Novel Applications and Extensions for Bayesian Additive Regression Trees (BART) in Prediction, Imputation, and Causal Inference

by

Yaoyuan Vincent Tan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Michael Elliott, Chair
Research Associate Professor Carol Flannagan
Associate Professor Jian Kang
Professor Brisa Sánchez
Professor Kerby Shedden

Yaoyuan Vincent Tan

vincetan@umich.edu

ORCID: 0000-0001-5950-9846

DEDICATION

To my Mom, Poh Wan Oh, who has been my emotional support throughout my MS and PhD studies as well as my aunt, Poh Kian Oh, for the same reasons.

ACKNOWLEDGEMENTS

This dissertation was written with extensive contributions and support from my thesis chair Dr. Michael R. Elliott as well as valuable comments from the members of my thesis committee Drs. Carol A.C. Flannagan, Jian Kang, Brisa Sánchez, and Kerby Shedden. Any mistakes or typographical errors are my own.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

The Bayesian additive regression trees (BART) is a method proposed by Chipman et al. (2010) that can handle non-linear main and multiple-way interaction effects for independent continuous or binary outcomes. It has enjoyed much success in areas like causal inference, economics, environmental sciences, and genomics. However, extensions of BART and application of these extensions are limited. This thesis discusses three novel applications and extensions for BART.

We first discuss how BART can be extended to clustered outcomes by adding a random intercept. This work was motivated by the need to accurately predict driver behavior using observable speed and location information with application to communication of key human-driver intention to nearby vehicles in traffic. Although our extension can be considered a special case of the spatial BART (Zhang et al., 2007), our approach differs by providing a relatively simple algorithm that allows application to clustered binary outcomes.

We next focus on the use of BART in missing data settings. Doubly robust (DR) methods allow consistent estimation of population means when either non-response propensity or modeling of the mean of the outcome is correctly specified. Kang and Schafer (2007) showed that DR methods produce biased and inefficient estimates when both propensity and mean models are misspecified. We consider the use of BART for modeling means and/or propensities to provide a “robust-squared” estimator that reduces bias and improves efficiency. We demonstrate this result, using simulations, for the two commonly used DR methods: Augmented Inverse Probability Weighting (AIPWT, Robbins et al., 1994) and penalized splines of propensity prediction (PSPP, Zhang and Little, 2009). We successfully applied our proposed model to two national crash datasets to impute missing change in deceleration values (delta-v) and missing Blood Alcohol Concentration (BAC) levels respectively.

Our final effort considers how a negative wealth shock (sudden large decline in wealth) affects the cognitive outcome of late middle aged US adults using the Health Retirement Study, a longitudinal study of US adults, enrolled at age 50 and older and surveyed biennially since 1992. Our analysis faced three issues: lack of randomization, confounding by indication, and censoring of the cognitive outcome by a substantial number of deaths in our subjects. Marginal structural models (MSM), a commonly used method to deal with censoring by death, is arguably inappropriate because it upweights subjects who are more likely to die, creating a pseudo-population which resembles one where death is absent. We propose to compare the negative wealth shock effect only among subjects who survived under both sets of treatment regimens – a special case of principal stratification (Frangakis and Rubin, 2002). Because the counterfactual survival status would be unobserved, we imputed their survival status and restrict analysis to subjects who were observed and predicted to survive under both treatment regimes. We used a modified version of penalized spline of propensity methods in treatment comparisons (PENCOMP, Zhou et. al, 2018) to obtain a robust imputation of the counterfactual cognitive outcomes. Finally, we consider several possible extensions of these efforts for future work.

CHAPTER I

Introduction

Since its introduction in 2007 and formal publication in 2010, Bayesian additive regression trees (BART) has enjoyed much success in a variety of applications including biomarker discovery in proteomic studies (*Hernández et al.*, 2015), estimating indoor radon concentrations (*Kropat et al.*, 2015), estimation of causal effects (*Leonti et al.*, 2010), genomic studies (*Liu et al.*, 2010), hospital performance evaluation (*Liu et al.*, 2015), prediction of credit risk (*Zhang and Härdle*, 2010), predicting power outages during hurricane events (*Nateghi et al.*, 2011), prediction of trip durations in transportation (*Chipman et al.*, 2010a), and somatic prediction in tumor experiments (*Ding et al.*, 2012). BART has also been extended to survival outcomes (*Bonato et al.*, 2011; *Sparapani et al.*, 2016), multinomial outcomes (*Kindo et al.*, 2016; *Agarwal et al.*, 2013), and heterogeneous outcomes (*Green and Kern*, 2012).

The primary reason for BARTs success is its ability to model non-linear main and multiple-way interaction effects without having to specify the type of non-linear or interaction mechanism. BART estimates multiple-way interactions 'automatically' by using regression trees which, in its simplest form (a constant mean parameter at the terminal nodes), can be viewed as an analysis of variance (ANOVA) model. To estimate the non-linear effects, BART uses a sum of regression trees. As the number of regression trees used in the sum increases, the non-linear effect estimation by BART improves. To keep BART from over-fitting, a strong prior is then placed on the tree structure of each regression tree to keep trees from growing too deep or too 'bushy' (trees with many terminal nodes).

Despite the flexibility, BART is still mostly applied to independent continuous or binary outcomes. Extensions and application of BART to situations outside of the independent continuous or binary outcomes setup are scarce. Two exceptions are *Zhang et al.* (2007), who extended BART using a spatial random intercept to merge two datasets in a statistical matched problem (*Rässler*, 2002) and *Low-Kam et al.* (2015), who modeled their terminal nodes of the regression tree as a cubic splines regression and used an autoregressive covariance matrix with truncated support on $[0, 1]$ to account for the correlation in their outcomes. These examples address complex extensions of BART to correlated continuous outcomes. Hence, in Chapter III of my thesis, I extended BART to correlated binary outcomes. For Chapter IV and V, I considered applications of BART to issues in the area of missing data and causal inference for longitudinal studies respectively.

I begin with a review chapter, where explicit details of how BART is formulated and implemented are discussed. Using a simple sum of two regression trees as an illustration, we will also attempt to answer a frequently asked question: "What is a sum of regression trees?" Included in this review chapter is also a brief discussion of why we think that application and extension of BART to models outside of the independent continuous and binary outcomes setting are lacking.

My next chapter was motivated by a project where the main aim was to determine whether a human driven vehicle would stop at an intersection before executing a left-turn. To answer this question, we used data where drivers would drive cars fitted with devices to capture various vehicle dynamics like speed, acceleration, turn signal use, etc. We used the vehicle speed collected to construct a prediction model to determine whether a driver would stop at an intersection before executing a left-turn. Preliminary work suggested that BART performed better and was more stable compared to many state-of-the-art machine learning methods, for example, Super Learner (*van der Laan and Polley*, 2010). Unfortunately, BART was designed for independent outcomes but in our data, each driver could take multiple left turns creating correlation among our binary outcomes. Thus far, there has been no literature extending BART to handle correlated binary outcomes. Hence, we introduced a random intercept to BART to handle clustered binary outcomes. The crucial idea lies in the fact that given a draw of the random intercept, the resulting model is once again BART and the BART algorithm can be applied to estimate the remaining parameters. We found that our proposed method, which we call "random intercept BART (riBART)", produced better empirical prediction properties compared to BART without the random intercept in simulations with correlated continuous or binary outcomes and when applied to our data.

Chapter IV focuses on the area of missing data. Under the missing at random (MAR) assumption, doubly robust (DR) estimators provide a consistent estimate of the mean when either the mean or propensity model is correctly specified. Unfortunately, *Kang and Schafer* (2007) showed using a simulation example that DR estimators could be highly biased and inefficient when both the propensity and mean model are modestly misspecified. We recognized that the misspecification of the propensity and mean model in Kang and Schafer's example mainly comes from the fact that common regression methods have difficulty in specifying a model that can handle non-linear main and multiple-way interaction effects. Hence, we propose to replace the usual regression models in DR estimators with BART and investigate whether such a strategy would improve the bias and efficiency of common DR estimators. We found that by replacing the model specification of the various DR estimators with BART greatly improved the robustness of these estimators to model misspecification. In addition, when applied to two publicly available datasets, we found that by comparing our proposed estimator with existing DR estimators, we could get a sense of the relationship of the outcome of interest with the various covariates in the data.

In Chapter V we turn our attention to a causal inference problem in the context of longitudinal studies. This work was motivated by the Health and Retirement Survey (*Sonnega et al.*, 2014) which is a longitudinal study of US adults, enrolled at age 50 and older. Enrolled subjects were surveyed biennially starting from 1992 with detailed modules on financial status and health. The primary aim of this work was to determine how the cognitive ability of late middle aged US adults is affected by a negative wealth shock, i.e. a sudden large decline in wealth. We faced three issues in this analysis. First, there is a lack of randomization for which subjects get a negative wealth shock; factors like socio-economic status and gender are likely confounders. Second, the risk of receiving a negative wealth shock may depend on prior cognitive ability, a situation commonly termed as "confounding by indication". Finally, and most importantly, death occurs at a 13% higher rate during follow-up in our data, causing a large proportion of our outcomes to be censored. A common approach is to employ Marginal Structural Models (MSM, *Robins et al.*, 2000) which accounts for confounding by indication and censoring by death by weighting using the inverse probability of the treatment received based on the previous values of the time-varying covariates and outcomes and inverse probability of death respectively. The issue with this approach – perhaps much under appreciated – is that by weighting using the inverse probability of death, subjects who are more likely to die would be upweighted creating a pseudo-population which resembles one where death is absent over time (*Chaix et al.*, 2012). We propose to compare the effect of a negative wealth shock on cognitive outcome only among subjects who would potentially survive under both sets of treatment regimes, a special case of principal stratification (*Frangakis and Rubin*, 2002). Because the survival status of the counterfactuals (for example, negative wealth shock survival status of subjects who did not get a negative wealth shock and vice versa) are unobserved, we imputed their survival status and restricted analysis to subjects who were observed and predicted to have survived. We then modified the penalized spline of propensity methods in treatment comparisons (PENCOMP, *Zhou et al.*, 2018) using BART to impute the counterfactual cognitive ability among this restricted set. This modified version of PENCOMP is doubly robust and eases the model specification burden on the researcher. Simulation studies suggested that our proposed method worked better than existing methods. Results from our data analysis also suggested a slightly different estimate of the effect of a negative wealth shock on cognitive ability compared to MSM.

CHAPTER II

Review

2.1 Bayesian additive regression trees

We next review in detail the Bayesian additive regression trees (BART) model proposed by *Chipman et al.* (2010b) for independent continuous and binary outcomes. Included in this review is a discussion of what a regression tree is and what a "sum of regression trees" mean. We also discuss how the prior distribution and hyperparameters are set as well as how the posterior distribution of BART is calculated.

2.2 Setup

Suppose we have n subjects indexed by k and we have outcomes Y_k. For continuous outcomes, $Y_k \in \mathbb{R}$, while for binary outcomes, $Y_k \in \{0, 1\}$. In addition to the outcomes, we have p predictors/covariates notated as $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^T$. The objective of BART is to estimate a flexible model to fit the following problem

$$Y_k = f(\mathbf{X}_k) + \epsilon_k \tag{2.1}$$

where $\epsilon_k \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$.

2.3 Continuous outcomes

2.3.1 Model and regression trees

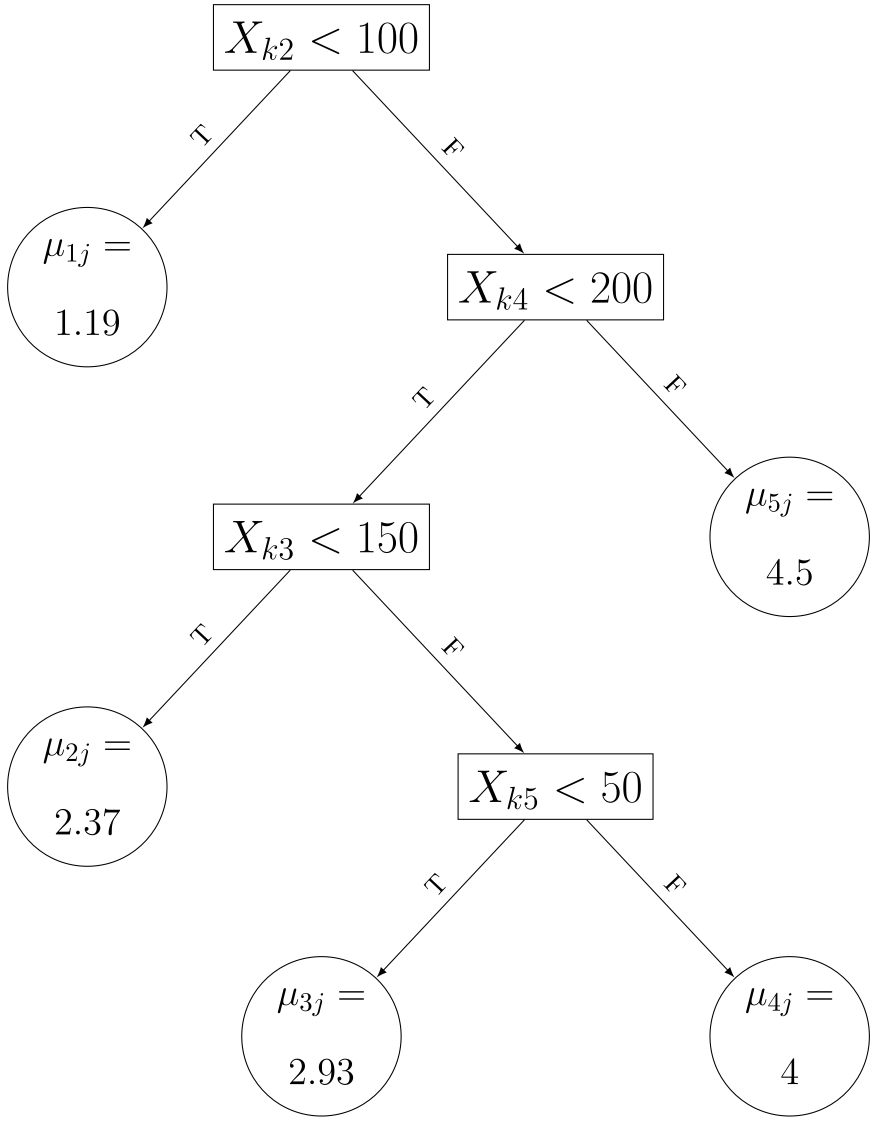
For continuous outcomes, BART estimates equation (2.1) as

$$Y_k = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k \quad \epsilon_k \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{2.2}$$

where T_j is the j^{th} binary tree structure and $\mathbf{M}_j = (\mu_{1j}, \ldots, \mu_{b_j j})^T$ is the set of b_j terminal node parameters associated with tree structure T_j. Typically, the number of trees m is fixed and no prior distribution is placed on m. Chipman et. al. suggested fixing m at 200 as this performs well in many situations. Alternatively, they suggested using cross-validation to determine m.

The binary tree T_j is made up of both internal nodes and terminal nodes. At each internal node, there is a decision rule that splits estimation of the mean of Y_k depending on the covariates $\mathbf{X}_k$. For example in Figure 2.1, the first internal node at the top of the tree drops the mean to the left if the corresponding covariate $X_{k2} < 100$ or to the right if $X_{k2} \geq 100$. At a terminal node (a node with no decision rules to split an outcome), the sample mean of the outcomes allocated to the terminal node can be calculated to obtain the parameter μ_{ij} at the terminal node. Thus, $g(\mathbf{X}_k, T_j, \mathbf{M}_j)$ can be viewed as the j^{th} function that assigns the mean μ_{ij} to the k^{th} outcome, Y_k.

Figure 2.1: Example of a regression tree where μ_{ij} is the mean parameter of the i^{th} node for the j^{th} regression tree.



We may view the regression tree in Figure 2.1 as an ANOVA model because it can be similarly expressed as

$$\begin{aligned} Y_k = {} & \mu_{1j} I\{X_{k2} < 100\} + \mu_{2j} I\{X_{k2} \geq 100\} I\{X_{k4} < 200\} I\{X_{k3} < 150\} \\ & + \mu_{3j} I\{X_{k2} \geq 100\} I\{X_{k4} < 200\} I\{X_{k3} \geq 150\} I\{X_{k5} < 50\} \\ & + \mu_{4j} I\{X_{k2} \geq 100\} I\{X_{k4} < 200\} I\{X_{k3} \geq 150\} I\{X_{k5} \geq 50\} \\ & + \mu_{5j} I\{X_{k2} \geq 100\} I\{X_{k4} \geq 200\} + \epsilon_k \end{aligned}$$

where $I\{.\}$ is the indicator function and $\epsilon_k \overset{i.i.d.}{\sim} N(0, \sigma^2)$. This representation as an ANOVA model clearly shows how a regression tree handles multiple-way interactions.

In equation (2.2), note that we have a sum of $g(\mathbf{X}_k, T_j, \mathbf{M}_j)$ or, a sum of regression trees. What is a sum of regression trees? We attempt to explain this using a simplified example. Suppose $p = 3$, $n = 10$, and we have the following data.

Table 2.1: Example data to explain sum of regression trees.

k	$\mathbf{Y}$	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$
1	Y_1	-182	235	-333
2	Y_2	54	339	244
3	Y_3	-106	-50	-682
4	Y_4	-80	-62	-320
5	Y_5	-123	198	-77
6	Y_6	175	108	-46
7	Y_7	-44	11	136
8	Y_8	-131	-10	-70
9	Y_9	-56	68	257
10	Y_{10}	7	324	282

Suppose again that we used two regression trees to fit this data i.e. $m = 2$, and we have the following two regression tree structures estimated in one of the Monte

Carlo Markov Chain (MCMC) draws (See Figures 2.2 and 2.3).

Figure 2.2: Regression tree, $j = 1$.

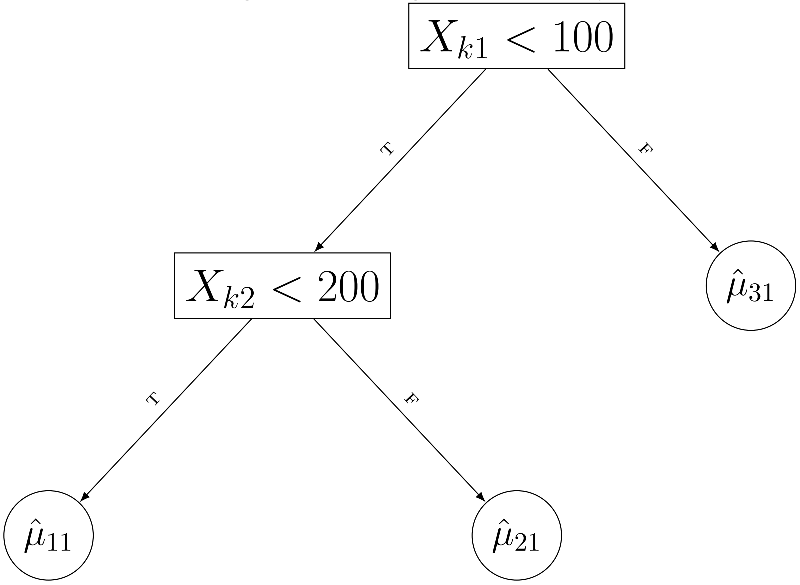
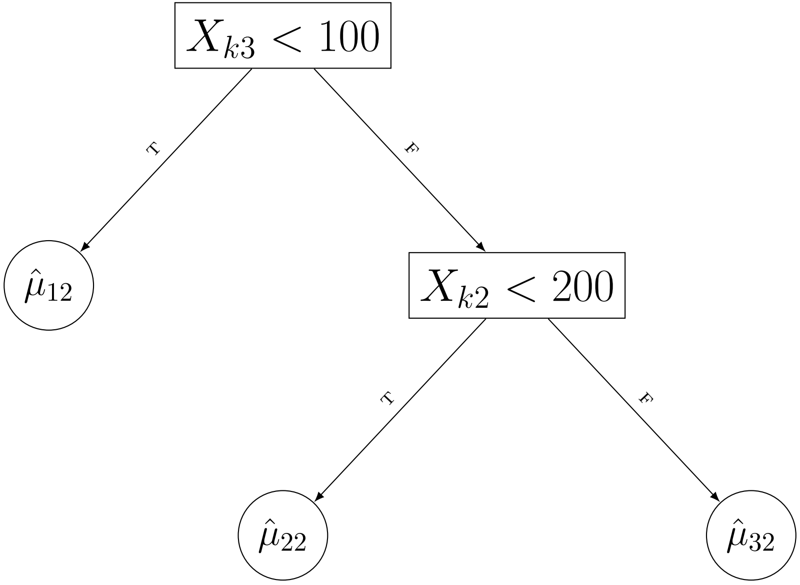


Figure 2.3: Regression tree, $j = 2$.



