Master Research Report: Multilevel Multivariate Imputation by Chained Equations through Bayesian Additive Regression Trees

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Heleen Brüggen

Word count: 1384
Candidate Journal: Computational Statistics & Data Analysis
FETC Case Number: 23-1778

Supervisors:

MSc. T. Volker — Utrecht University
Dr. G. Vink — Utrecht University
MSc. H. Oberman — Utrecht University

1 Introduction

Incomplete data is a common challenge in many fields of research. A frequent approach for dealing with incomplete data is listwise deletion, also known as complete-case analysis, which is to remove all incomplete cases from the data. However, this could possibly lead to biased results if the data is not Missing Completely At Random (MCAR), meaning the cause of the missing data is unrelated to the data [Austin et al., 2021, Enders, 2017, Kang, 2013, Rubin, 1976, van Buuren, 2018]. Furthermore, other approaches to dealing with incomplete data include: pairwise deletion, mean imputation and regression imputation, which also yield biased results [van Buuren, 2018]. Pairwise deletion, also known as available-case analysis, is to remove all incomplete cases from the analysis when considering a specific pair of variables. Pairwise deletion leads to unbiased results when correlation between variables are low and the data is MCAR. Mean imputation is to replace missing values with the mean of the observed values. Mean imputation will bias almost all estimates except the mean when the data is not MCAR. Regression imputation is to replace missing values with the predicted values from a regression model and is unbiased when the data is Missing At Random (MAR), meaning that the missing data is related to the observed data [Rubin, 1976], and the factor influencing the missingness is present in the data [van Buuren, 2018]. So, in order to use these ad hoc strategies to deal with incomplete data correctly, the missing data mechanism should be carefully determined. However, determining the missing data mechanism is often difficult and MCAR is in practise an unrealistic assumption [van Buuren, 2018]. We can imagine that when a system becomes more complex, thus increases the amount of parameters, testing these assumptions also increases in complexity.

One of these complex systems are multilevel data structures. Multilevel data is hierarchically structured, where, for example, students are nested within schools, or patients are nested within hospitals [Hox and Roberts, 2011, Hox et al., 2017]. Thus, in these types of data sets there are level-1 and level-2 variables. Level-1 variables relate to the individual within a class and level-2 variables relate to the class as a whole. The recommended statistical technique to analyzing these models are multilevel models as it accounts for the specific dependencies in the multilevel data sets [Hox and Roberts, 2011, Hox et al., 2017, Lüdtke et al., 2017]. It can contain both level-1 and level-2 variables, random intercepts, random slopes, and cross-level interactions [Hox and Roberts, 2011, Hox et al., 2017].

Multiple imputation (MI) is considered a valid method for dealing with incomplete data and allows us the seperate the missing data problem from the analysis problem [Audigier et al., 2018, Austin et al., 2021, Burgette and Reiter, 2010, Enders, 2017, Grund et al., 2021, Hughes et al., 2014, Mistler and Enders, 2017, Van Buuren, 2007, van Buuren, 2018]. MI imputes each missing value in the data set more than once given the observed data, thereby considering necessary variation associated with the missingness problem. The multiply imputed data sets are analyzed, and the corresponding inferences are pooled according to Rubin's rules [Austin et al., 2021, Carpenter and Kenward, 2013, Rubin, 1987, van Buuren, 2018]. Generally, multiple imputation operates under two frameworks: joint modeling (JM) and fully conditional specification (FCS) [Enders et al., 2018a,b, Hughes et al., 2014, Mistler and Enders, 2017, van Buuren, 2018]. JM employs a multivariate data distribution and regresses incomplete variables on complete variables to impute missing values. FCS, or chained equations, iteratively imputes one variable with missing values at a time through conditional univariate distributions regressing an incomplete variable complete and previously imputed variables [Enders et al., 2018a,b, 2016, Grund et al., 2018a, Hughes et al., 2014, Mistler and Enders, 2017, van Buuren, 2018]. For multilevel data JM is extend by defining a multivariate linear mixed model. FCS is extended by defining a series of univariate linear mixed models [Enders et al., 2018a, Mistler and Enders, 2017]. The implementation of JM and FCS in a multilevel context are equivalent in a two-level random intercept context with normally distributed data. However, they differ beyond that: JM is more capable of handling within- and between- cluster relationships, random intercepts and incomplete categorical variables, while FCS is better suited for random slopes and restricted to normally distributed variables [Enders et al., 2018b, 2016]. Also, they differ in their handling of missing level-2 data. Overall, FCS is believed to be more flexible than JM [Audigier et al., 2018, Burgette and Reiter, 2010, Grund et al., 2018a, Van Buuren, 2007] and, thus, may be better suited for multilevel data. In FCS, one needs to define conditional models for all variables with missing values [Enders et al., 2018a,b, 2016, Grund et al., 2018a, Hughes et al., 2014, Mistler and Enders, 2017, van Buuren, 2018] and the imputation models should at least be as general as the analysis model and preferably all-encompassing [Grund et al., 2018b]. However, the complexity of the multilevel analysis model is built step-wise with non-linearities [Hox and Roberts, 2011, Hox et al., 2017]. Thus, defining imputation models for a multilevel data set is quite challenging [Burgette and Reiter, 2010, Hox and Roberts, 2011, van Buuren, 2018].

