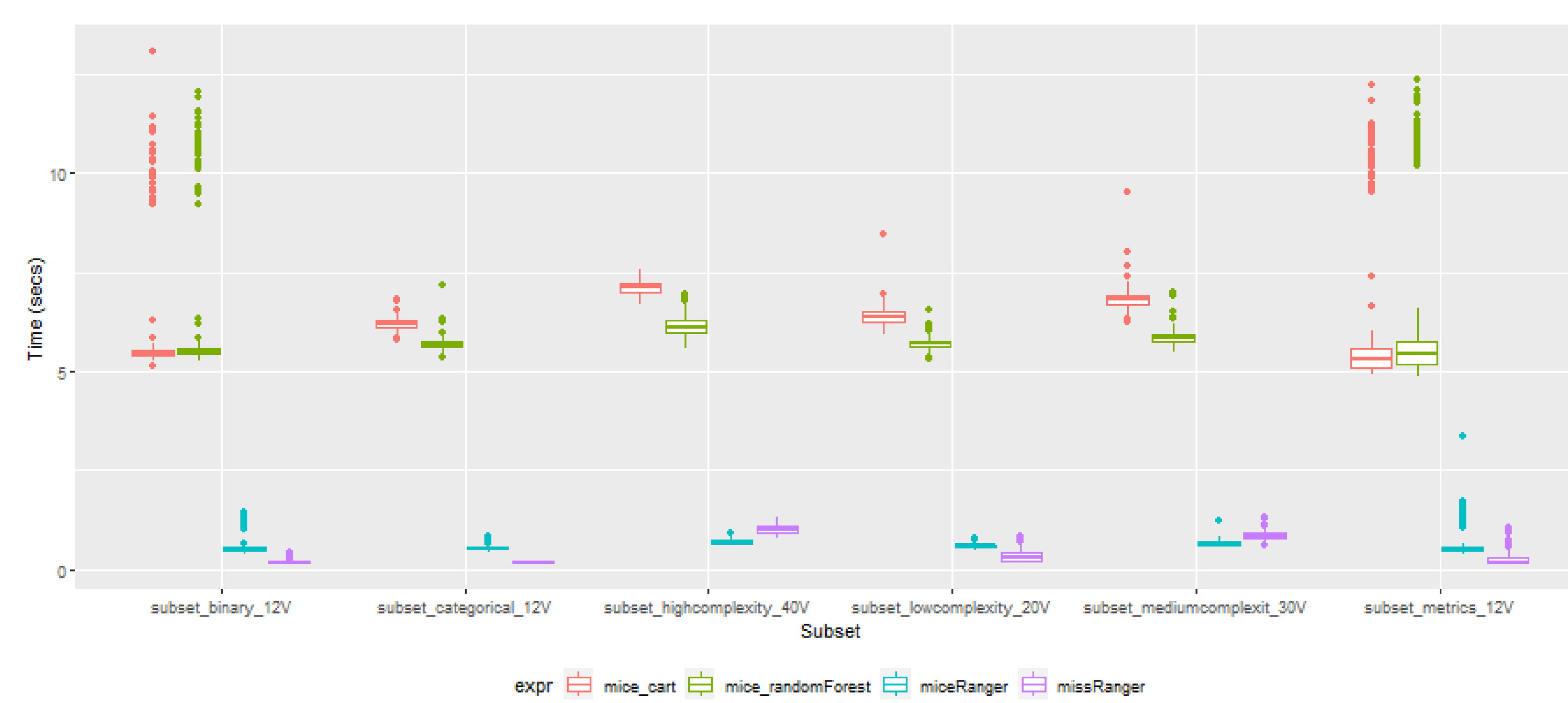

Fig. 2: Imputation time per subset per method

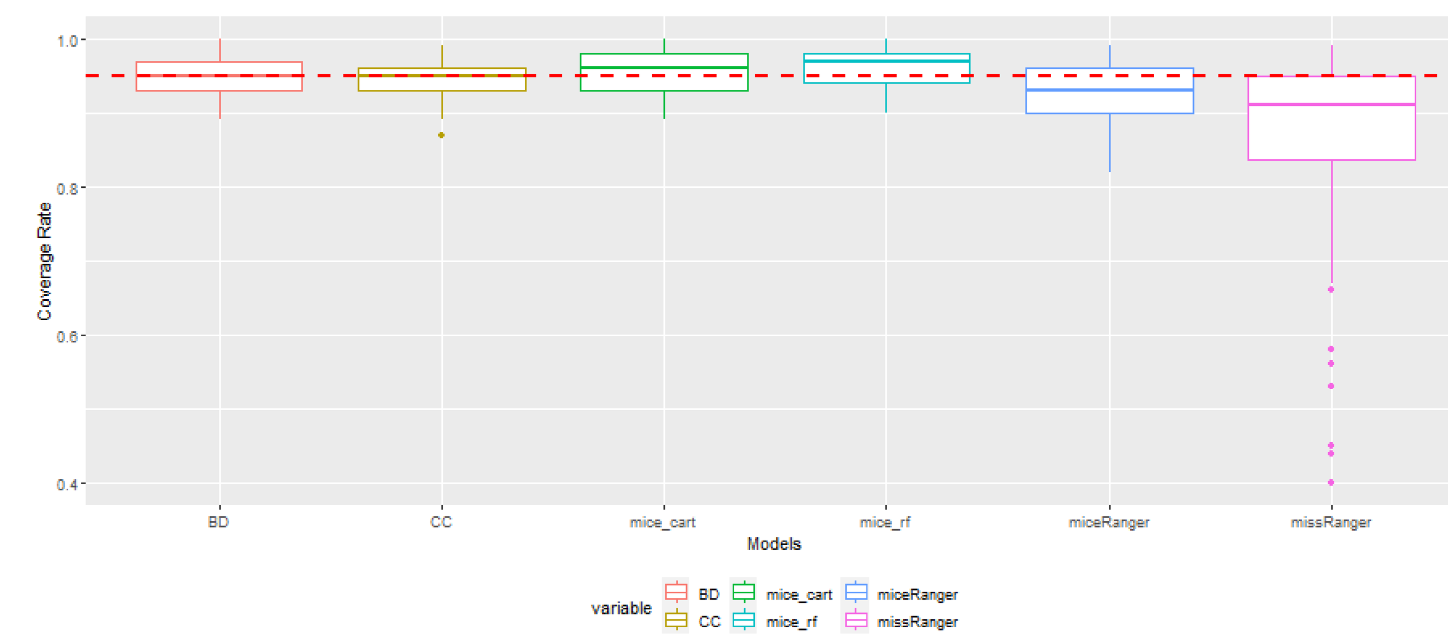

Fig. 3: Coverage rate by model

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, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology, 172*(9), 1070–1076. doi:10.1093/aje/kwq260

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Mayer, M. (2023). *Package 'missranger'*. Retrieved from https://cran.r-project.org/web/packages/missRanger/missRanger.pdf

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. Cambridge: MIT Press.

Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. doi:10.1093/bioinformatics/btr597

van Buuren, S. (2023). Package 'mice'. Retrieved from https://cran.r-project.org/web/packages/mice/mice.pdf

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. doi:10.18637/jss.v045.i03

Wilson, S. (2022). *Package 'miceranger'*. Retrieved from https://cran.r-project.org/web/packages/miceRanger/miceRanger.pdf

Wright, M. N. (2023). *Package 'ranger'*. Retrieved from https://cran.r-project.org/web/packages/ranger/ranger.pdf

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*(1). doi:10.18637/jss.v077.i01