For this hypothetical example, the resulting posterior estimation of $\sum_{j=1}^{2} g(\mathbf{X}_k, T_j, \mathbf{M}_j)$ can be summarized as follows

Table 2.2: Posterior estimation for $\sum_{j=1}^{2} g(\mathbf{X}_k, T_j, \mathbf{M}_j)$

k	$\mathbf{Y}$	$g(\mathbf{X}, T_1, \mathbf{M}_1)$	$g(\mathbf{X}, T_2, \mathbf{M}_2)$	$\sum_{j=1}^{2} g(\mathbf{X}, T_j, \mathbf{M}_j)$
1	Y_1	$\hat{\mu}_{21}$	$\hat{\mu}_{12}$	$\hat{\mu}_{21} + \hat{\mu}_{12}$
2	Y_2	$\hat{\mu}_{21}$	$\hat{\mu}_{22}$	$\hat{\mu}_{21} + \hat{\mu}_{22}$
3	Y_3	$\hat{\mu}_{11}$	$\hat{\mu}_{12}$	$\hat{\mu}_{11} + \hat{\mu}_{12}$
4	Y_4	$\hat{\mu}_{11}$	$\hat{\mu}_{12}$	$\hat{\mu}_{11} + \hat{\mu}_{12}$
5	Y_5	$\hat{\mu}_{11}$	$\hat{\mu}_{12}$	$\hat{\mu}_{11} + \hat{\mu}_{12}$
6	Y_6	$\hat{\mu}_{31}$	$\hat{\mu}_{12}$	$\hat{\mu}_{31} + \hat{\mu}_{12}$
7	Y_7	$\hat{\mu}_{11}$	$\hat{\mu}_{22}$	$\hat{\mu}_{11} + \hat{\mu}_{22}$
8	Y_8	$\hat{\mu}_{11}$	$\hat{\mu}_{12}$	$\hat{\mu}_{11} + \hat{\mu}_{12}$
9	Y_9	$\hat{\mu}_{11}$	$\hat{\mu}_{22}$	$\hat{\mu}_{11} + \hat{\mu}_{22}$
10	Y_{10}	$\hat{\mu}_{21}$	$\hat{\mu}_{32}$	$\hat{\mu}_{21} + \hat{\mu}_{32}$

where $\hat{\mu}_{ij} \sim h(R_{k_1 j} + R_{k_2 j} + \ldots + R_{k_{n_i}, j}, \theta)$, with $h(.)$ being the posterior distribution of μ_{ij}, θ being the set of prior hyperparameters for μ_{ij}, $R_{kj} = Y_k - \sum_{l \neq j} g(\mathbf{X}_k, T_l, \mathbf{M}_l)$ being the residual data taken in by $h(.)$ to obtain the posterior distribution of μ_{ij}, and n_i being the number of residuals R_{kj} allocated to the terminal node μ_{ij} by the j^{th} regression tree. For example, $\hat{\mu}_{21} \sim h(R_{11} + R_{21} + R_{10,1}, \theta)$ with $R_{11} = Y_1 - \hat{\mu}_{12}$, $R_{21} = Y_2 - \hat{\mu}_{22}$, and $R_{10,1} = Y_{10} - \hat{\mu}_{32}$; $\hat{\mu}_{12} \sim h(R_{12} + R_{32} + R_{42} + R_{52} + R_{62} + R_{82}, \theta)$, with $R_{12} = Y_1 - \hat{\mu}_{21}$, $R_{32} = Y_3 - \hat{\mu}_{11}$, $R_{42} = Y_4 - \hat{\mu}_{11}$, $R_{62} = Y_6 - \hat{\mu}_{31}$, and $R_{82} = Y_8 - \hat{\mu}_{11}$; etc. Note that during the posterior estimation of $g(\mathbf{X}_k, T_j, \mathbf{M}_j)$ for each j, the residuals $R_{k_1 j}, R_{k_2 j}, \ldots, R_{k_{n_i}, j}$ are used instead of $Y_{k_1}, \ldots, Y_{k_{n_i}}$. Hence, we estimate Y_k using the sum of the allocated parameters $\hat{\mu}_{ij}$ instead of their mean. To obtain $\hat{\mu}_{ij}$, an iterative process with $\frac{\bar{Y}}{m}$ as the initial value is used. From this illustration, it is clear that the sum of regression trees occur at the terminal node parameters and not the tree structure. In addition, as we increase the number of regression trees m to 200, this ‘additive’ property of BART allows estimation of non-linear effects easily without having a need to specify the form of non-linear relationship between the outcomes and predictors.

2.3.2 Prior distribution

In subsection 2.3.1, we assumed that the tree structure was specified. Of course, we would like the data to determine the tree structure. BART does this in a Bayesian framework, first specifying a prior on the tree structure, terminal node parameters, and variance. The joint prior distribution for (2.2) is

$$P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma]. \tag{2.3}$$

Assuming independence of ϵ_k and $(T_j, \mathbf{M}_j)$ and between all m tree structures and terminal node parameters, equation (2.3) can be decomposed as

$$\begin{aligned} P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma] &= [\prod_{j=1}^{m} P(T_j, \mathbf{M}_j)]P(\sigma) \\ &= [\prod_{j=1}^{m} P(\mathbf{M}_j|T_j)P(T_j)]P(\sigma) \\ &= [\prod_{j=1}^{m} \{\prod_{i=1}^{b_j} P(\mu_{ij}|T_j)\}P(T_j)]P(\sigma). \end{aligned}$$

where $i = 1, \ldots, b_j$ indexes the terminal node parameters in tree j. The prior distribution of $\mu_{ij}|T_j$ and σ^2 can be specified as

$$\begin{aligned} \mu_{ij}|T_j &\sim N(\mu_\mu, \sigma_\mu^2), \\ \sigma^2 &\sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2}), \end{aligned}$$

where $IG(\alpha, \beta)$ is the inverse gamma distribution with shape parameter α and rate parameter β. The prior for $P(T_j)$ can be specified using three aspects. The first is the probability that a node at depth $d = 0, 1, 2, \dots$ is an internal node, which is $\alpha(1+d)^{-\beta}$ where $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$. Here, α controls how likely a terminal node in the tree would split, with smaller α implying a lesser likelihood that a terminal node would split, and β controls the number of terminal nodes with a larger β decreasing the number of terminal nodes. The second aspect is the distribution used to choose which covariate is selected for the decision rule in an internal node. The final aspect is the distribution for the value of the selected covariate for the decision rule in an internal node. For the distribution in the second and third aspect of $P(T_j)$, the default distirbution used is the discrete uniform distribution for the available covariates. A more flexible distribution like the multinomial distribution with certain variables or values weighted higher can be used (*Kapelner and Bleich*, 2016).

2.3.3 Hyperparameters

The specification of these priors implies that the following hyperparameters need to be set: α, β, μ_μ, σ_μ, ν, and λ. These hyperparameters are constructed as a mix of apriori fixed and data-driven. For α and β, the default values of $\alpha = 0.95$ and $\beta = 2$ provide a balanced penalizing effect for the probability of a node splitting. For μ_μ and σ_μ, they are set such that $E[Y_k|\mathbf{X}_k] \sim N(m\mu_\mu, m\sigma_\mu^2)$ assigns high probability to the interval $(\min_k(Y_k), \max_k(Y_k))$. This can be achieved by defining v such that $\min_k(Y_k) = m\mu_\mu - v\sqrt{m}\sigma_\mu$ and $\max_k(Y_k) = m\mu_\mu + v\sqrt{m}\sigma_\mu$. For ease of posterior distribution calculation, Y_k is transformed by $\tilde{Y}_k = \frac{Y_k - \frac{\min_k(Y_k) + \max_k(Y_k)}{2}}{\max_k(Y_k) - \min_k(Y_k)}$. This results in $\tilde{Y}_k \in (-0.5, 0.5)$ where $\min_k(Y_k) = -0.5$ and $\max_k(Y_k) = 0.5$. This has the effect of allowing the hyperparamter μ_μ to be set as 0 and σ_μ to be determined as $\sigma_\mu = \frac{0.5}{v\sqrt{m}}$ where v is to be chosen. For $v = 2$, $N(m\mu_\mu, m\sigma_\mu^2)$ assigns a prior probability of 0.95 to the interval $(\min_k(Y), \max_k(Y))$ and is the default value. Finally for ν and λ, the default value for ν is 3 and λ is the value such that $P(\sigma^2 < s^2; \nu, \lambda) = 0.9$ where s^2 is the estimated variance of the residuals from the multiple linear regression with Y_k as the outcomes and X_k as the covariates.

2.3.4 Posterior distribution calculation

The prior distribution and hyperparameters would induce the posterior distribution

$$\begin{aligned} P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma | Y_k] \propto\ & P(Y_k | (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma) \\ & \times P((T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma) \end{aligned}$$

where $P(Y_k | (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma) \sim N(\sum_{j=1}^m g(\mathbf{X}_k, T_j, \mathbf{M}_j), \sigma^2)$ which can be simplified to two major posterior draws using Gibbs sampling. First, draw m successive

$$P[(T_j, \mathbf{M}_j) | T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma] \tag{2.4}$$

for $j = 1, \ldots, m$, where $T_{(j)}$ and $\mathbf{M}_{(j)}$ consist of all the tree structures and terminal nodes except for the j^{th} tree structure and terminal node; then, draw

$$P[\sigma | (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), Y_k] \tag{2.5}$$

from $IG(\frac{\nu+n}{2}, \frac{\nu\lambda + \sum_{k=1}^n (y_k - \sum_{j=1}^m g_k(\mathbf{X}_k, T_j, \mathbf{M}_j))^2}{2})$.

To obtain a draw from (2.4), note that this distribution depends on $(T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma)$ through

$$R_{kj} = Y_k - \sum_{w \neq j} g(\mathbf{X}_k, T_w, \mathbf{M}_w), \tag{2.6}$$

the residuals of the $m-1$ regression sum of trees fit excluding the j^{th} tree. Thus (2.4) is equivalent to the posterior draw from a single regression tree $R_{kj} = g(\mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k$

or

$$P[(T_j, \mathbf{M}_j)|\mathbf{R}_j, \sigma]. \tag{2.7}$$

We can obtain a draw from (2.7) by first integrating out $\mathbf{M}_j$ to obtain $P(T_j|\mathbf{R}_j, \sigma)$. This is possible since a conjugate prior on μ_{ij} was employed. We draw $P(T_j|\mathbf{R}_j, \sigma)$ using a Metropolis-Hastings (MH) algorithm where first, we generate a candidate tree T_j^* for the j^{th} tree with probability distribution $q(T_j, T_j^*)$ and then, we accept T_j^* with probability

$$\alpha(T_j, T_j^*) = \min\{1, \frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} \frac{P(\mathbf{R}_j|X, T_j^*, M_j)}{P(\mathbf{R}_j|X, T_j, M_j)} \frac{P(T_j^*)}{P(T_j)}\}. \tag{2.8}$$

A new tree T_j^* can be proposed given the previous tree T_j by four steps: (i) grow, where a terminal node is split into two new child nodes; (ii) prune, two terminal child nodes immediately under the same non-terminal node are combined together such that their parent non-terminal node becomes a terminal node; (iii) swap, the splitting criteria of two non-terminal nodes are swapped; (iv) change, the splitting criteria of a single non-terminal node is changed. Once we draw $P(T_j|\mathbf{R}_j, \sigma)$, we then draw $P(\mu_{ij}|T_j, \mathbf{R}_j, \sigma) \sim N(\frac{\sigma_\mu^2 \sum_i^{n_i} r_{ij}}{n_i\sigma_\mu^2+\sigma^2}, \frac{\sigma^2\sigma_\mu^2}{n_i\sigma_\mu^2+\sigma^2})$, where r_{ij} is the subset of elements in $\mathbf{R}_j$ allocated to the terminal node parameter μ_{ij} and n_i is the number of r_{ij}s allocated to μ_{ij}.

Complete details for the derivation of $P(\mu_{ij}|T_j, \mathbf{R}_j, \sigma)$, equation (2.5) as well as the explicit formula for equation (2.8) for the grow and prune steps can be found in Appendix A.

2.4 Binary outcomes

For binary outcomes, BART uses the probit link to model the relationship between $\mathbf{X}_k$ and Y_k. Formally,

$$P(Y_k = 1|\mathbf{X}_k) = \Phi[G(\mathbf{X}_k)] \tag{2.9}$$

where $\Phi[.]$ is the cumulative distribution function of a standard normal distribution and

$$G(\mathbf{X}_k) = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j). \tag{2.10}$$

The notation m, T_j, and $\mathbf{M}_j$ are similar to equation (2.2) and m by default is once again set at 200.

Because we employed a probit link, we may view the binary outcomes BART as the continuous outcomes BART with $\sigma \equiv 1$. Hence, only prior distributions for T_j and $\mu_{ij}|T_j$ need to be specified under binary outcomes BART. The same prior distributions as continuous outcomes BART can be used. The α and β hyperparameters are the same but the μ_μ and σ_μ hyperparameters are specified differently from continuous outcomes BART. To set the hyperparameters for μ_μ and σ_μ, Chipman et al. suggests $\mu_\mu = 0$ and $\sigma_\mu = \frac{3}{v\sqrt{m}}$ where $v = 2$ would result in an approximate 95% probability that draws of $G(\mathbf{X}_k)$ will be within $(-3, 3)$.

To draw the posterior distribution of T_j and μ_{ij}, we first use data augmentation (*Tanner and Wong*, 1987; *Albert and Chib*, 1993) to draw a continuous latent variable Z_k given Y_k. *Chipman et al.* (2010b) suggests drawing Z_k as

$$Z_k = \begin{cases} \max(N(G(\mathbf{X}_k), 1), 0) & \text{if } Y_k = 1 \\ \min(N(G(\mathbf{X}_k), 1), 0) & \text{if } Y_k = 0. \end{cases} \tag{2.11}$$

We differ slightly by drawing Z_k as

$$Z_k = \begin{cases} N_{(0,\infty)}(G(\mathbf{X}_k), 1) & \text{if } Y_k = 1 \\ N_{(-\infty,0)}(G(\mathbf{X}_k), 1) & \text{if } Y_k = 0. \end{cases} \tag{2.12}$$

where $N_{(a,b)}(\mu, \sigma^2)$ is the normal distribution with mean μ variance σ^2 truncated to (a, b). We then replace the continuous outcomes Y_k in equations (2.4) to (2.8) with Z_k and σ set to 1. Once the draws of T_js and μ_{ij}s are made, the estimate of $G(\mathbf{X}_k)$

can be updated followed by Z_k. The algorithm then iterates between the draws of Z_k, T_js, and μ_{ij}s until convergence.

2.5 Motivation for re-writing BART code and future work

In summary, the BART algorithm for continuous and binary outcomes can be visualized as follows:

```
INPUT: $Y_k$ outcome and $\mathbf{X}_k$ covariates.
OUTPUT: $\sum_{j=1}^{m} \hat{g}(\mathbf{X}_k, T_j, \mathbf{M}_j)$ and $\hat{\sigma}$ for continuous outcomes, $\hat{G}(\mathbf{X}_k)$ for binary outcomes.
BART algorithm($Y_k, \mathbf{X}_k$){
    1. If outcome is continuous, transform $Y_k$ to the range $(-0.5, 0.5)$. If outcome is binary, draw $Z_k$.
    2. Setup hyperparameters $\alpha$, $\beta$, $\sigma_\mu$, and for continuous outcomes $\nu$ and $\lambda$.
    3. Draw $(T_j, M_j)|T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma$ for $j = 1, \ldots, m$.
        • Draw $P[T_j|T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma]$ using Metropolis-Hastings algorithm.
            – Propose a new tree using either grow, prune, change, or swap.
            – Accept a new tree based on equation (2.8).
        • Draw $P[\mathbf{M}_j|T_j, T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma]$.
    4. If outcome is continuous, draw $P[\sigma|(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), Y_k]$. If outcome is binary, $\sigma$ is fixed at 1.
    5. Repeat steps 3 to 4 using the most updated parameters until convergence. For binary outcomes, update $Z_k$ before repeating steps 3 to 4.
}
```

Based on the above algorithm, there are four publicly available software packages that can implement the BART algorithm. They are

- *BayesTree* from *Chipman et al.* (2010b),
- *bartMachine* from *Kapelner and Bleich* (2016),
- Parallel BART from *Pratola et al.* (2014), and
- *dbarts* from *Chipman et al.* (2015).

The first three packages implement BART as a whole complete function i.e., there are no separate functions for 1-4. *dbarts* allows a single MCMC draw of 3 and 4. It is immediately clear that these implementations of BART are not modular in the sense that it is not easy to manipulate or modify any of the steps and substeps in the algorithm, especially for step 3. Due to this lack of modularity, extensions of BART to other outcomes or applying BART into other research areas would be tedious since the researcher will have to re-write the BART algorithm from scratch when often, an extension will only require a slight modification of one step or substep within the BART algorithm.

In order to provide the researcher flexibility in the implementation of BART, we re-coded the BART algorithm in *R* such that each substep in 3 is a separate function and step 4 is a separate function on its own. For step 3, this means that we have a separate function which can propose a new tree structure and another function which can accept or reject a new tree structure. Once the tree structure is fixed, we then have another function to draw the terminal nodes in the tree structure. Such flexibility can allow researchers to extend BART easily or modify different parts of the BART model to suit their own research application. In addition, by providing the codes in *R*, our implementation allows the researcher to easily follow the BART algorithm. To maintain efficiency, we then used Rcpp to re-write our *R* codes.

2.6 Discussion

In this chapter, we reviewed BART in great detail re-coded the BART algorithm to help us better understand the mechanism of BART. Our codes allows the m drawn tree structures at each MCMC to be extracted, hence, enchancing the interpretability of BART compared to existing methods. In terms of prediction performance compared to other existing machine learning methods like Lasso, Gradient boosting, Neural nets, and Random forests, *Chipman et al.* (2010b) already showed that BART was either comparable or performed better. Literature regarding the computation complexity of BART compared to these machine learning methods is a topic for future investigation.

CHAPTER III

Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian Additive Regression Trees

3.1 Introduction

In transportation statistics, a new area of research brought about by improvements in artificial intelligence and engineering is the creation of the autonomous (self-driving) vehicle. These vehicles have been tested on city streets in certain locations since 2009. A number of companies have deployed or announced plans for deployment of such vehicles (*Google*, 2015; *Mchugh, M.*, 2015; *Davies, A.*, 2015). A major hurdle for self-driving vehicles on public roads is that these vehicles will have to interact with human-driven vehicles for the foreseeable future. Human drivers do not always communicate their plans to other drivers well. For example, when making a turn, the turn signal is the only explicit means of communicating plans, and even they are used with less than perfect reliability. Hence, the ability to deploy driverless vehicles on a large scale will critically depend on the development of a good prediction model for human driving behavior.

Currently, driverless vehicles developed generally use onboard sensors to gather data from their surrounding environment to make driving decisions. We envision in the future that vehicles (both human driven and driverless) would be connected such that a driving intent model could first be evaluated on the human driver's vehicle and subsequently "communicated" to the driverless vehicle enabling it to make a better driving decision. Such vehicle-to-vehicle communication would become increasingly available as technology improves resulting in a connected environment. Under such a connected environment, developing a good prediction model for human driving behavior would make sense especially when the driving pattern of a human driven vehicle depends heavily on the unique tendencies of the human driver.

Building a prediction model that addresses all or most of the human driving behavior and driving intent is a massive and complex task. To keep this paper concise, we focus on the the development of a prediction model for a single driving behavior: whether a human driver would stop at an intersection before executing a left turn. We are particularly interested in left turn stops because in countries with right-side driving, for example, US, left turn crashes can result in severe passenger-side impacts. Since left turn maneuvers already present a challenge for human drivers, we expect this maneuver to present difficulty for the driverless vehicle. Placing this prediction scenario in the context of a connected environment, the driverless vehicle will be evaluating data from the human-driven vehicle, supplied from an adapted version of existing "black-box" technology that would broadcast speed and location information to driverless vehicles. The connected driverless vehicle would then combine this transmitted information together with the data it has gathered from its surrounding environment to make a driving decision.

To develop such a prediction model, we used a naturalistic driving study, the Integrated Vehicle Based Safety System (IVBSS) study *Sayer et al.* (2011). Naturalistic driving studies (including the IVBSS) involve the collection of driving data from vehicles as they are piloted on actual roads. These driving data are collected by a data acquisition system (DAS) installed on a study subject's vehicle or a research vehicle.

Typical data collected include vehicle speed, brake application, and miles traveled.

Prediction models in statistics typically rely on regression models that require estimation of covariate main effects and interactions, and, when predictors are continuous or on a fine ordinal scale, assessment of non-linearities. In the settings where understanding associations or, under appropriate assumptions, causal mechanism between predictors and outcomes are of interest, approximations for non-linearities and averaging over interactions might be used to develop summaries to ease interpretation. In prediction, since obtaining the most accurate forecast is the goal, estimating highly complex non-linearities, including the interactions, is at a premium, as long as these non-linearities are true signals and not noise.

Perhaps the most common method for modeling non-linearity is to use a polynomial transformation for a covariate, usually centered at the mean to reduce correlation. More sophisticated approaches use penalized splines or additive models that only require assumptions of smoothness (existence of derivatives) to obtain consistent estimates of a non-linear trend *Hastie and Tibshirani* (1990); *Ruppert et al.* (2003). Modeling of non-linear interactions between two or more predictors using thin-plate splines *Franke* (1982) can quickly become difficult, suffering from the "curse of dimensionality", as the data required to estimate high-dimensional surfaces become enormous. In the binary outcomes setting, methods such as classification and regression trees (CART; *Breiman et al.*, 1984) as well as more sophisticated machine learning techniques such as artificial neural networks (ANN; *Smith et al.*, 1993) and support vector machines (SVM; *Gammermann*, 2000) are commonly used. Although CART is able to model complex interactions naturally, it faces difficulty when modeling non-linear interactions. In contrast, ANN and SVM excel at modeling non-linearities but may face difficulties when modeling complex interactions.

Because our goal is prediction, we prefer regression methods that are able to account for non-linear main and multiple-way interaction effects. Bayesian additive regression trees (BART; *Chipman et al.*, 2010b) is one such model which allows flexible estiamtion of non-linear main and multiple-way interaction effects without much input from the researcher. Hence, we employed BART to predict whether a human-driven vehicle would stop before executing a left turn at an intersection. However, BART was designed for independent subjects, but we would like to evaluate the tendencies of each driver and decide whether including their tendency would improve the prediction of whether a human-driven vehicle would stop before executing a left turn. We are aware of two papers that extended BART to handle longitudinal or clustered observations: *Zhang et al.* (2007) used a spatial random intercept BART to merge two datasets, and *Low-Kam et al.* (2015) did so in a dose-finding toxicity study. *Zhang et al.* (2007) developed an imputation model for a statistical matching problem *Rässler* (2002) that used BART with a conditional auto-regressive distribution for the random intercept. Since the correlation our dataset was induced by repeated measurements and not spatial effects, the distribution *Zhang et al.* (2007) placed on the random intercept may not be appropriate. Moreover, they did not discuss how their model could be extended to clustered binary outcomes. *Low-Kam et al.* (2015) investigated the associations between the physico-chemical properties of nanoparticles and their toxicity profiles over multiple doses. The complex nature of their goal prompted them to first specify an autoregressive covariance matrix with truncated support on $[0, 1]$ to handle the correlated measurements, and then they specified a conditionally conjugate P-spline prior for the terminal nodes of the regression trees. The complexity of their method makes implementation to our dataset difficult since our outcomes are binary. Neither papers provided convenient software for implementing their methods.

