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

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Heleen Brüggen

Utrecht University

Word count: 2496
Candidate Journal: Computational Statistics & Data Analysis
FETC Case Number: 23-1778

Supervisors:

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

Contents

1 Introduction

Incomplete data is a common challenge in many fields of research. Frequently used ad hoc strategies to deal with missing data, such as complete case analysis or mean imputation, often lead to erroneous inferences in realistic situations. These strategies don't consider the multivariate nature of the data. Missingness can depend on observed data or even unobserved data, leading to biased estimates and inaccurate variance estimates when using one of these ad hoc strategies (Austin et al., 2021; Enders, 2017; Kang, 2013; Little and Rubin, 2002; van Buuren, 2018). Multiple imputation (MI; Rubin, 1987) is considered an effective method for dealing with incomplete data supported by a considerable amount of methodological research (Audigier et al., 2018; Austin et al., 2021; Burgette and Reiter, 2010; Enders, 2017; Grund et al., 2021; Hughes et al., 2014; Little and Rubin, 2002; Mistler and Enders, 2017; Van Buuren, 2007; van Buuren, 2018).

MI allows us to separate 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; Little and Rubin, 2002; Mistler and Enders, 2017; Van Buuren, 2007; van Buuren, 2018). Each missing value in the dataset is imputed multiple times by drawing values from their posterior predictive distribution conditional on the observed data and parameters from the imputation model. The imputation model is statistical model specifying the variables used for imputation. By imputing the data multiple times, the necessary variation associated with the missingness problem is considered. After imputation, each of the imputed datasets are analyzed according to the model of interest, i.e. the substantive analysis model of interest. Then, their corresponding inferences are pooled together according to Rubin's rules (Austin et al., 2021; Carpenter and Kenward, 2013; Rubin, 1987; van Buuren, 2018). One central feature of MI is the concept of congeniality; the imputation model should should be at least as general as the analysis model and preferably all-encompassing (Bartlett et al., 2015; Enders et al., 2018a; Grund et al., 2016, 2018b; Little and Rubin, 2002; Meng, 1994). If not, the imputation model will not capture every aspect of the data and the pooled analysis model estimates may be biased.

When MI is applied in a multilevel data context, concerns regarding the concept of geniality become more pronounced (Audigier et al., 2018; Dong and Mitani, 2023; Enders et al., 2020, 2018a,b, 2016; Grund et al., 2016, 2018a,b, 2021; Lüdtke et al., 2017; Mistler and Enders, 2017; Quartagno and Carpenter, 2022; Resche-Rigon and White, 2018; Taljaard et al., 2008; van Buuren, 2018). Multilevel data is hierarchically structured, where, for example, students are nested within classes within schools or patients within hospitals (Hox and Roberts, 2011; Hox et al., 2017). When analyzing multilevel data, this hierarchical structure should be taken into consideration. Ignoring it will underestimate the intra-class correlation (ICC) and standard errors, as conventional statistical analyses assume independence of observations (Hox and Roberts, 2011; Lüdtke et al., 2017; Taljaard et al., 2008; van Buuren, 2018). The ICC can be interpreted as the proportion of the total variance at level-2 (Gulliford et al., 2005; Hox and Roberts, 2011; Shieh, 2012). Accounting for this structure, can be done using multilevel models (MLMs; Hox and Roberts, 2011; Hox et al., 2017; Lüdtke et al., 2017). MLMs can contain both level-1, and level-2 variables, relating to the individual and class respectively, random intercepts, random slopes, and cross-level interactions (Hox and Roberts, 2011; Hox et al., 2017). For example, the length of a student is a level-1 variable, the teacher's experience is a level-2 variable, a random intercept would indicate that some classes have significantly shorter or taller students on average, a random slope would indicate that the relationship between the length of a students and the outcome variable differs between classes, and the effect of the length of the student can differ with the teacher's experience, i.e. a cross-level interaction (Hox and Roberts, 2011; Hox et al., 2017). Typically, the complexity of the multilevel analysis model is built step-wise with non-linearities, meaning the analysis model is not determined beforehand: predictores, random intercepts, random slopes, and cross-level interactions are added stepwise to the model (Hox and Roberts, 2011; Hox et al., 2017). Thus, including the hierarchical structure and final analysis model, along with the complicated non-linearities from cross-level interactions, in imputation models can be quite challenging (Burgette and Reiter, 2010; Hox and Roberts, 2011; van Buuren, 2018) and a very complex model might not converge (van Buuren, 2018).