Using non-parametric tree-based models as imputation models might solve this problem. Tree-based models use recursive partitioning to split the data into smaller subgroups based on the predictor variables maximizing the homogeneity of the subgroups. Tree-based models are non-parametric, which means that they do not assume a specific distribution of the data. Thus, they can handle non-linear relationships and interactions between the predictor variables well. Furthermore they handle continuous and catergorical variables simultaneously [Breiman et al., 1984, Burgette and Reiter, 2010, Chipman et al., 2010, Hill et al., 2020, James et al., 2021, Lin and Luo, 2019, Salditt et al., 2023].

In a single-level imputation context, the use of tree-based, non-parametric models like regression trees, random forests or Bayesian Additive Regression Trees (BART) simplified imputation models and performed better than parametric methods: the imputations showed better confidence interval coverage of the population parameters, lower variance and lower bias, especially in non-linear and interactive contexts [Burgette and Reiter, 2010, Silva and Gutman, 2022, Xu et al., 2016]. Others have also found lower normalized root mean squared error (NRMSE), which in essence encapsulates the bias of the imputations, when imputing with a random forest algorithm compared to MICE and KNN imputation [Stekhoven and Bühlmann, 2012, Waljee et al., 2013]. Furthermore, they also found that the algorithm reduced computational time and could handle multivariate data consisting of both continuous and categorical data simultaneously.

BART models have also been implemented in a multilevel prediction context. However, multilevel-BART models (M-BART) have predominantly been implemented with random intercepts and no random slopes and cross-level interactions [Chen, 2020, Tan et al., 2016, Wagner et al., 2020, Wundervald et al., 2022]. Wagner et al. [2020] have found that this random intercept M-BART model provided better predictions with a lower Mean Squared Error (MSE) compared to a parametric multilevel model, Tan et al. [2016] found higher Area Under the Curve values, and Chen [2020] found better predictions and better coverage compared to parametric models and a single-level BART model. Other researchers modeled the random intercept as an extra split on each terminal node within the BART algorithm and found a lower MSE compared to a standard BART model and parametric multilevel models [Wundervald et al., 2022]. Dorie et al. [2022] developed a multilevel BART model that included random intercepts and random slopes by combining BART with the Stan algorithm. However, the random intercept and slope are modeled by Stan, which is a parametric method. Their results showed that their algorithm 'stan4bart' showed better coverage of the population value and lower Root Mean Squared Error (RMSE) compared to BART models with varying intercept, BART models ignoring the multilevel structure, Bayesian Causal Forests (BCF), and parametric multilevel models.

In spite of these promising findings: tree-based model performing well in single-level imputation context [Burgette and Reiter, 2010, Silva and Gutman, 2022, Stekhoven and Bühlmann, 2012, Waljee et al., 2013, Xu et al., 2016] and M-BART models performing well in a multilevel prediction context [Chen, 2020, Dorie et al., 2022, Tan et al., 2016, Wagner et al., 2020, Wundervald et al., 2022], M-BART models have yet to be implemented in a multilevel multiple imputation context. Thus, my research question will be: *Can multivariate imputation by chained equations through a multilevel bayesian additive regression trees model improve the bias, variance and coverage of the estimates in a multilevel context compared to current practices?* Given the success of non-parametric models in single-level multiple imputation, I anticipate that employing multilevel BART models in a multilevel missing data context will reduce bias, accurately model variance, and improve estimate coverage compared to classical multilevel imputation through *2l.pmm, 2l.lmer, 2l.pan, 2l.jomo, rf* and *pmm* in MICE [Buuren and Groothuis-Oudshoorn, 2011].

This research report is organised as follows: in section 2, will contain some theoretical background and describe the methods in which I will implement the M-BART model in a multilevel imputation context and Section 3 will provide some preliminary results.