Motivated by the lack of an appropriate and straightforward method to implement BART to handle clustered binary outcomes, we propose an extension of BART to account for longitudinal binary observations. Our proposed method accounts for clustering by adding a random intercept to BART and we call this random intercept

BART (riBART). We proceed by first providing a review of BART in the next section followed by a discussion of how we extended BART to riBART in Section 3. In Section 4, we use a simulation study to compare the performance of riBART against BART, fixed effects BART, and linear regression models when applied to clustered datasets. We implement riBART on our dataset and compare its prediction performance with BART, fixed effects BART, random intercept linear logistic regression, and multiple linear logistic regression in Section 5. Finally, we conclude with a discussion and possible future work in Section 6.

3.2 Bayesian Additive Regression Trees

3.2.1 Continuous outcomes

Denote a continuous outcome Y_k with associated p covariates $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^T$ for $k = 1, \ldots, n$ subjects. BART models the outcome as

$$Y_k = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k \quad \epsilon_k \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{3.1}$$

where T_j is the j^{th} binary tree structure and $\mathbf{M}_j = (\mu_{1j}, \ldots, \mu_{b_j j})^T$ is the set of b_j terminal node parameters associated with tree structure T_j *Chipman et al.* (2010b). $g(\mathbf{X}_k, T_j, \mathbf{M}_j)$ can be viewed as the j^{th} function that assigns the mean μ_{ij} to the k^{th} outcome, Y_k. Typically, the number of trees m is fixed and no prior distribution is placed on m. *Chipman et al.* (2010b) suggested setting $m = 200$ as this performs well in many situations. Alternatively, cross-validation could be used to determine m *Chipman et al.* (2010b).

The joint prior distribution for Eq. (3.1) is $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma]$. Note that by the independence of ϵ_k and $(T_j, \mathbf{M}_j)$ as well as the independence between all m tree structures and terminal node parameters, the joint prior distribution

$P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma]$ can be decomposed as

$$\begin{aligned}
P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma] &= [\prod_{j=1}^{m} P(T_j, \mathbf{M}_j)]P(\sigma) \\
&= [\prod_{j=1}^{m} P(\mathbf{M}_j|T_j)P(T_j)]P(\sigma) \\
&= [\prod_{j=1}^{m}\{\prod_{i=1}^{b_j} P(\mu_{ij}|T_j)\}P(T_j)] \\
&\quad \times P(\sigma).
\end{aligned}$$

where $i = 1, \ldots, b_j$ indexes the terminal node parameters in tree j. This implies that we need to assign priors to T_j, $\mu_{ij}|T_j$, and σ in order to obtain the posterior distributions of T_j, μ_{ij}, and σ. *Chipman et al.* (2010b) suggested the following prior distributions on $\mu_{ij}|T_j$ and σ:

$$\begin{aligned}
\mu_{ij}|T_j &\sim N(\mu_\mu, \sigma_\mu^2), \\
\sigma^2 &\sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2}).
\end{aligned}$$

where $IG(\alpha, \beta)$ is the inverse gamma distribution with shape parameter α and rate parameter β. The prior distribution of $P(T_j)$ can be specified using three aspects: (i) the probability that a node at depth $d = 0, 1, 2, \ldots$ is an internal node given by $\alpha(1 + d)^{-\beta}$ where $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$ so that α controls how likely a terminal node in the tree would split, with a smaller α implying lesser likelihood a terminal node would split, and β controls the number of terminal nodes, and a larger β decreasing the number of terminal nodes; (ii) the distribution used to choose which covariate to be selected for the decision rule in an internal node; and (iii) the distribution for the value of the selected covariate for the decision rule in an internal node. *Chipman et al.* (2010b) suggests a discrete uniform distribution for the available covariates and values in both (ii) and (iii) respectively, although other more flexible distributions could be used *Kapelner and Bleich* (2016).

In *Chipman et al.* (2010b), $\alpha = 0.95$ and $\beta = 2$. For μ_μ and σ_μ, they are set such that $N(m\mu_\mu, m\sigma_\mu^2)$ assigns high probability to the interval $(\min_k(Y_k), \max_k(Y_k))$. This can be achieved by defining v such that $\min_k(Y_k) = m\mu_\mu - v\sqrt{m}\sigma_\mu$ and $\max_k(Y_k) = m\mu_\mu + v\sqrt{m}\sigma_\mu$. For convenience when implementing the posterior draws of T_j and μ_{ij}, *Chipman et al.* (2010b) suggested transforming the observed Y_k to $\tilde{Y}_k = \frac{Y_k - \frac{\min_k(Y_k)+\max_k(Y_k)}{2}}{\max_k(Y_k)-\min_k(Y_k)}$, and then treating $\tilde{Y}_k$ as the outcome. This has the effect of allowing the hyperparameter of μ_μ to be set as $\mu_\mu = 0$ and σ_μ to be set as $\sigma_\mu = \frac{0.5}{v\sqrt{m}}$ where v is to be chosen. For $v = 2$, $N(m\mu_\mu, m\sigma_\mu^2)$ assigns a prior probability of 0.95 to the interval $(\min_k(Y), \max_k(Y))$ and is the suggested value. Finally for ν and λ, *Chipman et al.* (2010b) suggested setting $\nu = 3$ and λ is the value such that $P(\sigma^2 < s^2; \nu, \lambda) = 0.9$ where s^2 is the estimated variance of the residuals from the multiple linear regression with Y_k as the outcomes and $\mathbf{X}_k$ as the covariates.

This setup induces the posterior distribution $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma | Y_k]$ which can be simplified to two major posterior draws using Gibbs sampling. First, draw m successive

$$P[(T_j, \mathbf{M}_j) | T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma] \tag{3.2}$$

for $j = 1, \ldots, m$, where $T_{(j)}$ and $\mathbf{M}_{(j)}$ consist of all the tree structures and terminal nodes except for the j^{th} tree structure and terminal node; and then, draw $P[\sigma | (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), Y_k]$.

To obtain a draw from Eq. (3.2), note that this distribution depends on $(T_{(j)}, \mathbf{M}_{(j)}, Y_k, \sigma)$ through

$$R_{kj} = Y_k - \sum_{w \neq j} g(\mathbf{X}_k, T_w, \mathbf{M}_w), \tag{3.3}$$

the residuals of the $m - 1$ regression sum of trees fit excluding the j^{th} tree. Thus, Eq. (3.2) is equivalent to the posterior draw from a single regression tree $R_{kj} =$

$g(\mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k$ or

$$P[(T_j, \mathbf{M}_j)|R_{kj}, \sigma]. \tag{3.4}$$

We can obtain a draw from Eq. (3.4) by first drawing from $P(T_j|R_{kj}, \sigma)$ using a Metropolis-Hastings (MH) algorithm outlined in *Chipman et al.* (1998). A new tree T_j^* can be proposed given the previous tree T_j by four steps: (i) grow, where a terminal node is split into two new child nodes; (ii) prune, where two terminal child nodes immediately under the same non-terminal node is combined together such that their parent non-terminal node becomes a terminal node; (iii) swap, where the splitting criteria of two non-terminal nodes are swapped; (iv) change, where the splitting criteria of a single non-terminal node is changed. Once we draw $P(T_j|R_{kj}, \sigma)$, we then draw $P(\mu_{ij}|T_j, R_{kj}, \sigma) \sim N(\frac{\sigma_\mu^2 \sum_i^{n_i} r_{ij} + \sigma^2 \mu_\mu}{n_i \sigma_\mu^2 + \sigma^2}, \frac{\sigma^2 \sigma_\mu^2}{n_i \sigma_\mu^2 + \sigma^2})$, where r_{ij} is the subset of elements in R_{kj} allocated to the terminal node with parameter μ_{ij} and n_i is the number of r_{ij}s in R_{kj} allocated to μ_{ij}. Note that $\mu_\mu = 0$ after transformation. Complete details for the derivation of $P(\mu_{ij}|T_j, R_{kj}, \sigma)$ and $P[\sigma|(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), Y_k]$ are provided in the supplementary materials available online. Explicit MH algorithm details for Eq. (3.4) can be found in Appendix A of *Kapelner and Bleich* (2016).

3.2.2 Binary outcomes

Extending BART to binary outcomes involve a modification of Eq. (3.1). First, let

$$G(\mathbf{X}_k) = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j). \tag{3.5}$$

Using the probit formulation, the binary outcomes Y_k can be linked to Eq. (3.5) using $P(Y_k = 1|\mathbf{X}_k) = \Phi[G(\mathbf{X}_k)]$ where $\Phi[.]$ is the cumulative density function of a standard normal distribution. This formulation implicitly assumes that $\sigma \equiv 1$. Assuming once again that all m tree structures and terminal node parameters are independent, this implies that we only need priors for T_j and $\mu_{ij}|T_j$. *Chipman et al.* (2010b) assumes that priors for T_j and μ_{ij} as well as the hyperparameters for α and β are the same as BART for continuous outcomes. However, for the hyperparameters of μ_μ and σ_μ, *Chipman et al.* (2010b) suggested that μ_μ and σ_μ should be chosen such that $G(\mathbf{X}_k)$ is assigned to the interval $(-3, 3)$ with high probability. This can be achieved by setting $\mu_\mu = 0$ and choosing an appropriate v in the formula $\sigma_\mu = \frac{3}{v\sqrt{m}}$. Similar to the continuous outcome case, *Chipman et al.* (2010b) suggested $v = 2$.

To draw from the posterior distribution $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)|Y_k]$, *Chipman et al.* (2010b) proposed the use of data augmentation *Albert and Chib* (1993); *Tanner and Wong* (1987). This method proceeds by first generating a latent variable Z_k according to

$$\begin{aligned}(Z_k|Y_k = 1, \mathbf{X}_k) &\sim N_{(0,\infty)}(G(\mathbf{X}_k), 1)\\ (Z_k|Y_k = 0, \mathbf{X}_k) &\sim N_{(-\infty,0)}(G(\mathbf{X}_k), 1),\end{aligned}$$

where $N_{(a,b)}(\mu, \sigma^2)$ is the truncated normal distribution with mean μ and variance σ^2 truncated to the range (a, b). Once Z_k is drawn, $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)|Z_k]$ is drawn next as in Eq. (3.2) to Eq. (3.4) with the latent variables Z_k replacing Y_k in Eq. (3.2) and σ fixed at 1. Note that at each iteration, $G(\mathbf{X}_k)$ will be updated with the new $(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)$ draws from $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)|Z_k]$ so that an updated draw of the latent variable Z_k can be obtained.

3.3 Random Intercept BART

3.3.1 Continuous outcomes

We now extend BART to account for repeated measurements. We start with the clustered continuous outcomes. We introduce to Eq. (3.1) a random intercept a_k, $k = 1, \ldots, K$. Here, k still indexes the subjects but $i = 1, \ldots, n_k$ indexes the observations within a subject. With the addition of a_k, Eq. (3.1) becomes

$$Y_{ik} = \sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j) + a_k + \epsilon_{ik}, \tag{3.6}$$

where $\epsilon_{ik} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $a_k \overset{i.i.d.}{\sim} N(0, \tau^2)$, and $a_k \perp \epsilon_{ik}$. We decompose the joint prior distribution (assuming σ^2 and τ^2 are a priori independent) as

$$\begin{aligned} P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma, \tau] = {} & [\prod_{j=1}^{m}\{\prod_{l=1}^{b_j} P(\mu_{lj}|T_j)\}P(T_j)] \\ & \times P(\sigma)P(\tau). \end{aligned}$$

Next, we place the same prior distributions as the independent BART model for T_j, $\mu_{lj}|T_j$ (this is μ_{ij} for the independent BART model), and σ^2. The prior distribution of τ^2 could be set as $\sim IG(1, 1)$ although other specifications are definitely possible. We explore some alternatives in our supplementary materials available online. We use the same hyperparameter values for α, β, μ_μ, and ν that *Chipman et al.* (2010b) suggested for the independent BART model. For σ_μ, we found that $\sigma_\mu = \frac{1.96}{v\sqrt{m}}$ worked better for reasons we shall discuss later in this section. For λ, we first estimated the outcomes Y_{ik} using multivariate adaptive regression splines (MARS; *Friedman*, 1991) with $\mathbf{X}_k$ as the predictors. We then estimated an initial random intercept, $\hat{a}_k^{(0)}$, by taking the mean of the MARS residuals for each k. Finally, we obtained an initial estimate of σ^2 using $s^{(0)2} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ik} - \hat{Y}_{ik}^{(0)} - \hat{a}_k^{(0)})^2}{N - N(1 - \sqrt{\frac{RSS}{GCV \times N}})}$, where $N = \sum_{k=1}^{K} n_k$, RSS and GCV are the residual sum of squares and generalized cross-validation value from MARS respectively, and $N(1 - \sqrt{\frac{RSS}{GCV \times N}})$ is the effective number of parameters in MARS. Then λ can be set as the value such that $P(\sigma^2 < s^{(0)2}; \nu, \lambda) = 0.9$. We call this model the random intercept BART (riBART).

To draw from the posterior distribution of riBART, we employ a Metropolis within Gibbs procedure. We first draw the Gibbs sample of σ, τ, and a_k separately from their respective posterior distribution. Then, using the updated a_k, we obtain $\tilde{Y}_{ik} = Y_{ik} - a_k$. Now $\tilde{Y}_{ik}|\mathbf{X}_k$ can be viewed as a BART model. The idea of viewing $\tilde{Y}_{ik}|\mathbf{X}_k$ as a BART model has been discussed in *Zhang et al.* (2007) and *Dorie et al.* (2016). To allow for convenient implementation of the posterior draws of T_j and $\mu_{lj}|T_j$, we transform the outcomes $\tilde{Y}_{ik}$ to $\check{Y}_{ik} = \frac{(2\times 1.96)[\tilde{Y}_{ik} - \frac{\min\limits_{i,k}(\tilde{Y}_{ik}) + \max\limits_{i,k}(\tilde{Y}_{ik})}{2}]}{\max\limits_{i,k}(\tilde{Y}_{ik}) - \min\limits_{i,k}(\tilde{Y}_{ik})}$. This transformation produced posterior draws for σ and τ with better repeated sampling properties across the range of our simulation studies compared to the usual transformation employed in BART, and suggests setting $\sigma_\mu = \frac{1.96}{2\sqrt{m}}$ so that $(\min\limits_{i,k}(\tilde{Y}_{ik}), \max\limits_{i,k}(\tilde{Y}_{ik}))$ has a prior probability of 0.95. We suspect this transformation produces better repeated sampling properties for the posterior draws of σ and τ because it controls the range of values $\check{Y}_{ik}$ would vary in. Further investigation beyond the scope of this paper is needed in order to determine why this is the case. After obtaining $\check{Y}_{ik}$, we use $\check{Y}_{ik}$ as the outcome in the BART algorithm to obtain the posterior distribution of T_j. In our implementation, we employed the grow and prune steps for the proposal of a new tree T_j^* for computational ease. Given T_j, we then draw μ_{lj}. Derivation of the Gibbs sampling distributions of σ, a_k, and τ are provided in the supplementary materials available online.

3.3.2 Binary outcomes

Extending riBART to binary outcomes proceed in a similar fashion. We add a_k to Eq. (3.5) to obtain

$$G_a(\mathbf{X}_{ik}) = \sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j) + a_k. \tag{3.7}$$

We once again assume $a_k \sim N(0, \tau^2)$. To link the sum of trees to the binary outcomes Y_{ik}, we use the probit link and write $P(Y_{ik} = 1|\mathbf{X}_{ik}) = \Phi[G_a(\mathbf{X}_{ik})]$. We suggest prior distributions similar to the continuous outcomes riBART for T_j, μ_{lj}, and τ^2. The same hyperparameters in BART for binary outcome can be used for α, β, μ_μ, and σ_μ.

To obtain the posterior draws of T_j, $\mathbf{M}_j$, a_k, and τ^2, we employ the data augmentation method suggested by *Albert and Chib* (1996). First, we draw a latent variable Z_{ik} according to

$$
\begin{aligned}
(Z_{ik}|Y_{ik} = 1, \mathbf{X}_{ik}) &\sim N_{(0,\infty)}(G_a(\mathbf{X}_{ik}), 1) \\
(Z_{ik}|Y_{ik} = 0, \mathbf{X}_{ik}) &\sim N_{(-\infty,0)}(G_a(X_{ik}), 1).
\end{aligned}
$$

We then draw τ followed by a_k. Next, we remove a_k from Z_{ik} to obtain $\tilde{Z}_{ik} = Z_{ik} - a_k$. $\tilde{Z}_{ik}|\mathbf{X}_{ik}$ can now be viewed as a continuous BART model and the usual BART algorithm can be applied with σ fixed at 1. In our implementation, we employed a further transformation of $\tilde{Z}_{ik}$ to $\check{Z}_{ik} = \frac{6[\tilde{Z}_{ik} - \frac{\min_{i,k}(\tilde{Z}_{ik}) + \max_{i,k}(\tilde{Z}_{ik})}{2}]}{\max_{i,k}(\tilde{Z}_{ik}) - \min_{i,k}(\tilde{Z}_{ik})}$. This keeps $\check{Z}_{ik}$ within the range of $(-3, 3)$, which we found produces posterior draws for τ with better repeated sampling properties across the range of our simulation studies. The posterior draw is then completed by updating Z_{ik} using the most recent posterior draws of $(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)$, and a_k.

3.4 Simulation Study

We conducted a simulation study to determine the in-sample performance of riBART compared to three alternative methods on a longitudinal dataset with correlated outcomes. The methods we considered were: (I) BART, (II) riBART, (III) fixed effects BART where variables indicating which row belonged to which subject was added as a predictor in BART, and (IV) multiple linear regression (MLR) for continuous outcomes or multiple linear logistic regression (MLLR) for binary outcomes. We focused on the prediction performance of the models by using the mean squared error (MSE; continuous) and area under the receiver operating characteristic curve (AUC; binary) produced by each model. In addition, we investigated the bias, root mean squared error (RMSE), 95% coverage, and average 95% credible interval length

(AIL) of $\sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j) + a_k$ abbreviated as $g(x) + a_k$ and σ (for continuous correlated outcomes only).

We generated our correlated outcomes dataset by first drawing the predictors using $X_{ikq} \overset{i.i.d.}{\sim}$ Uniform$(0, 1)$, $q = 1, \ldots, 10$. For continuous outcomes, we generated

$$\begin{aligned} Y_{ik} = {} & 10 \sin(\pi X_{ik1} X_{ik2}) + 20(X_{ik3} - 0.5)^2 + 10X_{ik4} \\ & + 5X_{ik5} + a_k + \epsilon_{ik} \end{aligned} \tag{3.8}$$

where $\epsilon_{ik} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $a_k \overset{i.i.d.}{\sim} N(0, \tau^2)$, and $a_k \perp \epsilon_{ik}$. For binary outcomes, we first generated

$$\begin{aligned} G_a(X_{ik}) = {} & 1.35[\sin(\pi X_{ik1} X_{ik2}) + 2(X_{ik3} - 0.5)^2] \\ & - 1.35X_{ik4} - 0.675X_{ik5} + a_k \end{aligned} \tag{3.9}$$

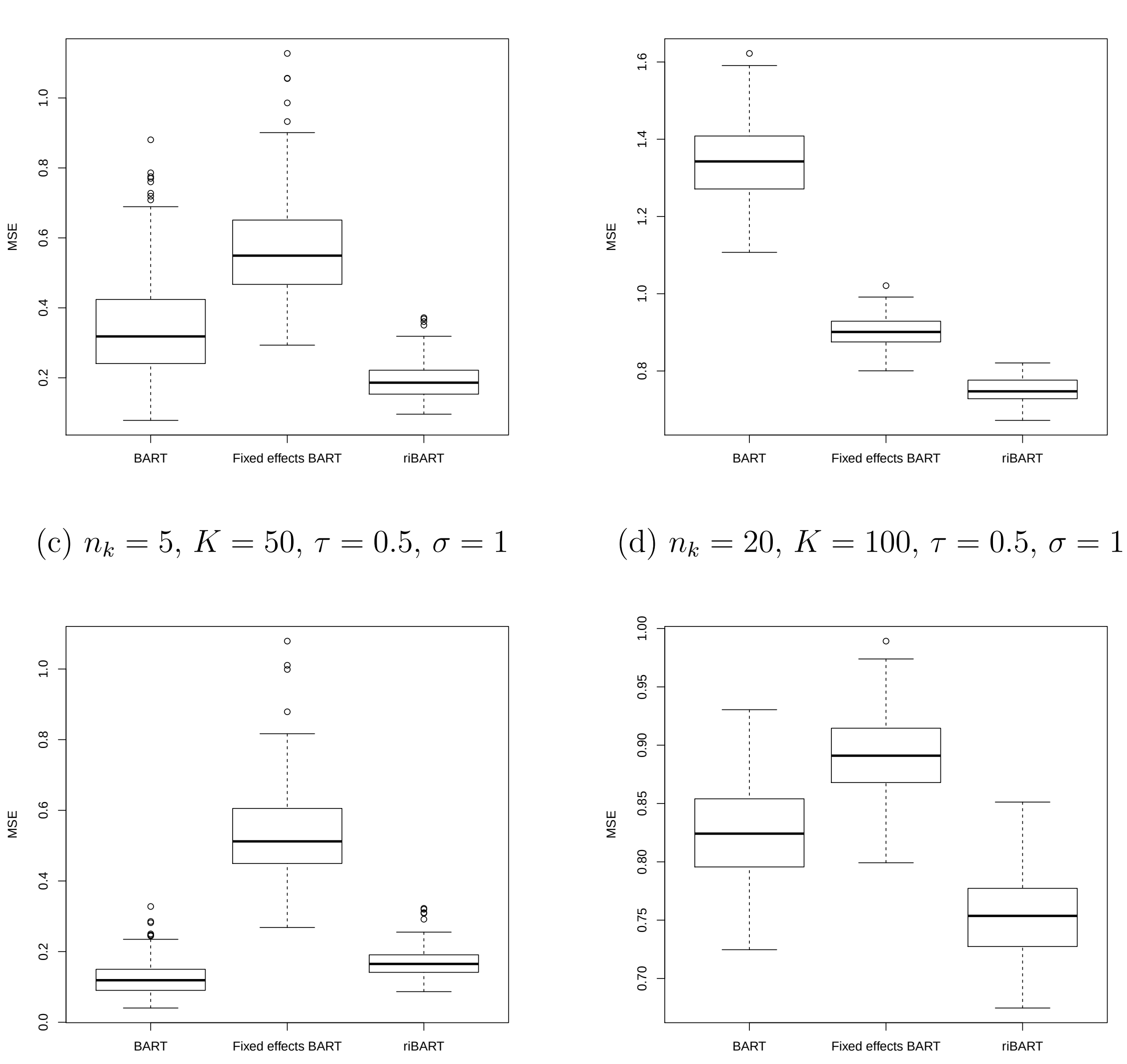
where $a_k \overset{i.i.d.}{\sim} N(0, \tau^2)$. Then, we generated the binary outcomes Y_{ik} by drawing $Z_{ik} \sim N(G_a(\mathbf{X}_{ik}), 1)$ and setting $Y_{ik} = 1$ if $Z_{ik} > 0$, otherwise $Y_{ik} = 0$. Eq. (3.8) and Eq. (3.9) suggest that only the first 5 predictors were important for prediction. The rest of the predictors were “junk” variables.