Fully conditional specification (FCS), otherwise known as chained equations, is a popular and flexible implementation of MI (Audigier et al., 2018; Burgette and Reiter, 2010; Grund et al., 2018a; Van Buuren, 2007) and employs univariate linear mixed models to account for the hierarchical structure of multilevel models (Enders et al., 2018a; Mistler and Enders, 2017; Resche-Rigon and White, 2018). FCS iteratively imputes each incomplete variable conditional on observed 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). Furthermore, it can impute non-linearities, such as cross-level interactions, by using 'passive imputation'

or defining a separate imputation model for the non-linearities (Grund et al., 2018b; van Buuren, 2018). However, including these non-linearities in FCS is still very complicated (Grund et al., 2018b, 2021; van Buuren, 2018). FCS can also handle random intercepts and slopes, yet, once again, correctly specifying an imputation model accounting for these random effects can be challenging (Grund et al., 2018b, 2021; van Buuren, 2018).

Non-parametric, tree-based models might alleviate these complexities when defining imputation models. They do not assume a specific data distribution. So, they implicitly model non-linear relationships and can simultanously handle continuous and categorical variables (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 imputation of single-level data, tree-based, non-parametric models like regression trees, random forests, or Bayesian Additive Regression Trees (BART) simplified the imputation process. They showed better model parameter estimates than parametric methods. Specifically, the imputations showed better confidence interval coverage of the 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). Waljee et al. (2013) also found lower missclassification error rate for the predicted class as well as lower imputation error when imputing with a random forest algorithm compared to multivariate imputation by chained equations (`MICE`) using linear, logistic, and polytomous logistic regression imputation models, K-nearest neighbors (KNN) and mean imputation.

In prediction, multilevel-BART models (M-BART) have predominantly only been implemented with random intercepts (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 MLM, Tan et al. (2016) found higher area under the curve (AUC) values compared to a singel-level BART model and linear random intercept model, and Chen (2020) found better predictions and better coverage of the parameter estimates compared to parametric models and a single-level BART model. Other researchers modeled the random intercept as an extra split on each terminal node and found a lower MSE compared to a standard BART model and parametric MLMs (Wundervald et al., 2022). Dorie et al. (2022) developed a multilevel BART model that included random intercepts, random slopes and cross-level interactions by modeling these random parts with a Stan (Lee et al., 2017) model and the fixed parts with a BART model. Their results showed that their algorithm `stan4bart` showed better coverage of the population values and lower root mean squared error (RMSE) compared to BART models with varying intercept, BART models ignoring the multilevel structure, bayesian causal forests, and parametric MLMs.

Despite these promising findings, 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 mulitlevel model parameter estimates compared to current practices?* Given the success of non-parametric models in single-level MI, I anticipate that employing M-BART models in a multilevel missing data context will reduce bias, accurately model variance, and improve estimate coverage compared to conventional implementations of multilevel MI, single-level MI, and complete case analysis in the R-package `MICE` (Buuren and Groothuis-Oudshoorn, 2011).

2 Method

2.1 Theoretical background

2.1.1 Bayesian Additive Regression Trees (BART)

BART is a sum-of-trees model proposed by Chipman et al. (2010) that has regression trees as its building blocks (Chipman et al., 2010; Hill et al., 2020; James et al., 2021). Regression trees recursively split the data into binary subgroups based on the predictors included in the model. At each step down the tree, these splits are based on the predictor that minimizes the variability within the subgroups from all predictors. Observations are then assigned to a certain subgroup according to these splits. This is continued until a certain stopping criterion is reached; for example, we desire a minimal number of observations with in a subgroup (Breiman et al., 1984; Hastie, 2017; James et al., 2021; Salditt et al., 2023). Recursive binary partitioning of the predictor space doesn't assume a specific data form. This making regression trees, and as a consequence, BART, non-parametric models (Breiman et al., 1984; Hastie, 2017; James et al., 2021; Salditt et al., 2023) and allows regression trees to model non-linearities and other complicated relationships well and automatically (Burgette and Reiter, 2010; Hill et al., 2020).

