Msc Thesis

Multilevel Multivariate Imputation by Chained through Bayesian Additive Regression Trees

Author
Heleen Brüggen

Affiliations
Methodology and Statistics for the Behavioural, Biomedical and Social Sciences, Utrecht University

Contact
h.bruggen@uu.nl

Introduction

Currently, there are two intricacies relating to multilevel multiple imputation:

- Firstly, because of the complicated hierarchical structure of the multilevel data, defining the imputation models also becomes complex and complex models might not converge as well.
- Secondly, the imputation models must be congenial - they must be at least as general as the analysis model. In multilevel analysis, there are many non-linear relationships included in the model and, on top of that, the analysis model is often not determined beforehand.

In this thesis, I wanted to solve these intricacies by implementing a non-parametric model - Bayesian Additive Regression Trees - as an imputation method within the MICE framework.

Theoretical background

Multiple imputation

1. Each missing value is imputed **m** times with values from their posterior predictive distribution conditional on the observed data and parameters from the imputation model.
2. Each of the imputed datasets is analyzed.
3. Their corresponding model parameters are pooled together.

Bayesian Additive Regression Trees (BART)

BART is a sum-of-trees model proposed by Chipman et al. (2010) with regression trees as its building blocks. Multiple trees are grown successively, However, each tree is constrained to explain only a part of the outcome variable through a regularization prior and they are perturbed at each iteration to avoid local minima.

Research question

How can multivariate imputation by chained equations through a multilevel bayesian additive regression trees model improve the bias, variance, and coverage of the multilevel model parameter estimates compared to current practices?

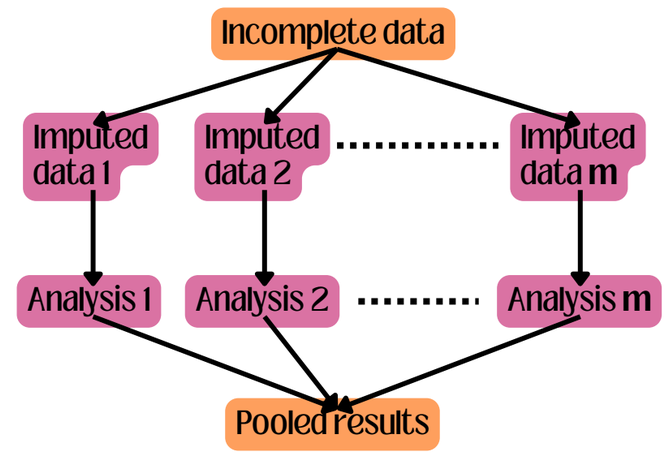

Figure 1.Schematic of main steps in multiple imputation

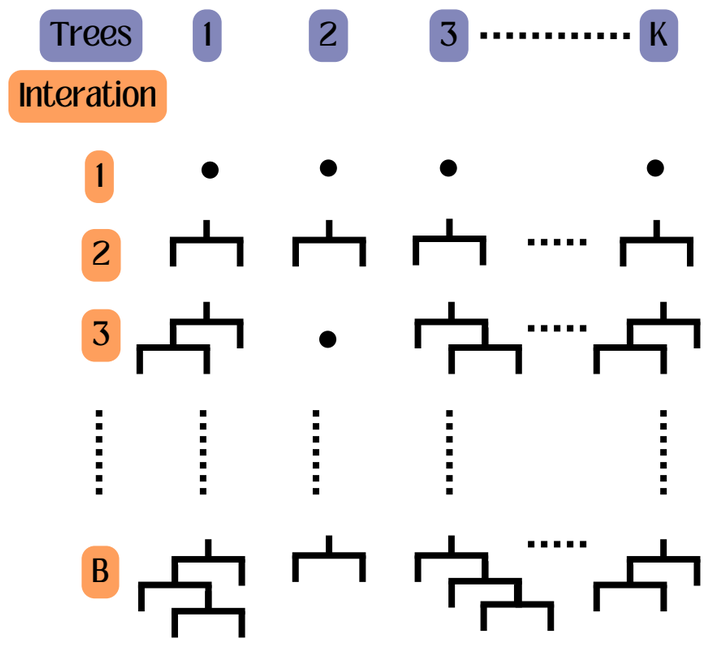

Figure 2. Schematic of the BART algorithm

Methodology

Data generating mechanism

7 level-1 variables | 2 level-2 variables | 3 cross-level interactions | 3 random slopes | Random intercept | Residual variance

Simulation design

Design factors	Values
Number of groups	30, 50
Group sizes	15, 50
Missingness mechanism	MCAR, MAR
Amount of missingness	0%, 50%

Missing data generation

Either no missing values are introduced (0%), or up to 5 missing values are introduced in 50% of the cases. When data is MAR, the probability of a value being missing depends on the observed values of all other variables, with variables x4 and z1 having a greater influence on this probability.