For the study design, we considered $K = 50$ clusters with $n_k = 5$ observations per cluster and $K = 100$ clusters with $n_k = 20$ observations per cluster. We also considered $\tau = 0.5$ and $\tau = 1$. This produces eight different simulation scenarios summarized in Tables 3.1 and 3.2. For each simulation, we conducted 1,000 burn ins followed by 5,000 posterior draws. Bias, RMSE, 95% coverage, AIL, MSE, and AUC were estimated from 200 simulations for each scenario. All our simulations were done in *R 3.1.1 R Core Team* (2015).

Figure 3.1 shows the boxplots of the MSEs for scenarios 1 to 4 while Figure 3.2 shows the boxplots of the AUCs produced for scenarios 5 to 8. For Figure 3.1, because the boxplots of the MSE for MLR were much larger compared to the rest of the methods, these boxplots were not presented in the manuscript. Interested readers may refer to our supplementary materials available online for the graphs including MLR results. For continuous correlated outcomes, riBART produces a clear advantage compared to BART and fixed effects BART when $K = 100$, $n_k = 20$, and $\tau = 1$. In other simulation scenarios, riBART does not seem to produce lower MSEs compared to BART and fixed effects BART. For binary correlated outcomes, the advantage of BART in terms of producing a better AUC is more apparent. We observed from Figure 3.2 that riBART produces the higher AUC compared to BART, fixed effects BART, and MLLR in all our simulation scenarios. This suggests that for continuous correlated outcomes, riBART may not yield an obvious prediction advantage except when the values of K, n_k, and τ are large. However, for binary correlated outcomes, riBART would produce an obvious prediction advantage regardless of K, n_k, and τ.

In terms of the inference for the parameters $\sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j) + a_k$ and σ, Table 3.1 suggests that for continuous correlated outcomes, the bias and RMSE for all methods would be similar under all scenarios for $g(x) + a_k$. However, the coverage for riBART would be closer to the nominal coverage of 95% under all scenarios. For σ, the bias produced by riBART was usually the smallest and coverage was usually the highest. These results suggest that riBART should be employed for continuous correlated outcomes if inference for $\sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j) + a_k$ or σ are desired. For binary correlated outcomes, the main focus of our paper, Table 3.2 suggests that riBART usually has the smallest bias compared with BART, fixed effects BART, and MLLR under all simulation scenarios. riBART also has the better coverage in our simulation scenario compared to the rest of the methods we considered. These results together with the AUC results from Figure 3.2 suggest that for binary correlated outcomes, riBART should be employed.

Figure 3.1: Boxplots of mean squared error (MSE) for continuous correlated outcomes produced by BART, Fixed effects BART, and riBART.



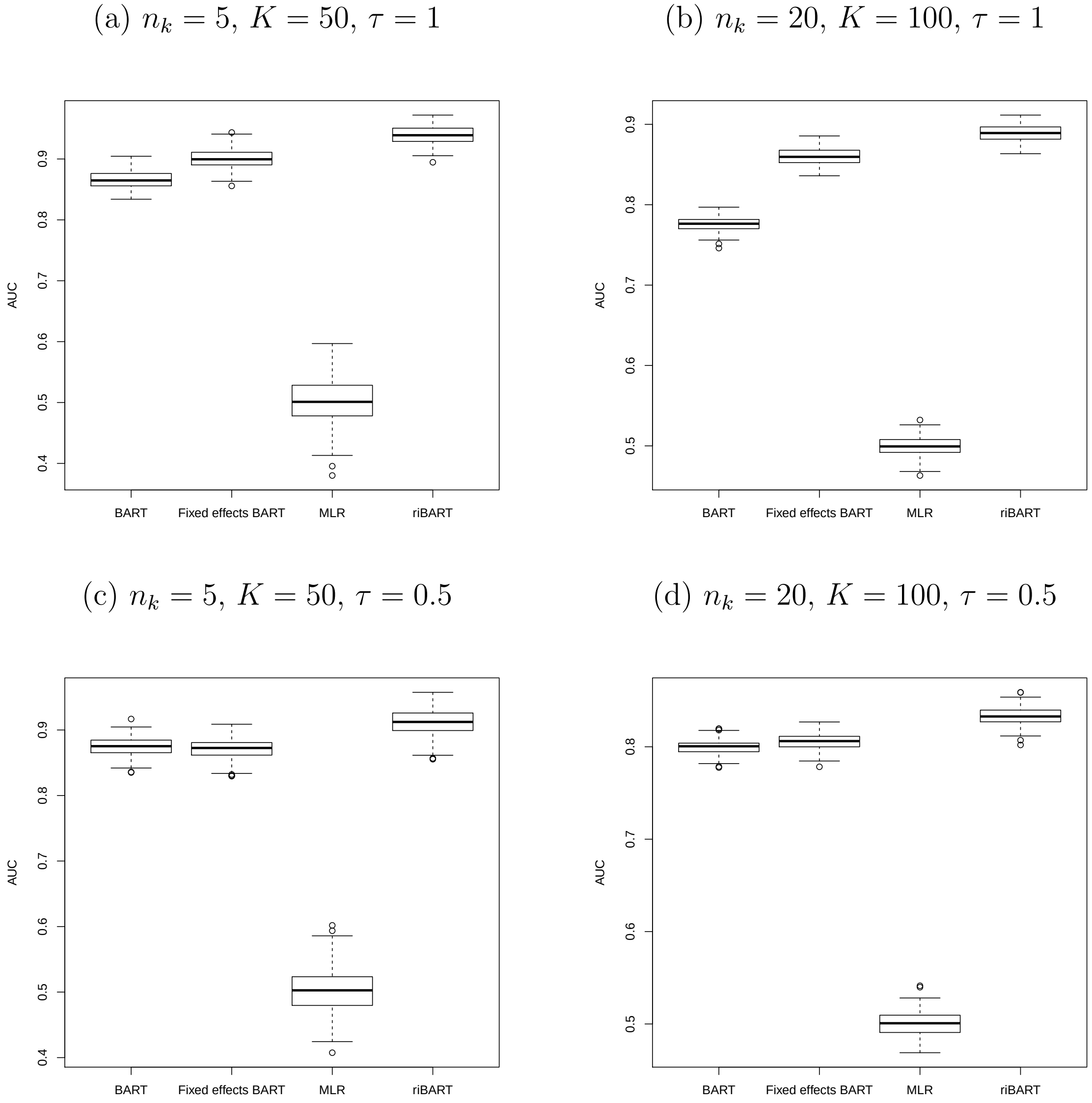
(a) $n_k = 5$, $K = 50$, $\tau = 1$, $\sigma = 1$

(b) $n_k = 20$, $K = 100$, $\tau = 1$, $\sigma = 1$

(c) $n_k = 5$, $K = 50$, $\tau = 0.5$, $\sigma = 1$

(d) $n_k = 20$, $K = 100$, $\tau = 0.5$, $\sigma = 1$

Figure 3.2: Boxplots of area under the receiver operating characteristic curve (AUC) for binary correlated outcomes produced by BART, Fixed effects BART, MLR, and riBART.



(a) $n_k = 5$, $K = 50$, $\tau = 1$

(b) $n_k = 20$, $K = 100$, $\tau = 1$

(c) $n_k = 5$, $K = 50$, $\tau = 0.5$

(d) $n_k = 20$, $K = 100$, $\tau = 0.5$

Table 3.1: Simulation results for continuous correlated outcomes. Bias and coverage of $\sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j) + a_k$ $(g(x) + a_k)$ and σ for BART, riBART, fixed effects BART, and multiple linear regression (MLR).

Scenario 1: continuous, $n_k = 5$, $K = 50$, $\tau = 1$, $\sigma = 1$								
	$g(x) + a_k$				σ			
	Bias	RMSE	Coverage (%)	AIL*	Bias	RMSE	Coverage (%)	AIL
BART	< 0.01	0.06	95.05	3.40	0.04	0.14	92.00	0.51
riBART	< 0.01	0.06	95.44	3.22	-0.04	0.07	99.50	0.41
Fixed effects BART	< 0.01	0.06	94.68	3.18	0.11	0.15	83.00	0.42
MLR	< 0.01	0.06	48.72	6.92	3.64	3.64	0.00	0.76
Scenario 2: continuous, $n_k = 20$, $K = 100$, $\tau = 1$, $\sigma = 1$								
	$g(x) + a_k$				σ			
	Bias	RMSE	Coverage (%)	AIL	Bias	RMSE	Coverage (%)	AIL
BART	< 0.01	0.02	82.72	2.50	0.32	0.33	0.00	0.10
riBART	< 0.01	0.02	92.77	1.81	-0.01	0.02	92.50	0.08
Fixed effects BART	< 0.01	0.02	89.57	1.78	0.06	0.06	34.50	0.11
MLR	< 0.01	0.02	45.74	6.42	3.69	3.70	0.00	0.27
Scenario 3: continuous, $n_k = 5$, $K = 50$, $\tau = 0.5$, $\sigma = 1$								
	$g(x) + a_k$				σ			
	Bias	RMSE	Coverage (%)	AIL	Bias	RMSE	Coverage (%)	AIL
BART	< 0.01	0.06	89.22	2.64	-0.24	0.25	37.50	0.41
riBART	< 0.01	0.06	94.80	3.05	-0.09	0.10	96.00	0.37
Fixed effects BART	< 0.01	0.06	94.66	3.09	0.07	0.12	90.00	0.40
MLR	< 0.01	0.06	49.32	6.91	3.56	3.56	0.00	0.74
Scenario 4: continuous, $n_k = 20$, $K = 100$, $\tau = 0.5$, $\sigma = 1$								
	$g(x) + a_k$				σ			
	Bias	RMSE	Coverage (%)	AIL	Bias	RMSE	Coverage (%)	AIL
BART	< 0.01	0.02	91.02	2.04	0.05	0.05	36.00	0.08
riBART	< 0.01	0.02	92.69	1.78	-0.01	0.02	91.00	0.08
Fixed effects BART	< 0.01	0.02	90.03	1.76	0.05	0.05	45.50	0.11
MLR	< 0.01	0.02	46.26	6.42	3.61	3.62	0.00	0.27

*AIL = Average interval length.

Table 3.2: Simulation results for binary correlated outcomes. Bias and coverage of $\sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j)+a_k$ ($g(x)+a_k$) for BART, riBART, fixed effects BART, and multiple linear logistic regression (MLLR).

Scenario 5: binary, $n_k = 5$, $K = 50$, $\tau = 1$				
	$g(x) + a_k$			
	Bias	RMSE	Coverage (%)	AIL*
BART	0.02	0.09	73.01	2.12
riBART	0.01	0.10	93.31	2.61
Fixed effects BART	0.03	0.09	62.77	1.61
MLLR	< 0.01	0.11	43.13	1.37
Scenario 6: binary, $n_k = 20$, $K = 100$, $\tau = 1$				
	$g(x) + a_k$			
	Bias	RMSE	Coverage (%)	AIL
BART	0.02	0.04	52.35	1.40
riBART	< 0.01	0.03	94.56	1.62
Fixed effects BART	0.02	0.04	53.60	1.08
MLLR	-0.01	0.04	32.54	1.01
Scenario 7: binary, $n_k = 5$, $K = 50$, $\tau = 0.5$				
	$g(x) + a_k$			
	Bias	RMSE	Coverage (%)	AIL
BART	< 0.01	0.08	92.51	2.13
riBART	< 0.01	0.08	95.32	2.22
Fixed effects BART	0.01	0.08	84.27	1.63
MLLR	-0.02	0.11	62.14	1.53
Scenario 8: binary, $n_k = 20$, $K = 100$, $\tau = 0.5$				
	$g(x) + a_k$			
	Bias	RMSE	Coverage (%)	AIL
BART	< 0.01	0.03	80.72	1.42
riBART	< 0.01	0.03	94.81	1.40
Fixed effects BART	0.01	0.03	78.53	1.05
MLLR	-0.02	0.05	51.40	1.18

*AIL = Average interval length.

3.5 Predicting Driver Stop before Left Turn Execution

Given the success of riBART in our simulation scenarios, especially for possibly correlated binary outcomes, we now turn to investigate whether this superior performance produced by riBART would propagate to our dataset.

3.5.1 Integrated Vehicle-Based Safety Systems (IVBSS) Study

The dataset we used to develop our prediction model was obtained from the Integrated Vehicle Based Safety System (IVBSS) study conducted by *Sayer et al.* (2011). This study collected naturalistic driving data from 108 licensed drivers in Michigan between April 2009 and April 2010. In the study, 16 late-model Honda Accords were fitted with cameras, recording devices, and several integrated collision warning systems. Each driver used a vehicle for a total of 40 days – 12 days baseline period with IVBSS switched off followed by 28 days with IVBSS activated. Since our objective was to develop a prediction model for human driving behavior, we used the 12 days baseline unsupervised driving data. In total, the 107 drivers made 1,822 left turns (One driver removed because he or she only made one left turn). Each driver took on average of 35 turns, with a range of 8 to 139 turns per driver. This suggests that riBART could potentially improve the prediction performance of our model compared to BART, while simultaneously producing an estimate of a driver's tendency to stop before executing a left turn.

3.5.2 Data preparation

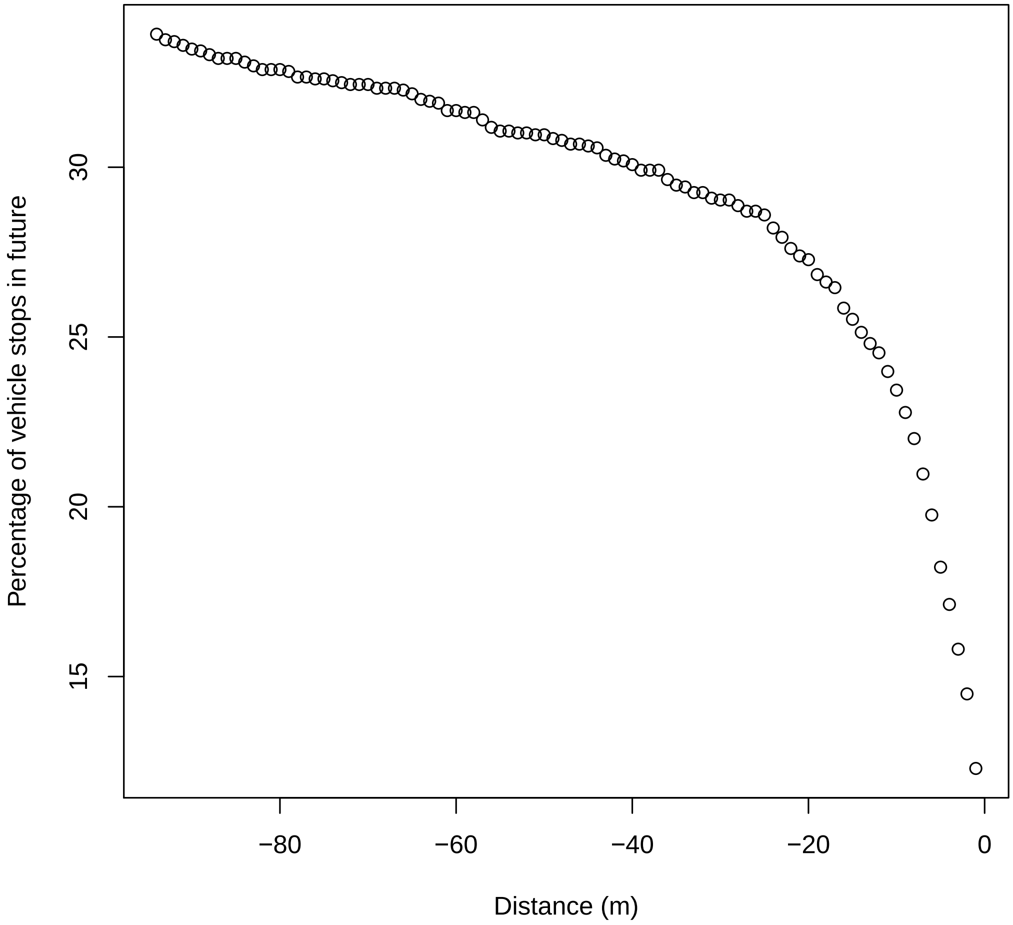
A detailed description of how we determined and prepared our dataset for analysis using riBART can be found in the Appendix C. We provide a brief description in the following paragraphs to aid discussion.

We begin by extracting both the speed of the vehicle (in m/s) and the distance traveled (in m) at 10 millisecond intervals starting from 100 meters away from the center of an intersection. To obtain a practical prediction model, we converted the time series of vehicle speeds to a distance series to provide a distance-varying definition for our binary outcomes of whether a vehicle would stop before executing a left turn in the future. Our outcome was whether a vehicle would eventually stop before executing a left turn, estimated repeatedly at 1 meter intervals before the intersection. We defined $Y_{ikd} = 1$ for the vehicle that would stop eventually before executing a left turn where d is the d^{th} meter from the center of an intersection and i indexes the turns for driver k, $i = 1, \ldots, n_k$. For the vehicles that would not stop before executing a left turn, we defined them as $Y_{ikd} = 0$. For example, if the vehicle's current location is -45 meters, the outcome is whether the vehicle will stop between -44 and -1 meter. If a vehicle stops and restarts, the outcome is reset: a vehicle that stops at -40 meters and then proceeds through the intersection will have an outcome of 1 (stopping) from -94 to -40 meters, and 0 (not stopping) from -39 to -1 meters.

Figure 3.3 shows the resulting profile of proportion of stops from -100 meters to the center of the intersection (0 meters). We can see that majority (about 65%) of the left turns did not stop before executing a left turn. At -100m, about 35% of the vehicles would stop before executing a left turn. As vehicles approach the center of an intersection, the proportion of vehicles that eventually stop decreases gradually until about -25m. Beyond -25m, there was a quick drop in the proportion of vehicles that stop suggesting that most vehicles 'decide' to stop about 25m away from the center of an intersection.

At any given distance, we could use the full profile of a vehicle's past speeds as the predictors, but these speeds may contain irrelevant information. Thus, we employed Principal Components Analysis (PCA) to summarize the distance series of vehicle speeds. A detailed description of our decision to use PCA can be found in *Tan et al.* (2017). In brief, we found that the principal components (PCs) of vehicle speed provided us with much more information than just dimension reduction. The first

Figure 3.3: Proportion of vehicles in our study that would be stopped ($\leq$ 1m/s) at some future point for each meter away from the center of an intersection.



three PC loadings were fairly similar meter by meter as the vehicle approaches the center of an intersection. In addition, these PCs seemed fairly interpretable as first, second, and third derivatives of the vehicle's location relative to the center of the intersection. The first PC could be loosely interpreted as average speed, second PC as acceleration, and third PC as jerk, change in acceleration. We only included the first two PCs as our predictors because the first two PC scores explained more than 99% of the variation in vehicle speed at all distances (See Figure 3.4). In addition, we found that adding PC scores beyond these did not produce a large improvement in prediction (See Figure 3.5).

To decide on our preliminary prediction method, we compared the AUC performance of the following models: logistic regression with polynomial transformation on the predictors, logistic regression with splines for the predictors, BART, and SuperLearner *van der Laan and Polley* (2010) with elastic net *Friedman et al.* (2010), logistic regression, K-Nearest Neighbor, generalized additive models *Hastie and Tibshirani* (1990), mean of the outcomes, and BART as the ensemble learners (results not shown here). BART easily outperformed all of the approaches with respect to AUC except the SuperLearner. For the SuperLearner, it sometimes somewhat outperformed BART at a far distance from the intersection but as the vehicle approaches the intersection, SuperLearner stabilized at or a little below BART. Given the unstable AUC performance of the SuperLearner, we focused our attention on extending BART to account for the clustering in our dataset.