Chipman et al. (2010) define the BART model as:

$$f(\mathbf{x}) = \sum_{k=1}^{m} g(\mathbf{x}; T_k, M_k), \tag{1}$$

where $f(\mathbf{x})$ is the overall fit of the model: the sum of m regression trees, $\mathbf{x}$ are the predictor variables, T_k is the k$^{\text{th}}$ tree and M_k is the collection of leaf parameters within the k$^{\text{th}}$ tree, i.e. the collection of values assigned to its terminal nodes (Chipman et al., 2006, 1998, 2010; Hill et al., 2020; James et al., 2021). The data are assumed to arise from a model with additive normally distributed errors: $Y = \sum_{k=1}^{m} g(\mathbf{x}; T_k, M_k) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$. 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 variation in the outcome variables to prevent overfitting. The prior is imposed over all parameters of the sum-of-trees model, specifically, $(T_1, M_1), \ldots, (T_m, M_m)$ and σ. However, the specification of the regularization prior is simplified by a series of independence assumptions:

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma) = \Big[\prod_k p(T_k, M_k)\Big] p(\sigma), \tag{2a}$$

$$= \Big[\prod_k p(M_k|T_k) p(T_k)\Big] p(\sigma), \tag{2b}$$

$$p(M_k|T_k) = \prod_j p(\mu_{jk}|T_k), \tag{2c}$$

where $\mu_{jk} \in M_k$. These assumptions state that the trees (T_k), leaf parameters $(\mu_j|T_k)$, and the standard deviation (σ) are independent of each other. Thus, priors only need to be specified for those parameters (Chipman et al., 2006, 1998, 2010; Hill et al., 2020). Chipman et al. (1998) define an independent prior for each tree. The probability that a node at depth d splits is defined as:

$$\alpha(1+d)^{-\beta}, \alpha \in (0,1), \beta \in [0, \infty), \tag{3}$$

where the default specification put forth by Chipman et al. (2006, 2010) is $\alpha = .95$ and $\beta = 2$. This specification sets the probability of a tree with 1, 2, 3, 4, and 5 nodes at .05, .55, .28, .09, and .03 respectively. Thus, smaller trees are favoured. Chipman et al. (2006, 2010) also provide a default specification for the prior for the leaf parameters. They propose to rescale the response value to the interval $[-.5, .5]$. Then, the leaf parameter prior is defined as:

$$\mu_{jk} \sim \mathcal{N}(0, \sigma_\mu^2), \text{with } \sigma_\mu^2 = \frac{.5}{t\sqrt{m}}, \tag{4}$$

where t is a preselected number and m is the number of trees. This prior shrinks the tree parameters μ_{jk} towards 0, decreasing the effect of the individual tree components. If t or m increase, more shrinkage is applied. Using the recommended $k = 2$ by Chipman et al. (2006, 2010) yields a 95% probability that $E[Y|\mathbf{x}]$ is within the range of the rescaled response variable. Chipman et al. (2006, 2010) propose the conjugate inverse chi-square distribution as the prior for the residual standard deviation $\sigma^2 \sim \nu\lambda/\chi^2_\nu$. They represent the degrees of freedom, λ, as the probability that the BART residual standard deviation, σ, is less than the estimated residual standard deviation from a linear regression model, $\hat{\sigma}_{\text{OLS}}$. Their default specification of the hyperparameters is $\nu = 3$ and $\Pr(\sigma < \hat{\sigma}_{\text{OLS}}) = .9$ (Chipman et al., 2006, 1998, 2010; Hill et al., 2020).