Missing data generation

1. Absolute bias
2. Coverage of 95% confidence intervals
3. Confidence interval width

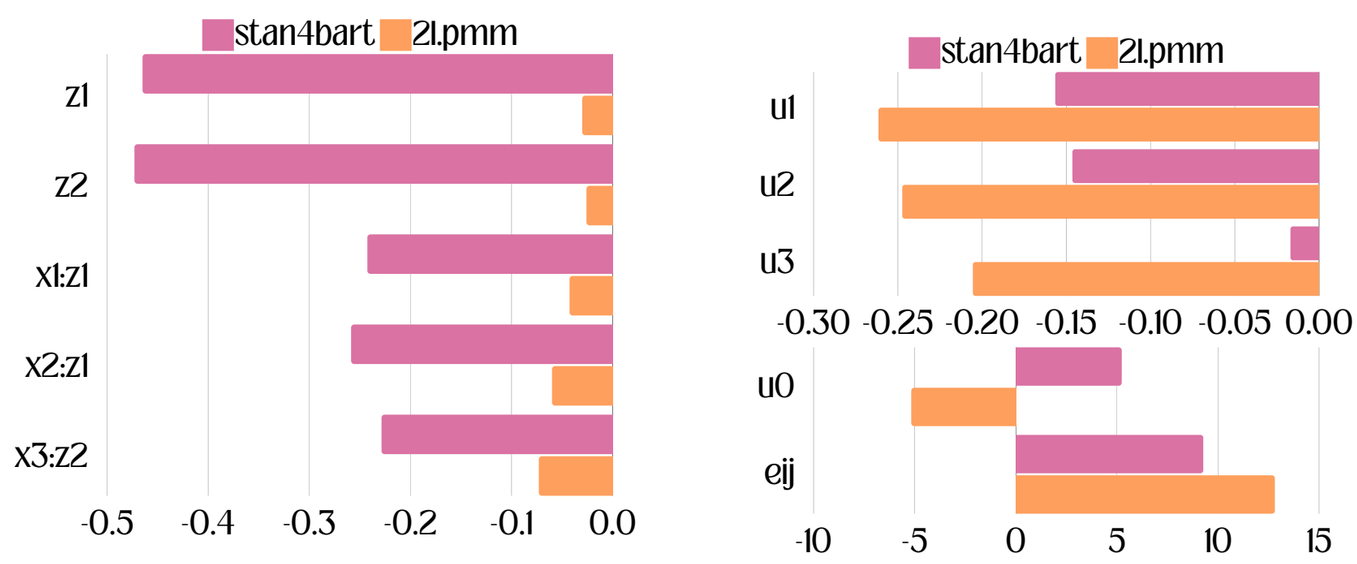

Figure 3.The absolute bias for the level-2 effects (z1 and z2), the cross-level interactions (x1:z1, x2:z2, and x3:z2), the random slopes (u1, u2, and u3) and the random intercept (u0) and residual variance (eij). stan4bart is the multilevel BART imputation method and 2l.pmm is the conventional 2-level predictive mean matching method.

Results/Findings

From figure 3, we can see from the absolute bias for the level-2 effects, that the conventional imputation method (2l.pmm) outperforms the multilevel-BART (stan4bart). On the other hand, the multilevel-BART model performs better in terms of the random structure of the model - the random slopes, intercept and residual variance).

Conclusion

The current implementation of a multilevel-BART imputation method did not improve on the current implementation of multilevel predictive mean matching.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). Mice : Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3).
Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1).
Dorie, V., Perrett, G., Hill, J. L., and Goodrich, B. (2022). Stan and BART for Causal Inference: Estimating Heterogeneous Treatment E↵ects Using the Power of Stan and the Flexibility of Machine Learning. Entropy, 24(12):1782.
Grund, S., Lüdtke, O., and Robitzsch, A. (2018b). Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. Organizational Research Methods, 21(1):111–149.
Hastie, T. J., editor (2017). Statistical Models in S. Routledge, 1st edition.
van Buuren, S. (2018). Flexible Imputation of Missing Data. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor & Francis Group, Boca Raton London New York, second edition edition.