Incorporating information from further distances into the estimation of the PCs might also introduce noise to our two PC predictors. Hence, we estimated 8 sets of the first and second PCs from the moving window of vehicle speeds with lengths 3 meters, 4 meters, . . ., 10 meters. We then computed the 10-fold cross validation AUC profile produced by each set with the first and second PCs as the predictor and BART as the model. We finally compared these 8 different AUC profiles and found that a

Figure 3.4: Principal Component loadings for the first and second PC from -95m to -90m, -70m to -65m, -45m to -40m, and -20m to -15m (left to right). The percentages indicate the proportion of variation explained by each PC.

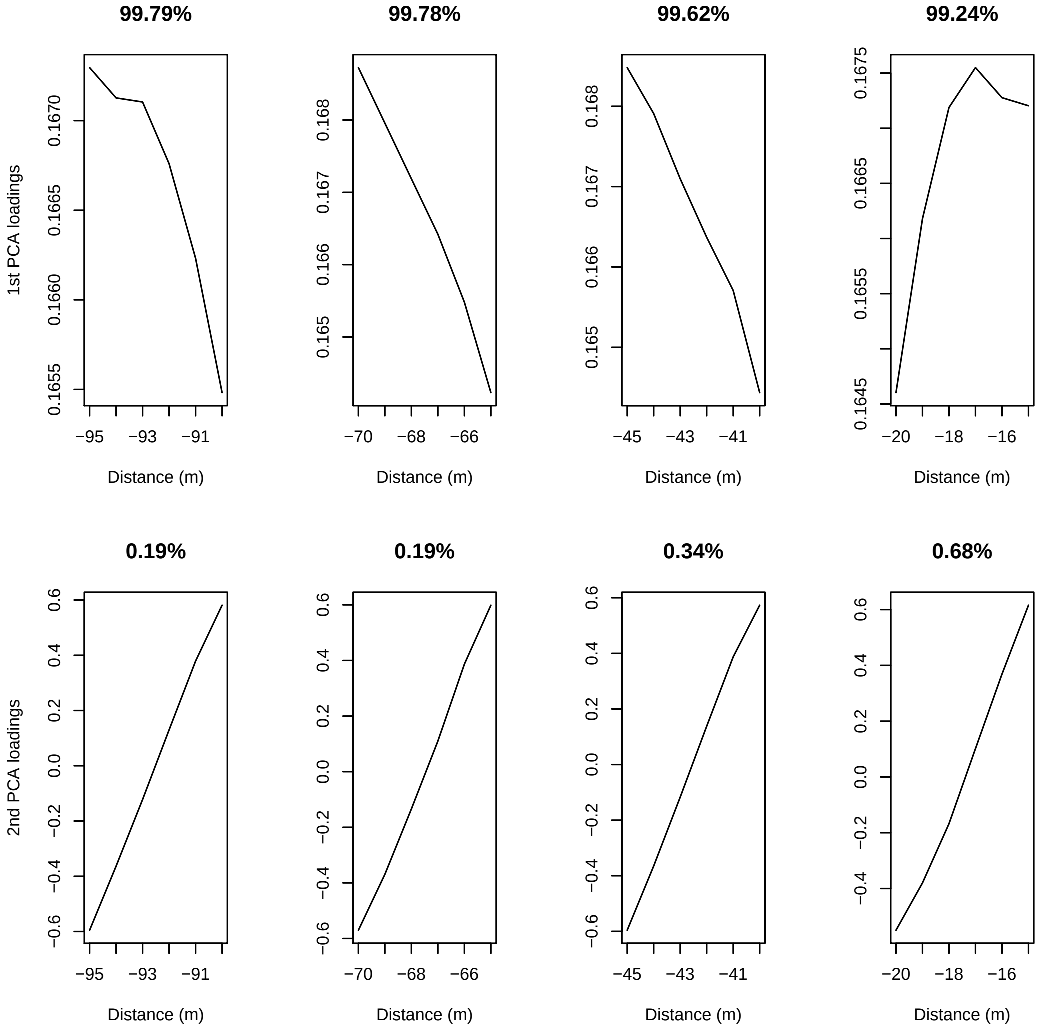


Figure 3.5: Comparing the Area Under the receiver operating characteristic Curve (AUC) profile gains of including each Principal Component (PC) in the logistic regression model.

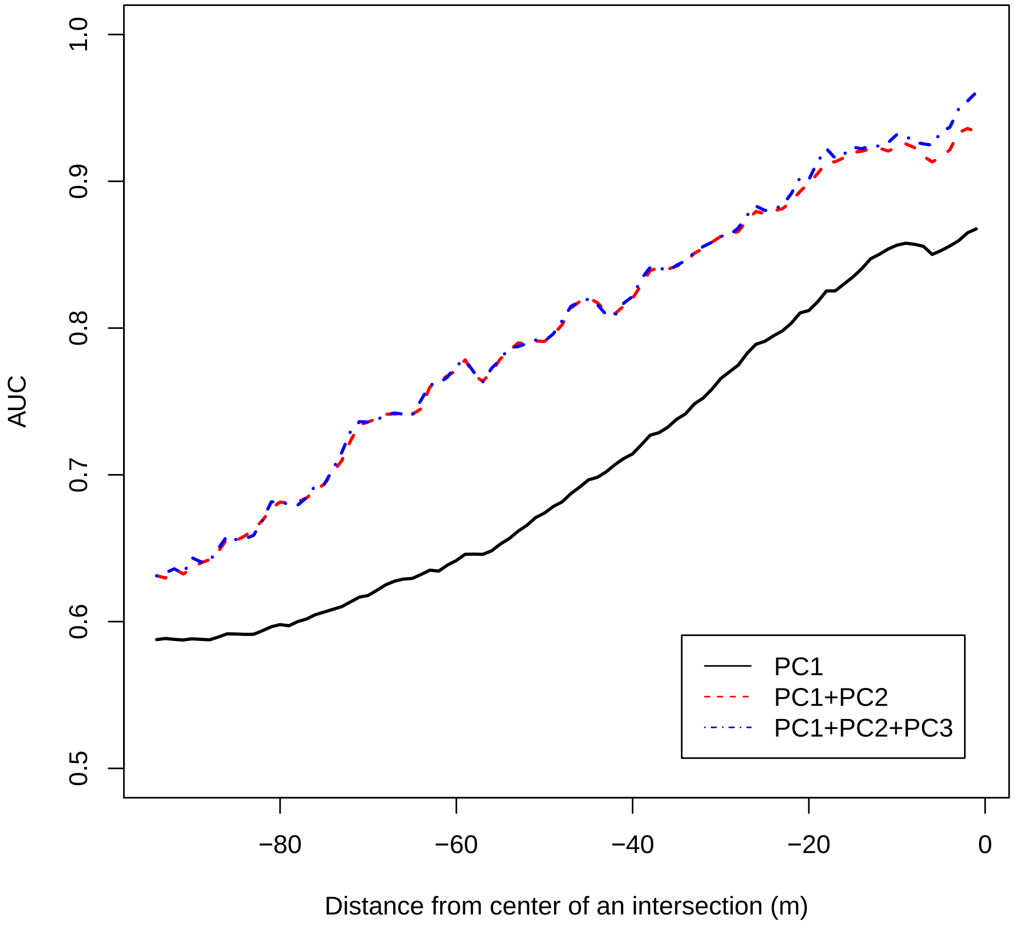


Table 3.3: Example of resulting matrix for our IVBSS study dataset.

d	k	i	X_{ikd1}	X_{ikd2}	X_{ikd3}
1	1	1	x	x	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	n_1	x	x	x
1	2	1	x	x	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	2	n_2	x	x	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	107	n_{107}	x	x	x
2	1	1	x	x	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	1	n_1	x	x	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
94	107	n_{107}	x	x	x

window length of 6 meters gave us the best balance between AUC value and window length. The result of this comparison can be found in Figure 6 of (*Tan et al.*, 2017).

Finally, we included a categorical predictor, the number of times the vehicle has stopped up to the current location, to adjust for the likely correlation within each turn. The categories for this predictor were: for -94m to -64m, 0 or ≥ 1; for -63m to -20m, 0, 1, or ≥ 2; and for -19m to -1m, 0, 1, 2, or ≥ 3. Table 3.3 illustrates the resulting data matrix before analysis.

3.5.3 Analysis

We fit riBART with a random effect at the driver level which incorporates within-driver correlation to our dataset. Because we fit riBART meter-by-meter, a slight clarification in notation of the riBART is needed. We model $P(Y_{ikd} = 1|\mathbf{X}_{ikd})$ as

$$P(Y_{ikd} = 1|\mathbf{X}_{ikd}) = \Phi[G(\mathbf{X}_{ikd})],$$

where $\mathbf{X}_{ikd} = (X_{ikd1}, X_{ikd2}, X_{ikd3})^T$, $k = 1, \ldots, K$ indexes the drivers, i indexes the turns for driver k, $i = 1, \ldots, n_k$, and $d = -94, \ldots, -1$ indexes the distance from the center of an intersection. The riBART model is then

$$G(\mathbf{X}_{ikd}) = \sum_{j=1}^{m} g(\mathbf{X}_{ikd}, T_{jd}, \mathbf{M}_{jd}) + a_{kd}, \tag{3.10}$$

where $a_{kd} \sim N(0, \tau_d^2)$. Note that we are estimating each model at distance d separately and assuming that there is a different random intercept for each driver at each d.

For comparison, we also ran BART, which ignores within-driver correlation; fixed effects BART, which ignores within-driver correlation but adjusts for the driver effect in the model; a random intercept linear logistic regression (riLogistic), which incorporates within-driver correlation but ignores non-linearity and complex interactions; and MLLR, which ignores within-driver correlation, non-linearity, and complex interactions. It may have been more straight forward to use polynomial or splines of our first two PCs together with a random intercept to obtain a model that handles non-linearity and driver correlations. Unfortunately, even simple models with a quadratic main effect or a single knot spline at the mean or median produced convergence errors for the random intercept GLM model. Hence, we did not include them as competitors against riBART. We obtained the linear logistic regression using the *glm* function in *R* while the random intercept linear logistic regressions were obtained using the *glmer* function from the *R* package *lme4*. We compared the in-sample AUC of the six methods and computed the 95% CI of the AUCs using the method of *Hanley and McNeil* (1982), which uses a linear approximation of the AUC to the Somer's D statistic to obtain an estimate of the variance of AUC. In addition, we investigated the proportion of depth of the 200 regression trees over 5,000 iterations for each meter as well as the marginal effects of each main effects and interaction to explore the additional features provided by riBART.

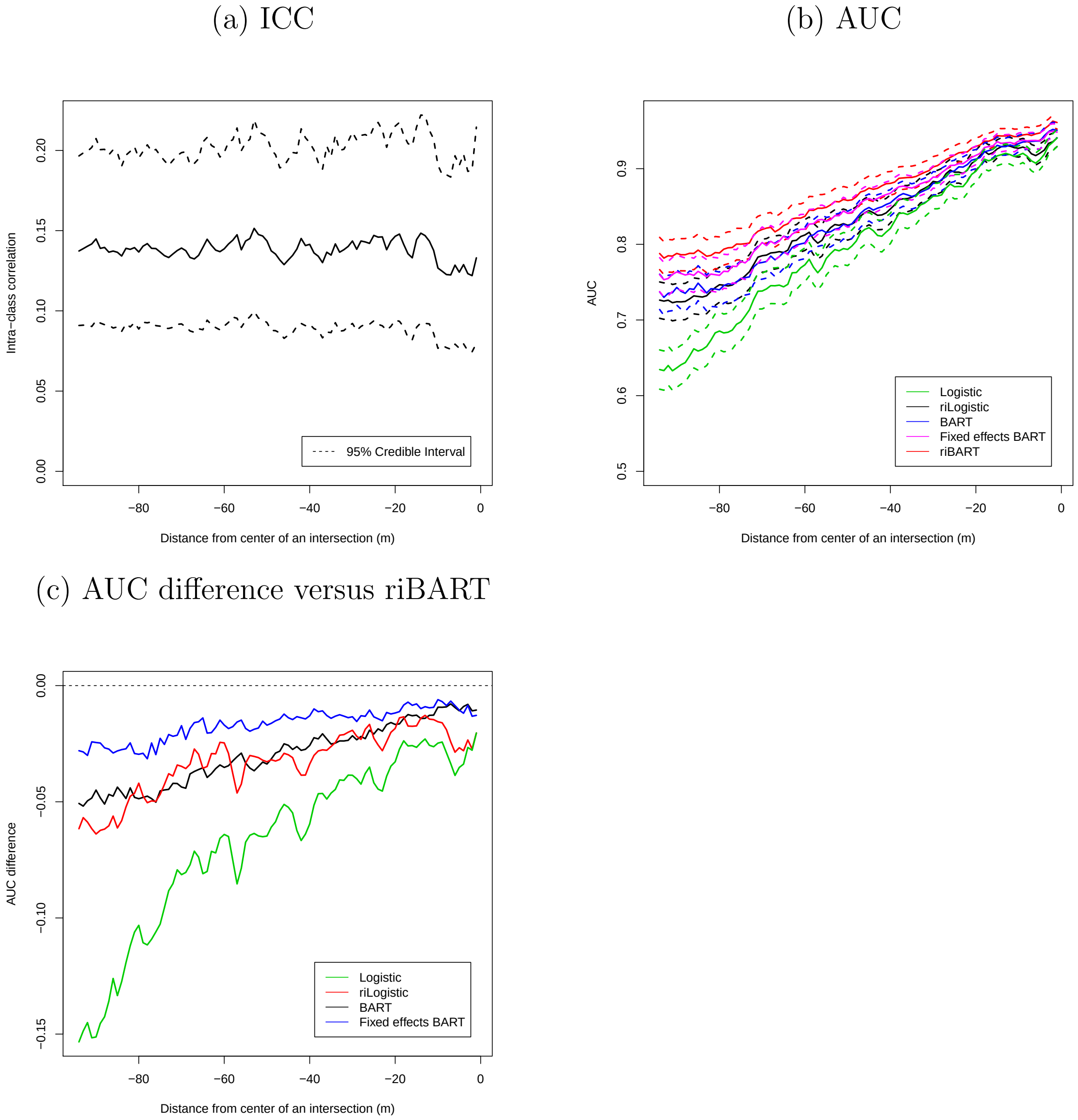
3.5.4 Results

Figure 3.6 shows (a) the the estimated intra-class correlation (ICC, $\frac{\tau^2}{\tau^2+1}$) profile; (b) the AUC profiles of riBART, BART, fixed effects BART, riLogistic, and MLLR; and (c) the AUC profile difference between riBART versus BART, riBART versus fixed effects BART, riBART versus riLogistic, and riBART versus MLLR.

The posterior mean profile of ICC was small, between about 0.12 and 0.15, and fairly stable as the vehicle approaches the center of an intersection. This suggests firstly that the variance parameter, τ, for the random intercept, a_k, is small for left turn stops and secondly that as the vehicle approaches the center of the intersection, the effect of individual 'habits' of the driver remained relatively stable throughout the left turn maneuver. For the AUC profile, we see evidence that riBART performed better than BART, fixed effects BART, riLogistic, and MLLR. The difference in AUC profile between riBART versus BART, riBART versus fixed effects BART, riBART versus riLogistic, and riBART versus MLLR remained negative throughout the left turn maneuver suggesting the superior prediction performance of riBART to the other prediction methods we considered.

At 94m away from the center of intersection, riBART produced an AUC estimate of 0.79 [95% C.I. (0.77, 0.81)]. Comparatively, fixed effects BART produced an AUC of 0.76 (0.74, 0.78), BART produced an AUC of 0.74 (0.71, 0.76), riLogistic produced an AUC of 0.73 (0.70, 0.75), and MLLR produced an AUC of 0.64 (0.61, 0.66). In situations where last-second decisions are needed for example, Automatic Emergency Braking, an AUC of 0.79 would not be enough. However, the application that we envision for our algorithm is to provide further information to an oncoming driverless vehicle and help it make better decisions in conjunction with its own sensor-based algorithms. As such, almost any AUC value greater than 0.50 should improve the decision made by the driverless vehicle. Most likely, a driverless vehicle would use this information to adjust its own speed (up or down) so that any potential conflict

Figure 3.6: (a) The intra-class correlation (ICC) profile of riBART as a factor of distance from the intersection; (b) Area under the receiver operating characteristic curve (AUC) profile of riBART, BART, and random intercept logistic regression (dotted lines are 95% Credible Interval); and (c) AUC difference profile between riBART versus BART and riBART versus random intercept linear logistic regression.

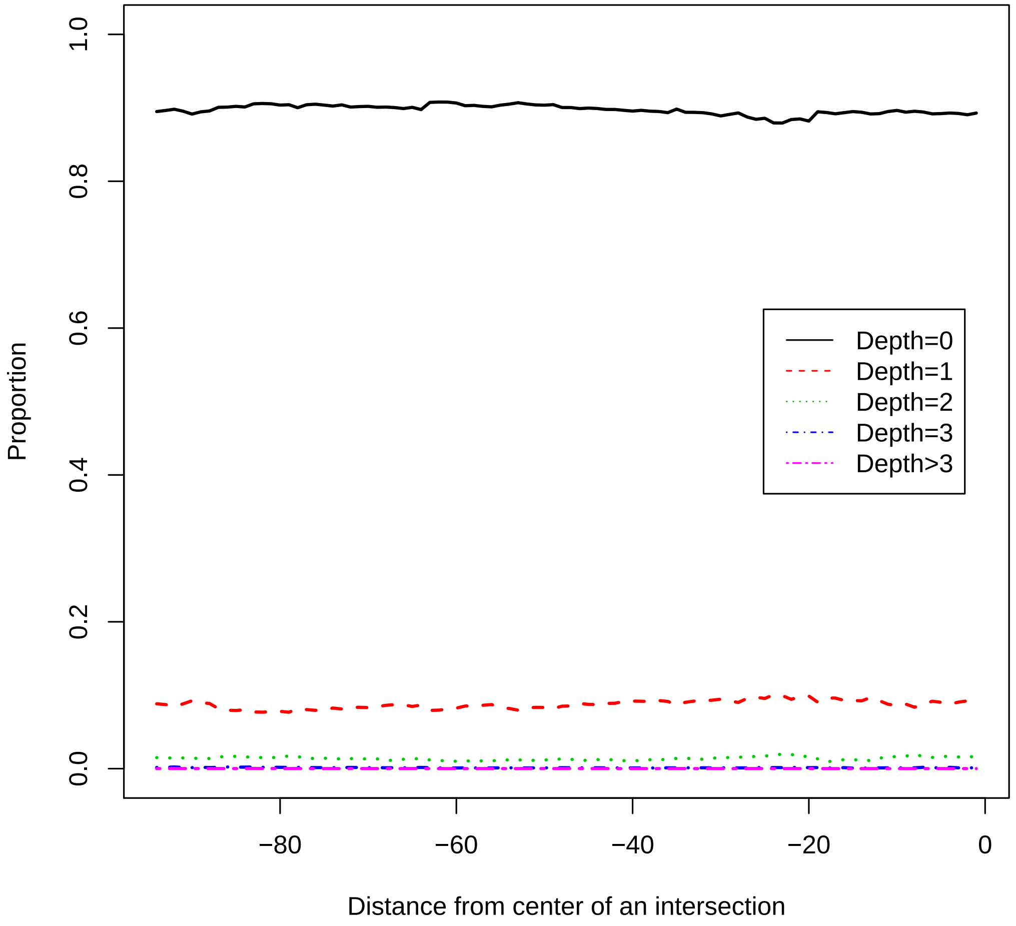


(a) ICC

(b) AUC

(c) AUC difference versus riBART

Figure 3.7: Proportion of depth of regression tree meter by meter.



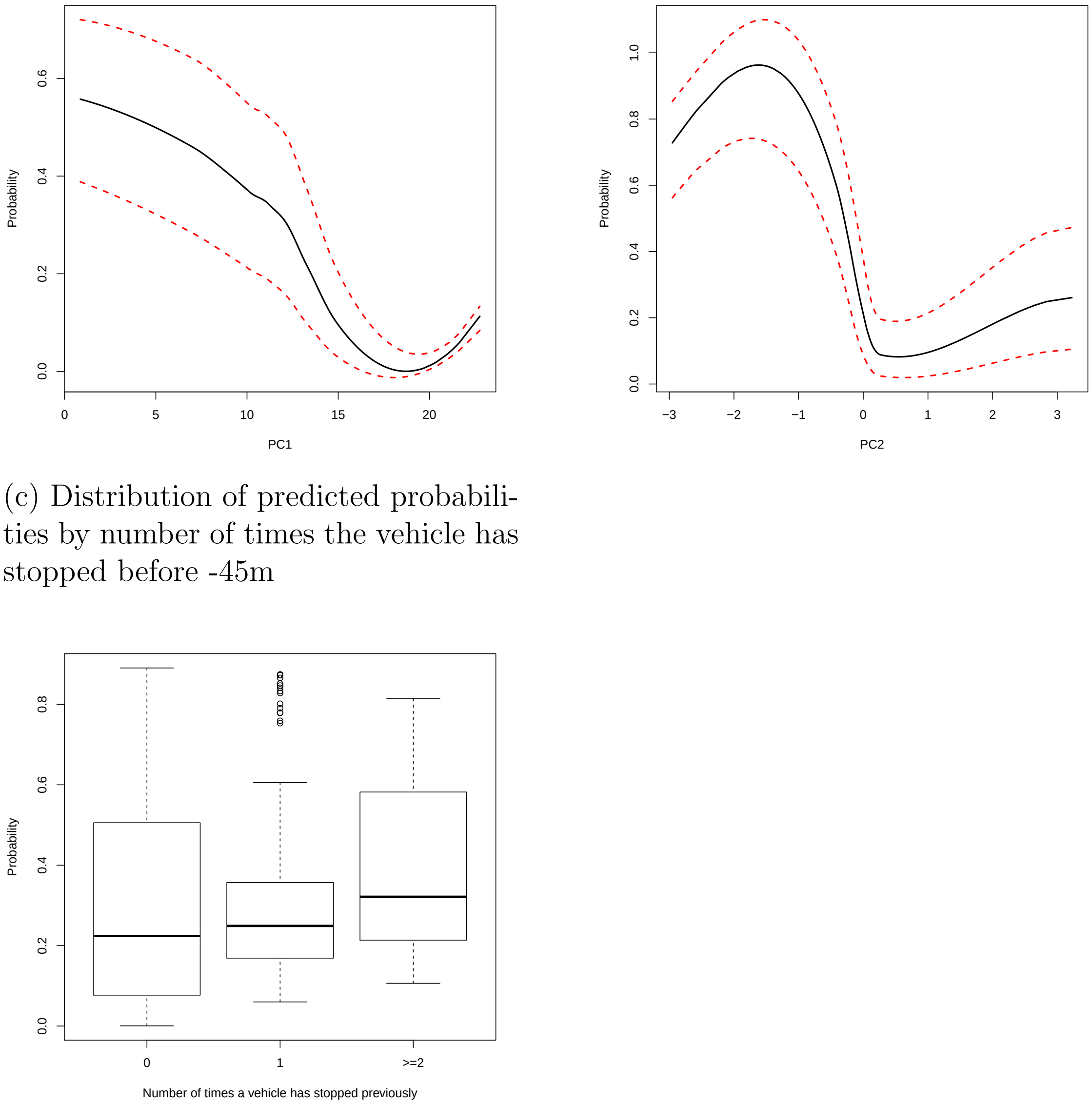
between it and the human-driven turning vehicle is less ambiguous (e.g., speeding up to pass before the turning vehicle would turn or slowing down to let the turning vehicle go).