BARTs are estimated using the Bayesian back-fitting Markov Chain Monte Carlo (MCMC) algorithm (Chipman et al., 2006, 1998, 2010; Hill et al., 2020; James et al., 2021). Each tree is intialized with a single root node with the mean response value divided by the number of trees ($\hat{f}_k^1(x) = \frac{1}{nK}\sum_{i=1}^{n} y_i$, with sample size n). Then, each pair (T_k, M_k) is updated considering the remaining trees, their associated parameters, and the residual standard deviation (σ) by sampling from the following conditional distribution:

$$(T_k, M_k)|T_{k'}, M_{k'}, \sigma, y. \tag{5}$$

However, this conditional distribution only depends on $(T_{k'}, M_{k'}, y)$ through the partial residuals:

$$r_i = y_i - \sum_{k'<k} \hat{f}^b_{k'}(x_i) - \sum_{k'>k} \hat{f}^{b-1}_{k'}(x_i), \text{with } i = 1, \ldots, n, \tag{6}$$

where $\hat{f}^b_k(x_i)$ is the prediction of the k^{th} tree in the b^{th} iteration for person i and sample size n. Thus, updating each pair (T_k, M_k) symplifies to proposing a new tree fit to the partial residuals, r_i, treating them as the data, by perturbing the tree from the previous iteration. Perturbations entail either *growing*, *pruning*, or *changing* a tree. *Growing* means adding additional splits, *pruning* removes splits, and *changing* changes decision rules. The algorithm stops after the specified number of iterations (Chipman et al., 2006, 1998, 2010; Hill et al., 2020; James et al., 2021).

2.1.2 Multilevel-BART (M-BART)

Chen (2020); Wagner et al. (2020) and Tan et al. (2016) define a M-BART model including a random intercept building on the work of Sela and Simonoff (2012) and Lin and Luo (2019). The M-BART algorithm breaks down the observed variable into fixed and random components. The fixed components are modeled by BART and the random components are modeled by a linear mixed effects model. The estimated random en fixed components are then combined and iteratively updated until convergence under the EM (Expectation-Maximization) framework (Chen, 2020; Tan et al., 2016; Wagner et al., 2020). The BART model (1) can be extended to include a random intercept by:

$$f(\mathbf{x}) = \sum_{k=1}^{m} g(\mathbf{x}; T_k, M_k) + \alpha_j, \tag{7}$$

where, now, $f(\mathbf{x})$ is the overall fit of the model incorporating random intercept α_j for cluster j. For a linear-mixed model, $Y = X\beta + Zu + \epsilon$, the proposed M-BART algorithm is as follows:

1. The random component Zu is intialized as a vector containing deviances between the cluster mean $\bar{Y}_j$ and the overall mean $\bar{Y}$, $Zu = \bar{Y}_j - \bar{Y}$.
2. The fixed effect component, $X\beta$, is extracted by subtracting the random component from the observed data, $Y - Zu$. Then, the fixed effect component is modeled by a single-level BART model with $Y - Zu$ as outcome and all predictors. An indicator variable, I, the mean of the posterior distribution of the predicted value for $\hat{y}$, is generated from this model.
3. Then, the indicator variables is used as the only predictor in a linear mixed model, $Y = I\lambda + Zu + \epsilon$, estimating the random component Zu.
4. The updated random component is then used to update the fixed component in step 2.

Step 2 and 3 of the algorithm are iterated until convergence (Chen, 2020; Tan et al., 2016). Dorie et al. (2024) implemented this algorithim within the R-package `dbarts` with the funtion `rbart_vi()`.

