## Mo levels mo problems: Reducing the complexity of multiple imputation in a multilevel context

Supervisors: Thom Volker, Hanne Oberman, Gerko Vink

Project number: UU-31

Short description:

Multiple imputation is the de facto standard in solving missing data problems. Although multiple imputation is generally a flexible approach that can deal with a wide range of missing data problems, it can still be a daunting task to specify conditional regression models for potentially many variables in the data. This is especially true if there are non-linear relations or interactions in the data, and the problem further increases with hierarchical data structures. The complexity of clustered data can be captured with multilevel modeling techniques. Multilevel models accommodate heterogeneous relationships between distinct groups, but can be difficult to specify correctly.

In a single-level context, the task of the modeler can be simplified considerably through the use of non-parametric imputation approaches as classification and regression trees (CART; Burgette & Reiter, 2010), Bayesian additive regression trees (BART; Xu et al., 2016) or random forests (Doove et al., 2014). These approaches have the attractive feature that they almost automatically capture complex relationships between variables, as non-linearity or interactions. Recently, the methodology of CART (Hajjem et al., 2011; Lin & Luo, 2019), BART (Chen, 2020) and random forests (Pellagatti et al., 2021) has been extended to allow for prediction in a multilevel context (see also Salditt et al., 2022). However, despite the promising results of single-level tree-based approaches in imputation, the multilevel extensions have neither been implemented nor evaluated as tools in an imputation technique.

In this project, you will implement one or several of the multilevel tree-based approaches in the R-package mice, and evaluate the implemented method(s) by means of simulation. The standard multilevel imputation techniques (see Enders et al., 2016; Audigier et al., 2018; and Lüdtke et al., 2017 for overviews) will serve as benchmark methods. Depending on your skills and interest different angles to the research problem can be taken, including but not limited to a Bayesian modeling approach, health center data applications and extensions to different machine learning methods. Upon successful completion, you will advance the state-of-the-art in multilevel imputation and ease the lives of many applied researchers working with multilevel data.

The thesis project is offered by the [Missing Data research group](https://www.uu.nl/en/organisation/methodology-and-statistics/missing-data). You will collaborate closely with the core development team of mice, see also [amices.org](https://amices.org/) and [Github](https://github.com/amices).

Required skills: programming in R, affinity with multilevel modeling, interest in machine learning

Literature:

Audigier, Vincent, Ian R. White, Shahab Jolani, Thomas PA Debray, Matteo Quartagno, James Carpenter, Stef Van Buuren, and Matthieu Resche-Rigon. "Multiple imputation for multilevel data with continuous and binary variables." *Statistical Science* 33, no. 2 (2018): 160-183.

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, *172*(9), 1070-1076.

Chen, S. (2020). *A New Multilevel Bayesian Nonparametric Algorithm and Its Application in Causal Inference* (Doctoral dissertation).

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational statistics & data analysis*, *72*, 92-104.

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological methods*, *21*(2), 222.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters*, *81*(4), 451-459.

Lin, S., & Luo, W. (2019). A new multilevel CART algorithm for multilevel data with binary outcomes. *Multivariate Behavioral Research*, *54*(4), 578-592.

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological methods*, *22*(1), 141.

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A. M. (2021). Generalized mixed‐effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *14*(3), 241-257.

Salditt, M., Humberg, S., & Nestler, S. (2022). Gradient Tree Boosting for Hierarchical Data. *Multivariate Behavioral Research*, 1-27.

Xu, D., Daniels, M. J., & Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, *17*(3), 589-602.