Figure 3.7 shows the proportion of depth of each regression tree meter by meter from -94m away from the center of an intersection to -1m away from the center of an intersection. About 90% of the regression trees employed by riBART were single terminal nodes for every meter, 9% were trees with one internal node with two child terminal nodes, and the rest, about 1%, had regression tree depths of more than 1. This suggests a rather strong penalization effect for the tree structure depth which was what the BART portion of riBART was aiming for. We also investigated the frequency of each main and interaction effect being used by each regression tree to give us a sense of which main or interaction effect was most used, hence an indication of effect importance (results not shown here). We found that the main effects were most frequently used (excluding single terminal node trees) followed by the two-way interactions and lastly the three-way interaction. These results suggest that the two most important variables could be the first two PCs.

Figure 3.8 shows the smoothed marginal effect plots of all the main effects at -45m (approximately halfway through the left turn). The clear non-linearity of the main effects and the reduced use of the interactions by riBART suggests that the substantial improvement provided by riBART over random intercept linear logistic regression came from the non-linear effects. Since PC1 can be loosely interpreted as the average speed, plot (a) suggest that at -45m, a higher average speed suggests a lower probability of stopping with a sharp decline in the probability when the average speed increases to around 12-13 m/s. As the average speed increases to about 17-18 m/s, the probability of stopping increases again. Smoothed marginal effect plots for PC1 from -94m to -1m can be found in the supplementary materials available online.

For PC2, since it could be loosely defined as the acceleration of the vehicle, plot (b) suggests that negative acceleration produces a higher probability of stopping while positive acceleration produces a lower probability of stopping halfway through the left turn maneuver. This result continues as the vehicle approaches the center of an intersection. The smoothed marginal effect plots for PC2 from -94m to -1m can be found in the supplementary materials available online. Note that for PC2, the PC loadings sometimes suggest deceleration instead of acceleration i.e. the slope for PC2 in Figure 3.4 is negative instead of positive. We have placed a condition (multiplying the loadings by -1 whenever this occurs) in our implementation to ensure that the heuristic interpretation of PC2 will always stay as acceleration.

Figure 3.8: Smoothed (a) marginal effect of PC1 (b) marginal effect of PC2; and (c) boxplots of the predicted probability of stopping stratified by the number of times a vehicle has stopped previously. Dotted red lines show smoothed 95% credible interval.


(a) PC1

(b) PC2

(c) Distribution of predicted probabilities by number of times the vehicle has stopped before -45m

Plot (c) shows the boxplot of the predicted probability of stopping stratified by the number of times a vehicle has stopped previously before -45m. From the stratified boxplots, we can see that as the number of times the vehicle has stopped previously increases, the vehicle is slightly more likely to be predicted to stop before executing a left turn.

In summary, Figure 3.8 suggests that vehicles with lower average speed, and/or slowing down quickly, and/or have stopped multiple times previously would be more likely to stop compared to vehicles with higher average speed, accelerating, and has not made a previous stop. This agrees with our understanding of how a vehicle would stop at an intersection before executing a left turn and suggests that riBART is producing sensible results.

3.6 Discussion

In this paper, we developed a model, riBART, to help engineers developing self driving vehicles predict whether a human-driven vehicle would stop at an intersection before executing a left turn. We achieved this by utilizing the model that did well in our preliminary analysis, BART, and extending it to account for the key feature in our dataset, clustered observations. Although existing methods extending BART to longitudinal datasets were available, our approach was more straight-forward and can be implemented on correlated binary outcomes. We have also provided codes that would implement riBART in our supplementary materials available online. Our codes could be used to explore some of the properties and features that riBART provided over the random intercept linear logistic regression. These results could help the researcher make sense of the marginal effects provided by each variable estimated using riBART.

Applying riBART to our dataset, substantial improvement in prediction compared to BART can be obtained when we take into account that different drivers have different ‘propensities to stop’ before executing a left turn at an intersection; that is, the inclusion of a random intercept improves prediction performance for our dataset compared to a model without a random intercept. This implies that future development of an operational algorithm should try to accommodate the similarities of stopping behavior for a given human driver through a learning algorithm. For example, devices that are able to transmit information about a driver’s propensity to stop could be installed on vehicles to improve the decision-making performance of the self driving vehicle.

To elaborate, we are assuming that this method would be used to create a prediction profile that would be broadcast to autonomous vehicles, thus utilizing all of the available information on the turning behavior both across and within vehicles. For a new vehicle to this system, we could treat the posterior means of the random intercepts in our dataset as a “quasi” distribution for the random intercept of the unseen driver. Alternatively, we could draw an initial random intercept distribution using the posterior distribution of the random intercept variance parameter. Once this driver makes a turn, their random intercept can be estimated and updated.

In our simulation study, we found that the 95% coverage for σ was reduced when the number of clusters and the number of observations within a cluster was large ($n_k = 20$, $K = 100$). The likely cause for the poor coverage is due to low variation in the posterior draw of σ resulting in reduced average 95% credible interval length. We believe this low variation in σ is due to the regression trees in BART getting stuck at certain tree structures. This phenomenon of regression trees getting stuck at certain tree structures has been discussed by *Pratola* (2016) previously. The difference here is that *Pratola* (2016) only reported observing regression trees being stuck when the true σ is small for regression trees. We argue that regression trees might also get stuck when the effective sample size, N, is large. This is because with a large N, deeper trees tend to produce a better fit for R_{kj} in Eq. (3.3). However, when a regression tree gets deep, the standard grow, prune, change, and swap steps will have trouble proposing new trees with radically different tree structures. This lack of radically different tree structures implies reduced variability in the tree structures, which is indirectly reflected by the lack of variation in σ.

This issue is separate from the development of BART in the correlated data context, and indeed would occur even when observations are independent. We illustrate this with an example using BART implemented via the *BayesTree* package in *R*. We generated $Y_k = 10\sin(\pi X_{k1}X_{k2}) + 20(X_{k3} - 0.5)^2 + 10X_{k4} + 5X_{k5} + \epsilon_k$ with $X_{kq} \overset{i.i.d.}{\sim} \text{Uniform}(0,1)$, $q = 1, \ldots, 5$ and $\epsilon_{ik} \overset{i.i.d.}{\sim} N(0,1)$. We then ran 200 simulations with $\sigma = 1$ and a sample size of 2,000. The resulting bias, RMSE, 95% coverage, and AIL for σ were -0.04, 0.04, 79%, and 0.09 respectively. We observe once again that although bias and RMSE were small, the 95% coverage for σ was far from nominal because the AIL was small. We think that this issue of a lack in variation of σ when the sample size is large could be solved by either increasing the number of regression trees used, re-calibrating the α and β parameters used to penalize each regression tree, or to include the rotate step proposed by *Pratola* (2016) in the proposal of a new regression tree in the MH algorithm of BART. As inference about σ is not the key focus of this paper, we leave investigation of this problem with BART to future work.

Although our analysis of left turn data found that the first two PCs appeared to be the most important predictors based on the frequency of the trees drawn, caution should be exercised when using riBART to decide whether a variable was important. This is because of the default discrete uniform prior we placed on the variables which forces the model to use the variables uniformly for prediction. If variable selection is desired, spike and slab priors could be considered but such an implementation would go beyond the scope of this work.

Our proposed model only included a random intercept but, there may be situations where the researcher believes that there may be more complicated linear random effect mechanisms occurring. In our application, estimating a “turn-level” random effect nested within the driver-level random effect is possible. Eq. (3.10) could be modified to become

$$G(\mathbf{X}_{ikd}) = \sum_{j=1}^{m} g(\mathbf{X}_{ikd}, T_{jd}, \mathbf{M}_{jd}) + a_{kd} + l_{ik},$$

where $a_{kd} \sim N(0, \tau_d^2)$, $l_{ik} \sim N(0, \tau^2)$, and $a_{kd} \perp l_{ik}$. To estimate this model, we would employ once again a Gibbs-sampling type method by drawing τ or l_{ik} conditional on the rest of the parameters and the observed data. By estimating τ and comparing it with τ_d, we could determine if we require additional variables to account for the dependencies in our outcome. This is because if τ was much larger compared to τ_d, this suggests that not all of the variation is captured by the driver level random intercept and there is still some variation left at the turn level. However, such a model is not practical for our prediction situation. This is because the estimated turn-level effect would only be useful for prediction for that turn – but once that turn is completed, we have no interest in predicting it. Other plausible areas for future research include extending BART and riBART to outcomes of other forms, for example, ordinal outcomes or counts.

CHAPTER IV

“Robust-squared” Imputation Models Using BART

4.1 Introduction

Missing data are common in many surveys and experiments. Data may be missing because of the subject’s refusal to provide information or survey drop-out, or by the design of the experiment or survey. If the amount of missing data is large, or if the missing data differ from the observed data and would change our conclusions if we had observed it, failure to account for missing data during analysis leads to biased parameter estimation and misleading conclusions. Missing data in surveys, including major US transportation safety-related surveys, is very common. The National Automotive Sampling System – Crashworthiness Data System (NASS-CDS) is representative of all police-reported towaway crashes in the US. A key measure of crash severity is the “instantaneous” change in velocity, delta-v. Because estimation of delta-v requires a careful crash investigation that is not always possible, it is commonly missing. Similarly, the Fatality Analysis Reporting System (FARS) releases information annually from all fatal motor vehicle crashes that occur on US public roads. Here, blood alcohol concentration (BAC) levels are often missing because subjects were not tested at the crash site.

Determining a dataset's missingness mechanism is the first step in handling missing data. There are three categories of missingness mechanism: missing completely at random (MCAR), where the data are missing by chance and are not related to observed or unobserved variables; missing at random (MAR), where the data are missing depending on some variables which are fully observed; and not missing at random (NMAR), where the data are missing depending on the variable that contains the missing value. In our examples, delta-v is often missing in vehicles that have either quite limited damage (so that the vehicle may be been driven off and not available for followup) or very severe damage (so that the algorithms used to estimate it do not have reliable inputs); while this might seem to imply NMAR, there are a number of observed measures such as towaway status, injury severity, and speed limit to make the MAR assumption more plausible. Similarly, BAC measures are often missing in subjects that did not appear to be intoxicated; again factors such as gender, age, time of day, and crash severity can strengthen what, without other covariates, would seem to imply an NMAR mechanism. Since MAR assumptions do not typically need to rely on unobservable parameters and can be reasonable given sufficient fully observed covariates, it is a common assumption that researchers adopt and shall be the focus of this paper.

Common methods to handle missing data under MAR is to impute the missing values via mean imputation, regression imputation, or hot deck (*Little and Rubin*, 2002, Chapter 4). Once the missing values are imputed, standard statistical techniques can be employed as though there were no missingness in the dataset. To obtain valid inferences, multiple imputation (MI) can then used to account for the imputation uncertainty. MI first generates D imputed datasets. Then, the within and between variability of the estimator are calculated and combined to give the total uncertainty of the imputed estimator (See *Little and Rubin*, 2002, Chapter 5). MI usually rely on modeling assumptions that might be incorrect or difficult to test.

Robins et al. (1994) proposed a robust method, the augmented inverse probability estimator (AIPWT), which separately models the response propensity and mean of the outcome as a function of observed data, and yields a consistent estimator if either model is specified correctly. Another robust method is the penalized splines of propensity prediction (PSPP; *Zhang and Little*, 2009), which is based on a Bayesian prediction framework different relative to the method proposed by *Robins et al.* (1994), but having the same property that either a correct specification of response propensity or mean model produces consistent estimates. These methods are usually called doubly robust (DR) estimators. Extensions to multiply robust estimators that allow multiple models to be specified and yield consistent estimates as long as at least one is correct have been developed as well (*Han and Wang*, 2013).

Unfortunately, DR estimators may not work that well in situations where both the propensity and mean models are misspecified. *Kang and Schafer* (2007) showed this using a simulation example where both the propensity and mean model were moderately misspecified. AIPWT and PSPP did worse in terms of bias compared to a method that only used a mean model for imputation. For real-life datasets, of course, true models are almost never known. Thus, modifying the AIPWT and PSPP so that these methods are robust to misspecification of both the propensity and mean model becomes important for AIPWT and PSPP to remain relevant outside theoretical and simulation settings.

Current literature modifying DR estimators so that they become robust to misspecification mainly focus on two observations. First, the propensity model can produce large weights and hence cause severely biased estimates in DR estimators when both propensity and mean models are incorrectly specified. Second, the propensity and mean model are misspecified because of the non-linear main and multiple-way interaction effects.

For the former observation, *Kang and Schafer* (2007) proposed to replace the logistic regression with the robit regression (*Liu*, 2004) where the robit regression replaces the logistic link with the Student-t distribution. In *Cao et al.* (2009), they recognized that good performance of the propensity model in AIPWT relies on the summation of the multiplication of the propensity score and response being close to the sample size. Hence, they suggested estimating the logistic regression with the restriction that the summation of the multiplication of the propensity score and response is approximately equal to the sample size. More recently, *Imai and Ratkovic* (2014) proposed the covariate balancing propensity score (CBPS) where they focused on balancing the moments of the covariates between missing and non-missing groups instead of searching for a better parametric approach.

In this paper, we capitalize on the fact that the PSPP is already robust to the misspecification of the mean model, since it only requires that residuals of the misspecified mean model be a smooth function of the probability of non-response. Hence, a robust estimator of the response propensity will yield an estimator with especially strong robustness properties. Specifically, we estimate the propensity model using Bayesian additive regression trees (BART; *Chipman et al.*, 2010b). BART models the conditional mean of $\mathbf{Y}$ given X as a sum of regression trees. Use of regression trees allows automatic incorporation of multi-way interactions; non-linear main effects and multi-way interactions can be incorporated through the summation of these trees.

The use of BART as the imputation model is not entirely new. *Xu et al.* (2016) suggested using BART for situations where there is sequential missingness while *Kapelner and Bleich* (2015) suggested an approach to estimate regression trees if there are missingness in the predictors. The novelty of our work is the combination of the AIPWT or PSPP with BART to create a doubly-robust estimator where the degree of the misspecification for the estimation of the propensity model is greatly reduced: hence, our "robust-squared" terminology.

We organize the rest of our manuscript as follows. In Section 2, we describe our missing data problem followed by a brief review of the AIPWT, PSPP, and BART. We present our proposed methods for extending AIPWT and PSPP followed by suggesting two imputation methods using BART directly in Section 3. In Section 4, we employ a simulation study to compare our proposed methods against AIPWT and PSPP. In Section 5, we compared various imputation methods on the estimation of the population mean of delta-v and unadjusted odds ratio of injury severity using the 2014 National Automotive Sampling System Crashworthiness Data System (NASS-CDS) dataset as well as estimation of the population mean of Blood Alcohol Concentration (BAC) and proportion of subjects with BAC more than .010 and .100 using the 2015 Fatality Analysis Reporting System (FARS) dataset. Section 6 concludes with a discussion and possible future work.

4.2 Review of existing Doubly Robust methods for MAR data

4.2.1 Description and Notation

Suppose we have a continuous outcome Y_k, $k = 1, \ldots, n$ and we are interested in estimation and inference of $E[\mathbf{Y}] = \mu$, the population mean. Let $R_k = 1$ denote the k^{th} element of $\mathbf{Y}$ is observed and $R_k = 0$ denote the k^{th} element is missing. We restrict to situations where missingness of Y_k depends on p fully-observed covariates $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^T$.

4.2.2 Robbins, Rotnitzky, Zhao (1994) augmented inverse probability estimator (AIPWT)

To address the missing data problem described above, *Robins et al.* (1994) proposed a double robust estimator by solving a set of estimating equations. In brief, μ

is estimated as

$$\hat{\mu}_{AIPWT} = \frac{1}{n}\sum_{k=1}^{n}\{\frac{R_k Y_k}{Z_k} - \frac{R_k - Z_k}{Z_k}m(\mathbf{X}_k, \hat{\beta})\} \tag{4.1}$$

where $m(\mathbf{X}_k, \hat{\beta})$ is the conditional mean of Y_k and Z_k is the conditional propensity of response. Typically the conditional mean of Y_k is estimated by multiple linear regression (MLR)

$$m(\mathbf{X}_k, \hat{\beta}) = E[Y_k|\mathbf{X}_k; \hat{\beta}] = \hat{\beta}_0 + \hat{\beta}_1 X_{k1} + \ldots + \hat{\beta}_p X_{kp}. \tag{4.2}$$

For Z_k, logistic regression is typically used,

$$Z_k = P(R_k = 1|\mathbf{X}_k) = \frac{\exp(\mathbf{X}_k\hat{\theta})}{1 + \exp(\mathbf{X}_k\hat{\theta})}. \tag{4.3}$$

The AIPWT estimator is doubly robust because

$$E[\hat{\mu}_{AIPWT}] = \mu + E[\{\frac{R_k}{Z_k} - 1\}\{Y_k - m(\mathbf{X}_k, \hat{\beta})\}], \tag{4.4}$$

and under MAR assumption, $E[\{\frac{R_k}{Z_k} - 1\}\{Y_k - m(\mathbf{X}_k, \hat{\beta})\}] = 0$ if either the mean or propensity model is correctly specified. Full details of the proof can be found in Appendix D.

4.2.3 Penalized splines of propensity prediction (PSPP)

Another commonly used double robust estimator is the PSPP (*Zhang and Little*, 2009). First, equation (4.3) is computed followed by imputing Y_k using

$$Y_k = s[Z_k|\phi] + f(X_{k1}, \ldots, X_{kp}, \eta) + \epsilon_k \tag{4.5}$$

where $\epsilon_k \sim N(0, \sigma^2)$ and $s[Z_k|\phi]$ is the penalized spline formulation with H fixed knots for Z_k (*Ruppert et al.*, 2003) and usually $f(X_{k1}, \ldots, X_{kp}, \eta) = \eta_0 + \eta_1 X_{k1} + \ldots + \eta_p X_{kp}$.

For $s[P(Z_k)|\phi]$, we consider a penalized linear mixed effect model using cubic splines. μ is estimated by taking the mean of Y_ks after imputation.

PSPP is doubly robust because when the mean model is specified correctly, the propensity model may be treated as random noise. Hence, PSPP is consistent for μ. Suppose the propensity model is specified correctly and we omit the mean model. By the balancing property of the propensity score, $E[Y_k|Z_k] = g(Z_k)$ for an unknown function $g(.)$. Using a cubic spline for $g(.)$ allows the robust estimation of $g(Z_k)$ i.e., $E[Y_k|Z_k] = g(Z_k) \xrightarrow{p} \mu$. *Zhang and Little* (2009) showed that this property can be extended to any misspecified form of $f(X_{k1}, \ldots, X_{kp}, \eta)$ so that $E[Y_k|Z_k, \mathbf{X}_k] = g(Z_k, \mathbf{X}_k) \xrightarrow{p} \mu$ if the propensity model is correctly specified. Details of this proof can be found in Appendix E.

4.3 Proposed methods

4.3.1 Bayesian additive regression trees

4.3.1.1 Continuous outcomes

Suppose a continuous outcome Y_k with associated p covariates $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^T$ for $k = 1, \ldots, n$ subjects. BART models the outcome as

$$Y_k = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k \quad \epsilon_k \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{4.6}$$

where T_j is the j^{th} binary tree structure and $\mathbf{M}_j = (\mu_{1j}, \ldots, \mu_{b_j j})^T$ is the set of b_j terminal node parameters associated with tree structure T_j (*Chipman et al.*, 2010b). The function $g(\mathbf{X}_k, T_j, \mathbf{M}_j)$ can be viewed as the j^{th} function that assigns the mean μ_{ij} to the k^{th} outcome, Y_k. Typically, the number of trees m is fixed and no prior distribution is placed on m. *Chipman et al.* (2010b) suggested setting $m = 200$ as this performs well in many situations. Alternatively, cross-validation could be used to determine m (*Chipman et al.*, 2010b).