2.2 Simulation study

2.2.1 Data generating mechanism

The population data-generating mechanism was based on the following MLM:

$$y_{ij} = \beta_{0j} + \sum_{k=1}^{7} \beta_{kj} X_{kij} + \epsilon_{ij}, \qquad X_{kij} \sim \mathcal{MVN}(0, \mathbf{\Sigma}_x), \tag{8a}$$

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{2} \gamma_{0q} Z_{qj} + \upsilon_{0j}, \tag{8b}$$

$$\beta_{kj} = \gamma_{k0} + \sum_{q=1}^{2} \gamma_{kq} Z_{qj} + \upsilon_{kj}, \qquad Z_{qj} \sim \mathcal{MVN}(0, \mathbf{\Sigma}_z), \tag{8c}$$

where y_{ij} is a continuous level-1 outcome variable for person i in group j and X_{kij} are k continuous level-1 variables and Z_{qj} are q continuous level-2 variables. They are multivariate normally distributed with means of 0 and variance-covariance matrix $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_z$, respectively:

$$\mathbf{\Sigma}_x = \begin{pmatrix} 6.25 & & & & & & \\ 2.25 & 9 & & & & & \\ 1.5 & 1.8 & 4 & & & & \\ 2.25 & 3.06 & 2.04 & 11.56 & & & \\ 1.5 & 1.8 & 1.2 & 2.04 & 4 & & \\ 1.125 & 1.35 & 0.9 & 1.53 & .9 & 2.25 & \\ 3.3 & 3.96 & 2.64 & 4.488 & 2.64 & 1.98 & 19.36 \end{pmatrix}, \tag{9a}$$

$$\mathbf{\Sigma}_z = \begin{pmatrix} 1 & \\ .48 & 2.56 \end{pmatrix}. \tag{9b}$$

The covariances between the variables were calculated as such that the correlation between the variables was .3, aligned with Cohen's (1990) medium effect size benchmark. The residuals are normally distributed ~~with a mean of 0 and a variance of 25,~~ as

$$\epsilon_{ij} \sim \mathcal{N}(0, 25). \tag{10}$$

The random intercept β_{0j} is determined by the overall intercept γ_{00}, the q group-level effects $\gamma_{0q}Z_{qj}$ and the group-level random residuals υ_{0j}. The overall intercept γ_{00} was set to 10 and the group-level effects γ_{01} and γ_{02} to .5. The k regression coefficients β_{kj} for the continuous variables X_{kij} depend on the intercepts γ_{k0}, the cross-level interactions $\gamma_{kq}Z_{qj}$, and the random slopes υ_{kj}. The k intercepts, or within-group effect sizes, γ_{kj} were varied in the simulations, the cross-level interactions γ_{11}, γ_{21}, and γ_{32} were set to .35.

$$\gamma_{00} = 10, \quad \boldsymbol{\gamma}_{0q} = \begin{pmatrix} .5 \\ .5 \end{pmatrix}, \quad \boldsymbol{\gamma}_{k0} = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \\ \gamma_{40} \\ \gamma_{50} \\ \gamma_{60} \\ \gamma_{70} \end{pmatrix}, \quad \boldsymbol{\gamma}_{kq} = \begin{pmatrix} .35 & 0 \\ .35 & 0 \\ 0 & .35 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{11}$$

The random slopes are multivariate normally distributed with a mean of 0 and a variance-covariance matrix $\mathbf{T}$ shown in equation 12a. Again, the covariances were calculated to yield a correlation of .3.

$$\boldsymbol{\upsilon}_j \sim \mathcal{MVN}(0, \mathbf{T}), \quad \mathbf{T} = \begin{pmatrix} t_{00} & & & & & & & \\ .3 & 1 & & & & & & \\ .3 & .3 & 1 & & & & & \\ .3 & .3 & .3 & 1 & & & & \\ 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{12a}$$

The variance of υ_{0j}, the group-level random residuals t_{00}, were scaled such that the specified ICC values as in table 1 was obtained. The following formula was used to calculate υ_{0j} following the variance decomposition from Rights and Sterba (2019):

$$\text{ICC} = \frac{\boldsymbol{\gamma}^{b\prime}\boldsymbol{\phi}^{b}\boldsymbol{\gamma}^{b} + \tau_{00}}{\boldsymbol{\gamma}^{w\prime}\boldsymbol{\phi}^{w}\boldsymbol{\gamma}^{w} + \boldsymbol{\gamma}^{b\prime}\boldsymbol{\phi}^{b}\boldsymbol{\gamma}^{b} + tr(\mathbf{T}\mathbf{\Sigma}) + \tau_{00} + \sigma^2}, \tag{13}$$