2 Method

2.1 Theoretical background

Bayesian Additive Regression Trees (BART) is a sum-of-trees model proposed by Chipman et al. [Chipman et al., 2010]. Regression trees are its building blocks [Chipman et al., 2010, Hill et al., 2020, James et al., 2021]. Regression trees model non-linearities well and automatically through recursive binary partitioning of the predictor space [Burgette and Reiter, 2010, Hill et al., 2020]. Recursive binary partitioning doesn't assume a specific data form; it divides the predictor space to maximize variance explanation by automatically identifying best fitting splits [Hastie, 2017, James et al., 2021, Salditt et al., 2023]. BART

models can be described as:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{1.1}$$
$$y_i = g(\mathbf{x}_i, T_1, M_1) + g(\mathbf{x}_i, T_2, M_2) + \cdots + g(\mathbf{x}_i, T_k, M_k) + \epsilon_i, \tag{1.2}$$

where y_i is the outcome variable for person i, $f(\mathbf{x}_i)$ is the sum-of-trees many regression trees, and ϵ_i is the error term; $\epsilon \sim \mathcal{N}(0, \sigma^2)$. $\mathbf{x}$ are the predictors included in the model, T_k is the k^{th} tree and M_k is the collection of leaf parameters within the k^{th} tree [Chipman et al., 2010, Hill et al., 2020, James et al., 2021]. Next to the sum-of-trees model, BART also includes a regularization prior that constrains the size and fit of each tree so that each contributes only a small part of the overall fit to prevent overfitting [Chipman et al., 2010, Hill et al., 2020, James et al., 2021]. The Bayesian back-fitting Markov Chain Monte Carlo (MCMC) algorithm is used to obtain estimates from BART. It updates each tree, conditional on the remaining trees, their associated parameters and σ, by fitting a new tree to the partial residuals, r_i, perturbing the tree from the previous iteration. The partial residuals, r_i, are defined as:

$$r_i = y_i - \sum_{k'<k} \hat{f}^b_{k'}(x_i) - \sum_{k'>k} \hat{f}^{b-1}_{k'}(x_i), \tag{2}$$

where $\hat{f}^b_{k'}(x_i)$ is the prediction of the k'^{th} tree in the b^{th} iteration for person i.
The M-BART model including a random intercept can be identified as:

$$y_{ij} = \sum_{k=1}^{m} f(\mathbf{X}_{ij}; T_k, M_k) + \alpha_j + \epsilon_{ij}, \tag{3}$$

where, now, y_{ij} is the outcome variable for person i in cluster j and α_j is the random intercept for cluster j.

2.2 Simulation study

3 Results

References

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., Van Buuren, S., and Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2).

Austin, P. C., White, I. R., Lee, D. S., and Van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Routledge, 1 edition.

Burgette, L. F. and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9):1070–1076.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). **Mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3).

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. Wiley, 1 edition.

Chen, S. (2020). *A New Multilevel Bayesian Nonparametric Algorithm and Its Application in Causal Inference*. PhD thesis, Texas A&M University.

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 Effects Using the Power of Stan and the Flexibility of Machine Learning. *Entropy*, 24(12):1782.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98:4–18.

Enders, C. K., Hayes, T., and Du, H. (2018a). A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. *Multivariate Behavioral Research*, 53(5):695–713.

Enders, C. K., Keller, B. T., and Levy, R. (2018b). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2):298–317.

Enders, C. K., Mistler, S. A., and Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2):222–240.

Grund, S., Lüdtke, O., and Robitzsch, A. (2018a). Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs. *Journal of Educational and Behavioral Statistics*, 43(3):316–353.

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.

Grund, S., Lüdtke, O., and Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, 53(6):2631–2649.

Hastie, T. J., editor (2017). *Statistical Models in S*. Routledge, 1st edition.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1):251–278.

Hox, J. and Roberts, J. K., editors (2011). *Handbook of Advanced Multilevel Analysis*. Routledge, 0 edition.

Hox, J. J., Moerbeek, M., and Van De Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. Routledge, Third edition. — New York, NY : Routledge, 2017. —, 3 edition.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):28.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402.

Lin, S. and 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., and Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1):141–165.

Mistler, S. A. and Enders, C. K. (2017). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 42(4):432–466.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Salditt, M., Humberg, S., and Nestler, S. (2023). Gradient Tree Boosting for Hierarchical Data. *Multivariate Behavioral Research*, pages 1–27.

Silva, G. C. and Gutman, R. (2022). Multiple imputation procedures for estimating causal effects with multiple treatments with application to the comparison of healthcare providers. *Statistics in Medicine*, 41(1):208–226.

Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Tan, Y. V., Flannagan, C. A. C., and Elliott, M. R. (2016). Predicting human-driving behavior to help driverless vehicles drive: Random intercept Bayesian Additive Regression Trees.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

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.

Wagner, J., West, B. T., Elliott, M. R., and Coffey, S. (2020). Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context. *Journal of Official Statistics*, 36(4):907–931.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8):e002847.

Wundervald, B., Parnell, A., and Domijan, K. (2022). Hierarchical Embedded Bayesian Additive Regression Trees.

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