The joint prior distribution for (4.6) is

$$P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma]. \tag{4.7}$$

Assuming ϵ_k and $(T_j, \mathbf{M}_j)$ are independent and all m tree structures and terminal node parameters are independent between each other, we decompose equation (4.7) to become

$$[\prod_{j=1}^{m}\{\prod_{i=1}^{b_j} P(\mu_{ij}|T_j)\}P(T_j)]P(\sigma) \tag{4.8}$$

where $i = 1, \ldots, b_j$ indexes the terminal node parameters in tree j. Assigning priors to T_j, $\mu_{ij}|T_j$, and σ completes the setup of BART. The posterior draw of $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma|Y_k]$ is achieved using a combination of Bayesian backfitting (*Hastie and Tibshirani*, 2000) and Metropolis within Gibbs algorithm. Details of the suggested priors and hyperparameters for T_j, $\mu_{ij}|T_j$, and σ as well as the Bayesian backfitting and Metropolis within Gibbs algorithm can be found in *Chipman et al.* (2010b).

4.3.1.2 Binary outcomes

Extending BART to binary outcomes involve a modification of (4.6). First, let

$$G(\mathbf{X}_k) = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j). \tag{4.9}$$

Using the probit formulation, the binary outcomes Y_k can be linked to (4.9) using $P(Y_k = 1|\mathbf{X}_k) = \Phi[G(\mathbf{X}_k)]$ where $\Phi[.]$ is the cumulative density function of a standard normal distribution. This implicitly assumes that $\sigma \equiv 1$. Assuming that all m tree structures and terminal node parameters are independent, this implies that we only need priors for T_j and $\mu_{ij}|T_j$. Further details regarding the prior distribution of binary outcomes BART can be found in *Chipman et al.* (2010b). To draw from the posterior distribution, *Chipman et al.* (2010b) proposed the use of data augmentation (*Albert and Chib*, 1993). This method proceeds by first generating a latent variable Z_k according to

$$\begin{aligned}(Z_k|Y_k = 1, \mathbf{X}_k) &\sim N_{(0,\infty)}(G(\mathbf{X}_k), 1)\\ (Z_k|Y_k = 0, \mathbf{X}_k) &\sim N_{(-\infty,0)}(G(\mathbf{X}_k), 1),\end{aligned}$$

where $N_{(a,b)}(\mu, \sigma^2)$ is the truncated normal distribution with mean μ and variance σ^2 truncated to the range (a, b). Once Z_k is drawn, it is used to replace Y_k in the algorithm to calculate the posterior distribution of continuous outcomes BART with σ fixed at 1. Note that at each iteration, $G(\mathbf{X}_k)$ will be updated with the new $(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)$ draws from $P[(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m)|Z_k]$ so that an updated draw of the latent variable Z_k can be obtained.

4.3.2 Modifying the augmented inverse probability estimator with BART

To modify the AIPWT, we replace Z_k in equation (4.1) with

$$Z_k^* = P(R_k = 1|\mathbf{X}_k) = \Phi[G(\mathbf{X}_k)]. \tag{4.10}$$

$G(\mathbf{X}_k)$ is estimated using equation (4.9). Next, we model $m(\mathbf{X}_k, \hat{\beta})$ as a sum of regression trees i.e replace equation (4.2) with

$$Y_k = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k, \tag{4.11}$$

where $\epsilon_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. This allows the propensity model and mean model to be approximately close to the true generating model if the true model contains non-linear main and/or multiple-way interaction effects (*Rockova and van der Pas*, 2017).

4.3.3 Modifying PSPP using BART: Penalized splines of BART propensity prediction (PSBPP)

We modify PSPP by replacing Z_k in equation (4.5) with equation (4.10). This gives

$$Y_k = \phi_0 + \sum_{l=1}^{L} \phi_l Z_k^{*l} + \sum_{h=1}^{H} \phi_{L+h}(Z_k^* - \tau_h)_+^L + f(X_{k1}, \ldots, X_{kp}, \eta) + \epsilon_k. \tag{4.12}$$

Since BART was used to estimate the propensity score, we call this the penalized splines of BART propensity prediction (PSBPP).

4.3.4 Imputing directly using BART

Kang and Schafer (2007) argued that using the mean model is more appropriate in situations where misspecifying both the propensity and mean model is high. Since BART has the potential to approximate models with non-linear main and multiple-way interaction effects closely, it may be more straight forward to impute Y_k directly using equation (4.11).

4.3.5 Adding the BART propensity score to BART

Although PSPP uses a spline to reduce model misspecification for the prediction of Y_k given Z_k, possible interaction with $\mathbf{X}_k$ might still be present. Hence, using BART at both stages of modeling may be worth considering where

$$Y_k = \sum_{j=1}^{m} g(Z_k^*, \mathbf{X}_k, T_j, \mathbf{M}_j) + \epsilon_k, \tag{4.13}$$

with $\epsilon_k \overset{i.i.d.}{\sim} N(0, \sigma^2)$, i.e. impute the missing Y_k outcomes using equation (4.11) with the addition of the BART estimated propensity score Z_k^* as a predictor.

4.4 Simulations

We used three simulation scenarios to investigate how misspecification due to incorrect model would affect the bias, root mean squared error (RMSE), 95% coverage, and average length of the 95% confidence interval (AIL) of PSPP, AIPWT, PSBPP, AIPWT with BART, BART, and BARTps. For reference, we included the usual sample mean estimator before partial removal of outcomes (BD), the complete case estimation of the sample mean (CC), as well as imputation using only the mean model (MLR).

4.4.1 Linear interaction in mean model

In scenario 1, we included a linear two-way interaction term in both the propensity and mean model. We generated 2 predictors as $X_{k1} \sim N(0, 0.5)$ and $X_{k2} = X_{k1} + W_k$ where $W_k \sim N(0.25, 0.5)$. The true propensity model was specified as

$$\text{logit}[P(M_k = 1|X_{k1}, X_{k2})] = \frac{1}{3}\{0.15 + 0.75(X_{k1} + X_{k2}) - 2X_{k1}X_{k2}\} \tag{4.14}$$

and the mean model as

$$Y_k = 10.8125 + 0.75(X_{k1} + X_{k2}) - 2X_{k1}X_{k2} + \epsilon_k \tag{4.15}$$

where $\epsilon_k \overset{\text{iid}}{\sim} N(0, 2^2)$. The resulting population mean for this model is 10.

We consider four types of model misspecification:

(i) Propensity model and mean model are specified correctly as equations (4.14) and (4.15),

(ii) Mean model is misspecified by dropping the interaction term in equation (4.15),

(iii) Propensity model is misspecified by dropping the interaction term in equation

(4.14), and

(iv) Both propensity and mean models are misspecified by dropping the interaction terms in equations (4.14) and (4.15).

For BD and CC, note that because these estimators do not involve the specification of a propensity or mean model when estimating the population parameter μ, the estimators will be the same under all situations. For MLR, since it does not involve the specification of a propensity model, the MLR estimate under situations (i) and (iii), and (ii) and (iv) will be the same. Because BART automatically takes care of non-linear main effects and non-linear multiple-way interaction effects, the PSBPP estimator under situations (i) and (iii), and (ii) and (iv) will be the same. For the AIPWT with BART, BART, and BARTps, because each of them rely on BART to estimate their propensity and mean model, the estimators for all four situations will be the same.

4.4.2 Quadratic interaction in mean model

In scenario 2, the propensity model is still equation (4.14), but the mean model is now

$$Y_k = 11.875 + 0.75(X_{k1} + X_{k2}) - 2(X_{k1}X_{k2})^2 + \epsilon_k \tag{4.16}$$

where $\epsilon_k \overset{\text{iid}}{\sim} N(0, 2^2)$. X_{k1} and X_{k2} are generated as in subsection 4.1 and the population mean for this model is still 10. This scenario allows us to see how a slight non-linear effect in the simple two-way interaction of the mean model would affect the results of the eight mean estimation methods. The misspecification of the four situations is similar to the previous section in that the misspecification will remove the two-way interaction term.

4.4.3 Kang and Schafer (2007) example

Our third scenario was the *Kang and Schafer* (2007) example. The propensity model is given by

$$\text{logit}[P(R_k = 1|U_{k1}, U_{k2}, U_{k3}, U_{k4})] = -U_{k1} + 0.5U_{k2} - 0.25U_{k3} - 0.1U_{k4}, \qquad (4.17)$$

where $U_{kj} \overset{\text{iid}}{\sim} N(0,1)$, $j = 1, \dots, 4$. The mean model is given by

$$Y_k = 210 + 27.4U_{k1} + 13.7(U_{k2} + U_{k3} + U_{k4}) + \epsilon_k \qquad (4.18)$$

where $\epsilon_k \overset{\text{iid}}{\sim} N(0,1)$. In the misspecification situations, we assume that the U_{kj}s are latent and we only observe X_{kj}s which are given by

$$\begin{aligned} X_{k1} &= \frac{\exp[U_{k1}]}{2}, \\ X_{k2} &= \frac{U_{k2}}{1 + \exp[U_{k1}]}, \\ X_{k3} &= [\frac{U_{k1}U_{k3}}{25} + 0.6]^3, \text{ and} \\ X_{k4} &= [U_{k2} + U_{k4} + 20]^2. \end{aligned}$$

For the four situations, we use U_{kj}s to estimate the propensity and mean model when both models are specified correctly. When the propensity model is specified correctly but the mean model is misspecified, we use U_{kj} to estimate the propensity model but replace the U_{kj} with X_{kj} when estimating the mean model. When the mean model is specified correctly but the propensity model is misspecified, we replace U_{kj} with X_{kj} to estimate the propensity model but use U_{kj} to estimate the mean model. When both propensity and mean model are misspecified, we replace U_{kj} with X_{kj}s to estimate both the propensity and mean model.

For each of the simulation scenarios, we further split them into four situations: 1.

both the propensity and mean models are correctly specified; 2. the mean model is misspecified but the propensity model is correctly specified; 3. the propensity model is misspecified but the mean model is correctly specified; and 4. both models are misspecified. 500 simulations were used to estimate the empirical bias, RMSE, 95% coverage, and AIL. For PSPP and PSBPP, we used the linear truncated basis with 20 equally spaced knots on the propensity score, Z_k or Z_k^*, to estimate the penalized splines. We estimated the penalized splines following the method described in Chapter 9 of *Ruppert et al.* (2003). The 95% confidence interval (CI) and the length of this interval were estimated using a modified bootstrap approach with 200 resamples (*Heitjan and Little*, 1991) which accounts for the uncertainty of the parameter estimates during imputation. Essentially, Rubin's combining rules were applied to the D bootstrap means from the resampled datasets. This modified bootstrap approach accounts for the uncertainty of the parameter estimates during imputation. In addition to bootstrap, we also performed MI using the posterior mean of the propensity score in equations (4.5), (4.12), and (4.13) as well as MI using a posterior draw of the propensity score in equations (4.5), (4.12), and (4.13). Finally, we considered sample sizes of 500, 1,000, and 5,000 to investigate how changes in sample size would affect the performance of each estimator.

4.4.4 Results

Table 4.1 shows the result under scenario 1 for a sample size of 1,000. The CC estimators were substantially biased under all four types of misspecification. When the propensity model was correctly specified, both PSPP and AIPWT were approximately unbiased, although PSPP had much smaller RMSE and better coverage. The MLR, PSPP, and AIPWT estimators performed very well in terms of bias and RMSE when the mean model was correctly specified. When both models were misspecified, MLR, PSPP, and AIPWT were biased with coverage of both models decreasing dramatically. For PSBPP and AIPWT with BART, we observed that specifying the propensity model of PSPP using BART had little effect on the bias, RMSE, 95% coverage, and AIL when either one or both the propensity and mean model were correctly specified. When both models were misspecified, PSBPP was able to produce nearly unbiased estimation of the population mean and relatively similar AIL. In contrast, AIPWT with BART had bias and relatively poor coverage compared to AIPWT when at least one of the models in AIPWT was specified correctly. AIPWT with BART only performed better than AIPWT when both models were misspecified. Still, some bias and below nominal coverage remained. AIPWT with BART was more biased with larger RMSE and poorer coverage compared to PSBPP under all situations. BART alone generally had performance similar to AIPWT with BART, if slightly poorer in terms of bias and RMSE. For BARTps, the bias was reduced compared to BART with only X_ks as the predictors. Addition of BART propensity scores Z_k^* improves the 95% coverage compared to BART; nominal coverage was achieved for BARTps.

As the sample size increases, the bias, RMSE, and AIL of all methods reduce, and nominal 95% coverage increases (See Tables 1 to 3 in Appendix F). MI results were similar to bootstrap results (See Tables 4 to 6 in Appendix F). Using a posterior draw of the propensity scores instead of posterior mean increased bias slightly for PSBPP and BARTps (See Tables 7 to 9 in Appendix F).

Table 4.2 shows the result under scenario 2 for a sample size of 1,000. This scenario was more challenging compared to scenario 1, with larger bias, RMSE, and AIL with smaller 95% coverage for all methods. For the PSPP and AIPWT method, when the propensity model was correctly specified or when the mean model was correctly specified, we started to see substantial increases in the bias, RMSE, and AIL with a substantial reduction in the 95% coverage. When both models were misspecified, we started to see very poor performance: bias, RMSE, and AIL further increased

Table 4.1: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 1,000.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.2	0.34	0	0.09	94.2	0.34
CC	0.51	0.53	0.6	0.42	0.51	0.53	0.6	0.42
MLR	0	0.12	99	0.62	0.45	0.46	10	0.57
PSPP	0.01	0.14	99.8	0.78	0.05	0.13	97.4	0.61
AIPWT	0	0.12	94.4	0.47	0.04	0.18	87.2	0.6
PSBPP	0	0.13	99.2	0.64	-0.06	0.15	98.4	0.71
AIPWT with BART	0.11	0.17	78.2	0.44	0.11	0.17	78.2	0.44
BART	0.14	0.19	87.4	0.57	0.14	0.19	87.4	0.57
BARTps	0.07	0.14	95.8	0.6	0.07	0.14	95.8	0.6
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.2	0.34	0	0.09	94.2	0.34
CC	0.51	0.53	0.6	0.42	0.51	0.53	0.6	0.42
MLR	0	0.12	99	0.62	0.45	0.46	10	0.57
PSPP	0	0.12	99	0.63	0.22	0.26	85	0.78
AIPWT	0	0.12	92.8	0.46	0.43	0.45	5	0.43
PSBPP	0	0.13	99.2	0.64	-0.06	0.15	98.4	0.71
AIPWT with BART	0.11	0.17	78.2	0.44	0.11	0.17	78.2	0.44
BART	0.14	0.19	87.4	0.57	0.14	0.19	87.4	0.57
BARTps	0.07	0.14	95.8	0.6	0.07	0.14	95.8	0.6

with further reduction in the 95% coverage. For the PSBPP, the bias, RMSE, and AIL were similar to PSPP when either both models were correctly specified or only one model was correctly specified, although when the mean model was misspecified, PSBPP produced a better nominal 95% coverage. When both models were misspecified, PSBPP performed the best compared to all the other six methods with modest bias and approximately correct nominal coverage. For AIPWT with BART, BART was able to help the AIPWT estimator when both propensity and mean models were misspecified but when either one or both models were correctly specified, AIPWT with BART performed worse compared to AIPWT. In addition, the performance of AIPWT with BART when both propensity and mean models were misspecified was not as good compared to PSBPP. BART and AIPWT with BART performed similarly with BARTps having reduced bias and RMSE with improved the 95% coverage compared to BART. BARTps was still biased and nominal coverage was somewhat poor.

Similar to the linear interaction in mean model scenario, we found that as sample size increases, the bias, RMSE, and AIL of all methods reduce while 95% coverage increases (See Tables 10 to 12 in Appendix F). MI results echo those observed using bootstrap (See Tables 13 to 15 in Appendix F) while MI results using a posterior draw of the propensity score produced an increase in bias for PSBPP and BARTps methods (See Tables 16 to 18 in Appendix F).

Table 4.3 shows the result under the *Kang and Schafer* (2007) example for a sample size of 1,000. For the PSPP and AIPWT methods, we found that misspecification of the mean model increased the bias, RMSE, and AIL of these methods slightly more than misspecification of the propensity model does. When both models were misspecified, both models performed badly with the AIPWT estimator being highly unstable, producing a bias and RMSE more than the CC estimator. The standard MLR imputation performed fairly well even when the mean model was misspecified.

Table 4.2: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 1,000.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	91.8	0.86	0	0.24	91.8	0.86
CC	1.21	1.23	0.2	0.63	1.21	1.23	0.2	0.63
MLR	0	0.26	99	1.32	1.24	1.25	0	0.8
PSPP	0	0.26	98.8	1.33	0.21	0.44	81.2	2
AIPWT	0	0.26	91.2	0.93	0.22	0.72	67	1.68
PSBPP	0	0.26	98.6	1.33	0.13	0.35	94	2.16
AIPWT with BART	0.45	0.51	29.8	0.77	0.45	0.51	29.8	0.77
BART	0.52	0.57	42	0.97	0.52	0.57	42	0.97
BARTps	0.41	0.47	63.4	1.07	0.41	0.47	63.4	1.07
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	91.8	0.86	0	0.24	91.8	0.86
CC	1.21	1.23	0.2	0.63	1.21	1.23	0.2	0.63
MLR	0	0.26	99	1.32	1.24	1.25	0	0.8
PSPP	0	0.26	98.6	1.33	0.72	0.77	61.8	1.69
AIPWT	0	0.25	91	0.92	1.21	1.22	0	0.59
PSBPP	0	0.26	98.6	1.33	0.13	0.35	94	2.16
AIPWT with BART	0.45	0.51	29.8	0.77	0.45	0.51	29.8	0.77
BART	0.52	0.57	42	0.97	0.52	0.57	42	0.97
BARTps	0.41	0.47	63.4	1.07	0.41	0.47	63.4	1.07

Table 4.3: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 1,000.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.07	1.11	95.2	4.38	0.07	1.11	95.2	4.38
CC	-9.96	10.09	0	5.97	-9.96	10.09	0	5.97
MLR	0.07	1.11	99.4	6.38	-0.74	1.63	98	7.78
PSPP	0.06	1.11	99.4	6.38	-0.07	1.21	99.2	6.66
AIPWT	0.06	1.11	95.6	4.38	0.07	1.66	94.2	6.01
PSBPP	0.07	1.11	99.4	6.38	1.46	1.95	96.8	7.4
AIPWT with BART	-0.05	1.12	95.2	4.42	-0.31	1.19	93.8	4.61
BART	-0.13	1.12	99.6	6.38	-0.59	1.29	99.2	6.5
BARTps	0	1.11	99.4	6.46	0.39	1.23	99.2	6.8
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.07	1.11	95.2	4.38	0.07	1.11	95.2	4.38
CC	-9.96	10.09	0	5.97	-9.96	10.09	0	5.97
MLR	0.07	1.11	99.4	6.38	-0.74	1.63	98	7.78
PSPP	0.07	1.11	99.4	6.38	-2.12	2.52	77.2	6.29
AIPWT	-0.08	2.28	95.6	5.1	-35.69	477.13	41.2	196.51
PSBPP	0.07	1.11	99.4	6.38	-1.13	1.73	99	7.84
AIPWT with BART	-0.06	1.12	95.2	4.42	-0.45	1.24	93.2	4.62
BART	-0.13	1.12	99.6	6.38	-0.59	1.29	99.2	6.5
BARTps	-0.05	1.12	99.6	6.46	-0.52	1.27	99.2	6.7

For the PSBPP and AIPWT with BART, PSBPP performed better in terms of bias, RMSE, 95% coverage, and AIL when both the propensity and mean models are correctly specified or when only the mean model is correctly specified. When only the propensity model is correctly specified or when both models are misspecified, PSPP and AIPWT with BART had similar (slightly below nominal) coverage; AIPWT with BART had reduced bias, RMSE, and smaller AIL. Compared to AIPWT and PSPP, AIPWT with BART and PSBPP respectively showed improvements in performance when both models were misspecified. BART and BARTps generally performed well under all of the misspecification scenarios with BARTps having the better performance.

We note that as sample size increases, the bias, RMSE, and AIL of all methods reduce (See Tables 19 to 21 in Appendix F). The 95% coverage of all methods remained relatively similar as the sample size increased except for PSPP and AIPWT where coverage decreased as sample size increased. MI results produced similar conclusions with bootstrap (See Tables 22 to 27 in Appendix F).

4.5 Applications to Missing Data in Transportation Research

4.5.1 Imputing Delta-v in 2014 National Automotive Sampling System Crashworthiness Data System dataset

The NASS-CDS dataset is an annual three-stage representative probability sample of passenger vehicle crashes sponsored by the National Highway and Transportation Safety Authority (NHTSA). To be eligible, a crash must: (1) be police reported, (2) involve a harmful event (property damage and/or personal injury) resulting from a crash, and (3) involve at least one towed passenger car or light truck or van in transport on a traffic way. When a crash is selected, NASS-CDS investigators obtain police reports and conduct interviews with the occupants to collect information such as drivers age and sex, severity of injury measured using the KABCO scale (K=fatal; A=incapacitating Injury; B=non-incapacitating injury; C=possible injury; O=no injury; *Hedlund*, 2008), and the principal direction of impact from the crash. Often, the variable that estimates instantaneous change in velocity (delta-v), is missing. This variable is important because many studies have shown that delta-v is a strong predictor for the severity of injuries in tow-away crashes.

The 2014 NASS-CDS dataset contains 3,660 non-rollover passenger vehicle crashes. We converted all continuous variables to categorical and coded missingness in a variable as a level. We removed variables that had more than 80% missing, were derived from other variables in the dataset, or were 100% missing for vehicles missing delta-v. Simple descriptive statistics of the variables in our dataset stratified by missingness in delta-v can be found in Tables 1 to 9 of Appendix G. Out of the 44 variables, only climate, body type of vehicle, whether the trajectory data was reconstructed, make of the vehicle, model year, number of occupants, pre-event movement, road alignment, road surface type, number of seriously injured occupants, and driver's age, height, and weight were not statistically different between non-rollover passenger vehicles missing total delta-v and not missing delta-v.