where $\boldsymbol{\gamma}^{b}$ and $\boldsymbol{\gamma}^{w}$ are the level-1 and level-2 fixed effects, $\boldsymbol{\phi}^{b}$ and $\boldsymbol{\phi}^{w}$ are the variance-covariance matrices of a vector with 1, for the intercept, and all level-2 predictors and all cluster-mean-centered level-1 predictors respectively, τ_{00} is the variance of the random intercept, $\mathbf{T}$ is the variance-covariance matrix of the random intercept and slopes, $\mathbf{\Sigma}$ is the variance-covariance matrix of a vector containing 1, for the intercept, and the level-1 variables, and σ^2 is the residual variance. The value for τ_{00} was calculated using

the function `uniroot` in R (R Core Team, 2023).

2.2.2 Simulation design

Table 1 shows the variations considered in the simulation study. They are either grounded in prior research or deemed realistic in real-world applications (Enders et al., 2020, 2018b; Grund et al., 2018b; Gulliford et al., 1999; Hox et al., 2017; Murray and Blitstein, 2003). For each combination of varying parameters, 100 datasets were simulated. 5 different imputation methods were compared:

1. conventional single-level imputation,
2. conventional multilevel imputation,
3. single-level BART imputation,
4. multilevel BART imputation accounting for random intercepts (Chen, 2020; Tan et al., 2016; Wagner et al., 2020),
5. multilevel BART imputation accounting for random effects and cross-level interactions (Dorie et al., 2022).

They were compared on the bias, MSE and coverage of the pooled estimates after fitting the analysis model in equations 8a, 8b, and 8c to the imputed datasets. The analysis models were fitted using the R-package `lme4` (Bates et al., 2015) and the estimates were pooled together using the R-package `mice` (Buuren and Groothuis-Oudshoorn, 2011).

The first and second methods were implemented with the R-packages `mice` and `miceadds` (Robitzsch (<https://orcid.org/0000-0002-8226-3132>) et al., 2024). The conventional single-level imputation were implemented with the imputation method `pmm` (predictive mean matching) and the conventional multilevel imputation was implemented with the `2l.pmm` method for level-1 variables and `2lonly.mean` for level-2 variables.

Table 1: Simulation design

Parameter	Values
Number of clusters (j)	30, 50
Within-cluster sample size (n_j)	15, 50
Intraclass Correlation (ICC)	.5
Missing data mechanism	MAR
Amount of missingness	0%, 50%
Within-group effect size (γ)	.5

The third, single-level BART, fourth, random intercept BART and fifth method, multilevel BART methods were implemented by writing functions in R (R Core Team, 2023) for the package `mice`. The functions `bart` and `rbart_vi` from the `dbarts` package were used for the single-level and random intercept BART imputation methods (Dorie et al., 2024). The function `stan4bart` from the package `stan4bart` was used for the multilevel BART imputation method accounting for random effects and cross-level interactions (Dorie, 2023). The functions were written such that they can be used as imputation methods in the `mice` package.

2.2.3 Missing data generation

As can be seen in table 1, the missing data mechanism was Missing At Random (MAR) and 50% of the data was missing. The missing data was generated using the function `ampute` from the package `mice`.

For the MAR mechanism, patterns of missingness were defined with missing values for 1 to 5 missing values out of 10 veriables per case. These patterns had the same relative frequency of occurence in the data sets. The weighted sum of scores on the observed variables was used to predict the probability of missingness for a case. The weights of the variables *x4* and *z1* were set to 2 and 1.5 respectively when they remained observed in a specific pattern, while the weights of the other variables that remained observed in a specific pattern are set to 1. The type of missingess was set to `''RIGHT''` meaning that cases with a higher weighted sum of scores had a higher probability of becoming incomplete. So, this means that cases with higher values on *x4* and *z1* were more likely to become incomplete.

referee ook aan Oberman & Vink we doen niet veel met suggesties :)

2.2.4 Evaluation

The estimates from the analysis models were evaluated in terms of relative bias and mean squared error (MSE) and coverage of 95% confidence intervals (Morris et al., 2019): geen ci's?

$$\text{Bias} = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_t - \theta), \tag{14a}$$

$$\text{MSE} = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_t - \theta)^2, \tag{14b}$$

$$\text{Coverage} = \Pr(\hat{\theta}_{\text{low,i}} \leq \theta \leq \hat{\theta}_{\text{upp,i}}) = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} 1(\hat{\theta}_{\text{low,i}} \leq \theta \leq \hat{\theta}_{\text{upp,i}}), \tag{14c}$$

where $\hat{\theta}_t$ is the estimated parameter in simulation t, θ is the true value, and n_{sim} is the number of simulated datasets. The lower and upper bounds of the 95% confidence intervals are denoted as $\hat{\theta}_{\text{low,i}}$ and $\hat{\theta}_{\text{upp,i}}$ respectively. The coverage is the proportion of the 95% confidence intervals that contain the true value.

3 Results

4 Discussion

5 Conclusion

6 Appendix

Listing 1: Imputation function for single-level BART

```
mice.impute.bart <- function(y, ry, x, wy = NULL, use.matcher = FALSE, donors = 5L,
...) {
    install.on.demand("dbarts", ...)
    if (is.null(wy)) {
        wy <- !ry
    }

    # Parameter estimates
    fit <- dbarts::bart(x, y, keeptrees = TRUE, verbose = FALSE)

    yhatobs <- fitted(fit, type = "ev", sample = "train")[ry]
    yhatmis <- fitted(fit, type = "ev", sample = "train")[wy]

    # Find donors
    if (use.matcher) {
        idx <- matcher(yhatobs, yhatmis, k = donors)
    } else {
        idx <- matchindex(yhatobs, yhatmis, donors)
    }

    return(y[ry][idx])
}
```

Listing 2: Imputation function for random intercept BART

```
mice.impute.2l.rbart <- function(y, ry, x, wy = NULL, type, use.matcher = FALSE,
donors = 5L, ...) {
install.on.demand("dbarts", ...)
if (is.null(wy)) {
    wy <- !ry
}

clust <- names(type[type == -2])
effects <- names(type[type != -2])
X <- x[, effects, drop = FALSE]

model <- paste0(
    "y ~ ", paste0(colnames(X), collapse = " + ")
)

fit <- dbarts::rbart_vi(formula = formula(model), group.by = clust, data = data.
frame(y, x), verbose = FALSE, ...)

yhatobs <- fitted(fit, type = "ev", sample = "train")[ry]
yhatmis <- fitted(fit, type = "ev", sample = "train")[wy]

# Find donors
if (use.matcher) {
    idx <- matcher(yhatobs, yhatmis, k = donors)
} else {
    idx <- matchindex(yhatobs, yhatmis, donors)
}

return(y[ry][idx])
}
```

Listing 3: Imputation function for multilevel BART with random effects and cross-level interactions

```
mice.impute.2l.bart <- function(y, ry, x, wy = NULL, type, intercept = TRUE, use.
matcher = FALSE, donors = 5L, ...) {
install.on.demand("stan4bart", ...)
if (is.null(wy)) {
    wy <- !ry
}

if (intercept) {
    x <- cbind(1, as.matrix(x))
    type <- c(2, type)
```

```
        names(type)[1] <- colnames(x)[1] <- "(Intercept)"
    }

    clust <- names(type[type == -2])
    rande <- names(type[type == 2])
    fixe <- names(type[type > 0])

    lev <- unique(x[, clust])

    X <- x[, fixe, drop = FALSE]
    Z <- x[, rande, drop = FALSE]
    xobs <- x[ry, , drop = FALSE]
    yobs <- y[ry]
    Xobs <- X[ry, , drop = FALSE]
    Zobs <- Z[ry, , drop = FALSE]

    # create formula
    fr <- ifelse(length(rande) > 1,
        paste0("+ (1 +", paste(rande[-1L], collapse = "+")),
        " + (1 "
    )
    randmodel <- paste0(
        "y ~ bart(", paste0(fixe[-1L], collapse = " + "), ")",
        fr, "| ", clust, ")"
    )
fit <- eval(parse(text = paste("stan4bart::stan4bart(", randmodel,
    ", data = data.frame(y, x),
        verbose = -1,
        bart_args = list(k = 2.0, n.samples = 1500L, n.burn = 1500L, n.thin = 5L))",
    collapse = ""
)))

    yhatobs <- fitted(fit, type = "ev", sample = "train")[ry]
    yhatmis <- fitted(fit, type = "ev", sample = "train")[wy]

    # Find donors
    if (use.matcher) {
        idx <- matcher(yhatobs, yhatmis, k = donors)
    } else {
        idx <- matchindex(yhatobs, yhatmis, donors)
    }

    return(y[ry][idx])
}
```

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.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and for the Alzheimer's Disease Neuroimaging Initiative* (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1).

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., George, E., and Mcculloch, R. (2006). Bayesian Ensemble Learning. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

Cohen, J. (1990). Statistical power analysis for the behavioral sciences. *Computers, Environment and Urban Systems*, 14(1):71.

Dong, M. and Mitani, A. (2023). Multiple imputation methods for missing multilevel ordinal outcomes. *BMC Medical Research Methodology*, 23(1):112.

Dorie, V. (2023). *Stan4bart: Bayesian Additive Regression Trees with Stan-Sampled Parametric Extensions*.

Dorie, V., Chipman, H., McCulloch, R., Dadgar, A., Team, R. C., Draheim U., G., Bosmans, M., Tournayre, C., Petch, M., Valle, R. d. L., Johnson G., S., Frigo, M., Zaitseff, J., Veldhuizen, T., Maisonobe, L., Pakin, S., and Daniel G., R. (2024). Dbarts: Discrete Bayesian Additive Regression Trees Sampler.

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., Du, H., and Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, 25(1):88–112.

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. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, 48(2):640–649.

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.

Gulliford, M., Adams, G., Ukoumunne, O., Latinovic, R., Chinn, S., and Campbell, M. (2005). Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, 58(3):246–251.

Gulliford, M. C., Ukoumunne, O. C., and Chinn, S. (1999). Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, 149(9):876–883.

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.

Lee, D., Carpenter, B., Li, P., Morris, M., Betancourt, M., Maverickg, Brubaker, M., Trangucci, R., Inacio, M., Kucukelbir, A., Buildbot, S., Bgoodri, Seantalts, Arnold, J., Tran, D., Hoffman, M., Margossian, C., Modrák, M., Adler, A., Sakrejda, K., Stukalov, A., Lawrence, M., Goedman, R. J., Van Horn, K. S., Vehtari, A., Gabry, J., Casallas, J. S., and Bales, B. (2017). Stan-dev/stan: V2.17.1. Zenodo.

Lin, S. and Luo, W. (2019). A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research*, 54(4):578–592.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

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.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, pages 538–558.

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.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Murray, D. M. and Blitstein, J. L. (2003). Methods To Reduce The Impact Of Intraclass Correlation In Group-Randomized Trials. *Evaluation Review*, 27(1):79–103.

Quartagno, M. and Carpenter, J. R. (2022). Substantive model compatible multilevel multiple imputation: A joint modeling approach. *Statistics in Medicine*, 41(25):5000–5015.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Resche-Rigon, M. and White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649.

Rights, J. D. and Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, 24(3):309–338.

Robitzsch (<https://orcid.org/0000-0002-8226-3132>), A., Grund (<https://orcid.org/0000-0002-1290-8986>), S., and Henke, T. (2024). Miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'.

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.

Sela, R. J. and Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2):169–207.

Shieh, G. (2012). A comparison of two indices for the intraclass correlation coefficient. *Behavior Research Methods*, 44(4):1212–1223.

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.

Taljaard, M., Donner, A., and Klar, N. (2008). Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. *Biometrical Journal*, 50(3):329–345.

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.