We were interested in the population mean of the 2014 total delta-v and the unadjusted odds ratio of the police reported injury severity (any injury or severe injury) as a function of delta-v (between 15kph and 35kph, and more than 35kph, versus less than 15kph). To estimate the unadjusted odds ratio, we imputed the missing delta-v values and then categorized delta-v as: less than 15kph, between 15kph and 35kph, and more than 35kph. We ran a simple logistic regression with this categorized delta-v as the predictor and the police reported injury severity as the outcome. We compared the estimate and 95% confidence interval produced by CC, MLR, PSPP, AIPWT, PSBPP, and BARTps. To obtain the estimate and 95% confidence interval for all six methods, we employed the finite Bayesian bootstrap method developed by *Zhou et al.* (2016). This procedure allows us to compute a valid estimate and 95% confidence interval for our dataset while non-parametrically accounting for the sample design in the imputation.

The result of our analysis is given in Table 4.4. The population mean of delta-v estimated by PSBPP and BARTps were similar, more than 21.7 kph while MLR, CC, PSPP, and AIPWT suggested that the population delta-v was about 21.5 kph. The 95% confidence interval of PSBPP and BARTps were also slightly wider compared to MI, CC, PSPP, and AIPWT. For the odds ratios, PSPP and PSBPP tended to agree with each other under any injury, CC and AIPWT suggested somewhat similar results, while BARTps and MLR results were more similar. All methods suggested a significant association between delta-v and presence of injury with higher delta-v levels associated with a higher odds of experiencing injury in a non-rollover passenger vehicle crash. For severe versus non-severe injury, we observe similar results as injury

Table 4.4: Estimated population mean, and unadjusted odds ratios of injury severity, any injury (OR_{NULL}) or severe injury (OR_{SEV}), where reference group is delta-v less than 15 kph ($X < 15$).

	$Y_{\text{delta-v}}$		OR_{NULL}		OR_{NULL}	
Method	Estimate	95% CI	$15 \leq X \leq 35$	95% CI	$X > 35$	95% CI
CC	21.57	(20.64 , 22.47)	1.72	(1.15 , 2.43)	5.88	(2.94 , 8.79)
MLR	21.57	(20.64 , 22.47)	1.37	(1.08 , 1.76)	2.78	(1.92 , 3.52)
PSPP	21.55	(20.06 , 22.99)	1.86	(1.23 , 2.84)	7.93	(3.66 , 13.33)
AIPWT	21.5	(20.01 , 23.33)	1.5	(1 , 2.12)	5.87	(3.39 , 9.25)
PSBPP	21.75	(18.29 , 25.61)	1.86	(1.11 , 3.03)	7.67	(2.39 , 13.74)
BARTps	21.9	(18.51 , 24.79)	1.62	(1.08 , 2.36)	2.95	(1.54 , 5.18)
	OR_{SEV}		OR_{SEV}			
Method	$15 \leq X \leq 35$	95% CI	$X > 35$	95% CI		
CC	2.31	(1.49 , 3.64)	17.99	(9.31 , 30.42)		
MLR	1.43	(1.19 , 1.69)	6.08	(4.25 , 8.17)		
PSPP	3.19	(1.82 , 5.21)	33.73	(16.16 , 60.17)		
AIPWT	1.58	(1 , 2.21)	14.67	(9.19 , 21.77)		
PSBPP	3.3	(1.51 , 6.8)	33.23	(9.63 , 71.67)		
BARTps	1.77	(1.18 , 2.64)	7.48	(4.09 , 12.27)		

versus no injury in that PSPP and PSBPP suggested similar results, CC and AIPWT suggested similar results, and BARTps and MLR suggested similar results. Again all methods suggested a significant association between delta-v and presence of injury with higher delta-v levels associated with a higher odds of experiencing injury in a non-rollover passenger vehicle crash. Given that CC results and AIPWT results were similar and BARTps and MLR results were similar, we suspect there to be non-linear main and interaction effects between delta-v and the NASS-CDS variables as well as non-linear main and interaction effects between the missingness of delta-v and the NASS-CDS variables.

4.5.2 Imputing Blood Alcohol Concentration levels in 2015 Fatality Analysis Reporting System dataset

The FARS releases information annually from all fatal motor vehicle crashes that occur on US public roads. Information collected include age, surface conditions, gross weight of vehicle, type of road, and accident type. Of the information collected, BAC, which is used to identify alcohol involvement in fatal crashes, is often missing. The fact that alcohol involvement is more commonly reported in fatal crashes compared to personal injury and property-damage-only crashes makes this issue more concerning because high levels of missingness in BAC hinders the investigation of the trend and extent of alcohol involvement in fatal crashes, the successful identification of high-risk groups for countermeasures, and evaluation of drunk-driving prevention programs.

Due to the importance of the BAC measure, NHTSA considered several approaches to remedy the missing data problem before deciding to use MI in 2002 (*Subramaniam*, 2002). Although MI was a great improvement from previous imputation methods (*Klein*, 1986), misspecification of the model in MI could lead to biased results. Replacing the imputation methods with DR estimators like PSPP and AIPWT could further bias results if the propensity and mean model were not specified correctly. Hence, we applied our proposed methods to the 2015 FARS dataset to impute BAC levels and compared the imputation results with existing MI results provided by the FARS dataset.

Details of how the publicly available imputed BAC values were clculated for the 2015 dataset can be found in *Rubin et al.* (1998) Section 3. We modified this imputation strategy slightly. First, we used the imputed 2015 BAC FARS dataset to determine all the 55,502 "actively-involved" subjects eligible for imputation (See *Rubin et al.*, 1998, , Section 2). We restrict our attention to passenger vehicles as defined in Section 3 of *Rubin et al.* (1998) which gave us 19,425 subjects. We recoded continuous variables as categorical variables and coded missing entries as a category in all variables. We removed variables that had more than 80% missing, derived from other variables in 2015 FARS, or 100% missing for subjects missing BAC values. Simple descriptive statistics of the variables in our dataset stratified by missingness in BAC can be found in Tables 10 to 21 of Appendix G. All variables except whether crash occurred within the boundaries of a work zone were significantly different between subjects missing BAC and subjects not missing BAC.

We impute BAC values as follows:

1. We employed binary BART to predict BAC=0 ($Y = 0$) versus $BAC > 0$

($Y = 1$) using all available predictors (See Tables 33 to 36 in Appendix G for all predictors employed).

2. We set the predicted BAC=0 values as 0 and focus on the set of observed $BAC > 0$ and predicted $BAC > 0$. For the observed $BAC > 0$, we employed a Box-Cox transformation (*Box and Cox*, 1964) using all available predictors to obtain the Box-Cox transformation parameter $\hat{\lambda}$. We used $\tilde{\lambda} = \hat{\lambda} + 1$ as suggested by *Rubin et al.* (1998).

3. We next imputed the Box-Cox transformed BAC value for the predicted $BAC > 0$ using the following methods, PSPP, AIPWT, PSBPP, and BARTps. For the transformed BAC values that were predicted to be negative, we set them as 0. For transformed BAC values that were predicted to be positive, an inverse transformation was applied to the predicted transformed BAC values to obtain the predicted BAC value in the original scale.

4. We drew 200 resampled datasets and repeated Steps 1-3 on each dataset. Rubin's combine rules were used to estimate the imputation uncertainty.

For the estimate of interest, we examined the population mean of the BAC value, the proportion of BAC more than .010 g/100 ml, and the proportion of BAC more than .100 g/100 ml among passenger vehicles in 2015.

Table 4.5 gives the result of our analysis. MLR was calculated using the imputed BAC values provided in the 2015 FARS dataset. Comparing CC and MLR, we can see that CC likely overestimates the population mean of BAC as well as the proportion of subjects with BAC more than .010 and .100 g/ 100 ml. MLR estimates that the population mean BAC value was 4% with the proportion of subjects with BAC more than .010 estimated at 24% and for the proportion of subjects with BAC more than .100 estimated at 18%. MLR results were significantly different from the imputed values estimated by PSPP and AIPWT. PSPP and AIPWT suggested that

Table 4.5: Estimated population mean of BAC, proportion of $BAC > .010$, and proportion of $BAC > .100$. All values in precentages.

	Mean		$BAC > 1\%$		$BAC > 10\%$	
Method	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
CC	5.72	(5.53, 5.91)	0.34	(0.33, 0.35)	0.26	(0.26, 0.27)
MLR	3.97	(3.83, 4.11)	0.24	(0.24, 0.25)	0.18	(0.18, 0.19)
PSPP	3.07	(2.89, 3.26)	0.18	(0.17, 0.19)	0.14	(0.13, 0.15)
AIPWT	3.12	(2.10, 4.14)	0.16	(0.15, 0.16)	0.13	(0.13, 0.14)
PSBPP	3.08	(2.88, 3.28)	0.18	(0.17, 0.19)	0.14	(0.13, 0.15)
BARTps	3.13	(2.95, 3.27)	0.19	(0.18, 0.19)	0.15	(0.14, 0.15)

the population mean BAC value was about 3.1% while the proportion of subjects with BAC more than .010 was estimated at about 18% and 16% respectively while the proportion of subjects with BAC more than .100 was estimated at about 14%. PSBPP and BARTps were similar compared to PSPP and AIPWT. The significant difference between MLR versus the doubly robust and robust-squared methods suggest that there is likely some non-linear relation between BAC and the variables in the FARS dataset. The non-significant difference in the results produced by PSPP, PSBPP, and BARTps further suggests that the relationship between missingness in BAC and the rest of the FARS variables is linear without any interactions.

4.6 Discussion

In many situations, researchers would not know the true propensity and mean model and thus both models have a high chance that they will be misspecified, limiting the value of the doubly-robust property. Even if the misspecification was mild for example, removal of the two-way interaction terms when the true mean model included a linear two-way interaction term or quadratic two-way interaction term, the resulting bias may be almost as large as a complete case analysis. Hence we consider use of a highly flexible estimation method – specifically Bayesian Additive Regression Trees or BART – to reduce the risk of model misspecification. We consider the use of

BART in propensity score estimation when using the penalized spline of propensity prediction (PSPPB) or when using the augmented inverse probability weighted estimator (AIPWT with BART). We also consider direct imputation using BART, and a "double flexible" robust method that adds a BART-estimated propensity score to the BART imputation, so that both the mean and propensity are estimated in the PSPP using BART (BARTps).

By using BART, we were able to demonstrate the reduction in bias and RMSE of the double robust estimators when both propensity and mean models were misspecified, with little loss in efficiency when either one or both of the mean and propensity models can be correctly specified by standard linear or logistic regression. Our simulation study suggests that PSPP with BART performs considerably better than AIPWT with BART under settings with missing interaction terms. However, when both the propensity and mean model are complex, BARTps tends to perform better. Hence, we suggest PSBPP and BARTps as the preferred methods for imputing datasets under MAR, while acknowledging that these recommendations are empirically based on simulations that are somewhat limited in nature.

We also found in our simulation results that MI using a posterior draw of the propensity score in equations (4.12) and (4.13) increased bias compared to using the posterior mean of the propensity score for linear and quadratic interaction scenarios. This is because the propensity model in both scenarios tended to create datasets where there is not much overlap in the predictors for response and non-response. Hence, the researcher might want to rely on bootstrap to obtain the uncertainty of PSBPP and BARTps during analysis.

Although we focused our attention on MAR for a continuous outcome, extension to a binary outcome is possible using generalized additive models or generalized linear mixed models for the PSPPB setting, or use of latent variables models (e.g, probit models) for PSPPB or the BARTps setting. The MAR assumption remains a restriction in these “robust” estimation methods; extensions to NMAR mechanisms remains a topic for further research.

CHAPTER V

Accounting for selection bias due to death in estimating the effect of wealth shock on cognition for the Health and Retirement Study

5.1 Introduction

Late middle age adults commonly experience chronic health conditions like high blood pressure or diabetes as well as declining cognitive abilities. Factors known to be associated with accelerated decrease in cognitive abilities include smoking, high alcohol consumption, physical inactivity, high dietary intake of sodium and saturated fats, low dietary intake of fruits and vegetables (*Lee et al.*, 2010; *Stuck et al.*, 1999); hypertension, elevated serum cholesterol, diabetes, obesity, cerebrovascular and cardiovascular disease (*Plassman et al.*, 2010); depression, lower socioeconomic status, and exposure to acute stressful life events and chronic perceived stress (*Krieger*, 2001). In particular, the acute stress of a sudden decrease in wealth – "a negative wealth shock" – may have a negative impact on the cognitive ability of late middle aged adults. Because income typically exceeds consumption at this stage in life, sudden decreases in wealth during this period not only decrease the amount of wealth saved for retirement, but there are fewer remaining years left to replenish the lost wealth (*Butrica et al.*, 2010). The stress of losing substantial wealth during the savings period of the life cycle coupled with the pressure to replenish the lost wealth can lead to stress-related health conditions which in turn reduces the cognitive ability of an individual (*Shrira et al.*, 2011). In addition, individuals who have received a negative wealth shock may have to reduce consumption of health-enhancing goods and services which in turn leads to poor management of existing chronic conditions, further reducing cognitive abilities (*Friedman*, 1956).

Three issues arise when trying to estimate the causal effect of a negative wealth shock on cognitive ability. The first of these is the lack of randomization: negative wealth shocks are not randomly distributed in the population, but rather are confounded by factors such as gender and socio-economic status. The second issue is confounding by indication: the risk of the wealth shock at any point in time may depend on the prior cognitive ability up to the point. Finally, we face the fact that a sufficiently large fraction of the sample and the population will die during our follow-up, leading to "censoring by death". Those observed to have survived a negative wealth shock include those who would survive under either condition together with those that would survive only if they experienced a negative wealth shock (if any), while those observed to have survived in the absence of a negative wealth shock include those that would survive under either condition together with those that would survive only in the absence of a negative wealth shock. These "missing values" associated with cognition among the deceased are different from the measure of cognition being "missing" due to dropout, where the cognitive ability measure exists but is unobserved. As with wealth shock, death is not a random occurrence, and is positively associated with demographic measures that increase the risk of a negative wealth shock, increased cognitive ability decline, and the experience of a negative wealth shock. Hence, the measure for cognitive ability may be confounded by death if not considered appropriately.

Methods have been developed to deal with these barriers to causal inference. To deal with the lack of randomization, we might hope that, conditional on available covariates, negative wealth shocks would truly be random. In this case, conditioning on the probability of receiving a negative wealth shock as a function of these covariates – the propensity scores (*Rosenbaum and Rubin*, 1983) – can be used to remove the effect of confounding, either by regression, matching, or weighting (*Imbens and Rubin*, 2015). For the second issue – confounding by indication – marginal structural models (MSM, *Robins et al.*, 2000) and more recently, penalized spline of propensity methods in treatment comparisons (PENCOMP, *Zhou et al.*, 2018), have been used to account for confounding by the time-dependence association of the cognitive measures, either by weighting using the inverse probability of treatment actually received based on the previous values of the time-varying covariates and outcomes (MSM), or by imputation of the missing counterfactual values (PENCOMP). For censoring by death, MSMs have typically been extended by multiplying the treatment assignment weights with the inverse of the predicted probability of death. The issue with this approach – perhaps under appreciated – is that the resulting pseudo-population is not only balanced with respect to exposure "assignment", but also "immortal", in the sense that those more likely to die are upweighted so that the population over time resembles that would have been obtained in the absence of death up till time t (*Chaix et al.*, 2012). This is arguably not a sensible population for inference, at least from a policy and public health perspective.

A more refined approach would be to compare the difference in the effect of negative wealth shock on cognitive ability among subjects who would have survived whether they experienced a negative wealth shock or not. This approach is consistent with the potential outcomes approach of *Neyman* (1934) and *Rubin* (1974), which defines causal effects as the within-subject difference of an outcome at a particular time under different exposure or treatment regimen, averaged over the population. This idea is not new (*Elliott et al.*, 2006) and can be viewed as a specific example of the principal stratification (PS) method discussed in *Frangakis and Rubin* (2002). Our innovation here is to embed this in a longitudinal setting where confounding by indication is present. We view this as a large missing data problem where survival status and, among survivors, unobserved outcomes under a given treatment pattern are imputed. We extend the method proposed in Example 3 of *Elliott and Little* (2015), which provides a Bayesian MSM approach to compare two treatments at two time points. This approach was further extended by PENCOMP in *Zhou et al.* (2018) which, like augmented inverse probability weighting (AIPWT, *Robins et al.*, 1994), has a doubly-robust property in that if either the mean or propensity model is correctly specified, consistent estimates of the causal effect will be obtained. We modified PENCOMP slightly using Bayesian additive regression trees (BART), a flexible model to ease the burden of model specification by the researcher, and apply this to our proposed method.

We organize our paper as follows. We set up the framework for our problem, and provide a brief review of of MSM, PENCOMP, and Bayesian additive regression trees (BART) in Section 2. We develop our proposed method in Section 3. We then explore some of the empirical properties of our proposed method compared to a naïve method and MSM using a simulation study in Section 4. Section 5 describes the HRS data and the results of our negative wealth shock analysis. Section 6 concludes with a discussion of the implication of our results as well as future work.

5.2 Review of Relevant Methods

5.2.1 Setup and notation

Let $V = \{V_1, V_2, \ldots, V_p\}$ be p baseline covariates, Z_t be the treatment allocation at time $t = 1, \ldots, T$ where $Z_t = 1$ indicates a subject receiving a negative wealth shock at t and $Z_t = 0$ indicates no negative wealth shock, and $W_t = \{W_{1t}, W_{2t}, \ldots, W_{qt}\}$

be q covariates that may vary with time, but are unaffected by a given treatment regimen. For example, fixed covariates by definition would belong to this class. Let $Y_{Z_1,\ldots,Z_t}$ be the potential outcome under treatments $Z_1,\ldots,Z_t$ and $X_{Z_1,\ldots,Z_t} = \{X_{Z_1,\ldots,Z_t,1}, X_{Z_1,\ldots,Z_t,2},\ldots,X_{Z_1,\ldots,Z_t,r}\}$ be the time-varying covariates affected by treatments $Z_1,\ldots,Z_t$. Similarly, we define the potential survival indicator $S_{Z_1,\ldots,Z_{t-1}}$, for survival at time t. The survival outcome at t measures whether a subject would survive after being exposed to treatment $Z_1,\ldots,Z_{t-1}$; hence, the lagged notation for the potential survival outcome, $S_{Z_1,\ldots,Z_{t-1}}$. v, z_t, w_t, $y_{z_1,\ldots,z_t}$, $x_{z_1,\ldots,z_t}$, and $s_{z_1,\ldots,z_t}$ indicate the observed baseline, treatment allocation, time varying covariates unaffected by a given treatment regimen, outcome, time-varying covariates affected by a given treatment regime, and survival status variables respectively. As in *Pool et al.* (2018), we assume that a negative wealth shock is an “absorbing state” so that once a subject receives a negative wealth shock at time t, i.e. $Z_t = 1$, the subject is “forever” shocked, i.e. $Z_{t+1} = \ldots = Z_T = 1$. Note that this need not be the case for a more general set up where we could have $Z_t = 0$ when $Z_j = 1$ for any $j = 1,\ldots,t-1$. In our context, the potential outcomes for time $t = 2$ are then $Y_{Z_1=0,Z_2=0} = Y_{00}$, $Y_{Z_1=0,Z_2=1} = Y_{01}$, and $Y_{Z_1=1,Z_2=1} = Y_{11}$; similarly, $X_{Z_1=0,Z_2=0} = X_{00}$, $X_{Z_1=0,Z_2=1} = X_{01}$, and $X_{Z_1=1,Z_2=1} = X_{11}$ for time-varying covariates under the various treatment regimes; and $S_{Z_1=0} = S_0$, $S_{Z_1=1} = S_1$ for survival states. Subjects who die at time t have structurally missing data for outcomes and covariates i.e., $S_0 = 0$ implies that $Y_{00} = Y_{01} = NA$ and $X_{00} = X_{01} = NA$, while $S_1 = 0$ implies that $Y_{11} = NA$ and $X_{11} = NA$, where ‘NA’ indicates a structurally missing observation.

5.2.2 Marginal structural model

To estimate the causal effect for confounding by indication and censoring by death problems, MSM makes the following assumptions. First, MSM assumes that

$$P(S_{z_1,\ldots,z_{t-1}}|z_1,\ldots,z_{t-1},y_{z_1},\ldots,y_{z_1,\ldots,z_{t-1}},x_{z_1},\ldots,x_{z_1,\ldots,z_{t-1}},w_1,\ldots,w_{t-1},v)>0. \tag{5.1}$$

and

$$P(Z_t|z_1,\ldots,z_{t-1},y_{z_1},\ldots,y_{z_1,\ldots,z_{t-1}},x_{z_1},\ldots,x_{z_1,\ldots,z_{t-1}},w_1,\ldots,w_{t-1},v)>0 \tag{5.2}$$

for any z_t i.e. the probability of survival under treatment profile $z_1,\ldots,z_{t-1}$ and the probability of treatment allocation for time t is bounded away from 0. This is an extension of the standard positivity assumption to allow that at least some subjects will survive under a given treatment regimen. Second, MSM assumes that there is no interference between subjects i.e. the potential outcome of subject i, $Y_{i,Z_1,\ldots,Z_t}=Y_{i,z_1,\ldots,z_t}$, is independent of whatever treatment regimen subject j is allocated to $i\neq j$. Third, MSM assumes no unmeasured confounding and sequential randomization condition

$$Y_{Z_1,\ldots,Z_t}\perp Z_t|z_1,\ldots,z_{t-1},y_{z_1,\ldots,z_{t-1}},\ldots,y_{z_1},x_{z_1,\ldots,z_{t-1}},\ldots,x_{z_1},w_1,\ldots,w_{t-1},v.$$

Finally, MSM assumes that the model specifications for Equations 5.1, 5.2, and

$$Y_{z_1,\ldots,z_t}|z_1,\ldots,z_t,y_{z_1,\ldots,z_{t-1}},\ldots,y_{z_1},x_{z_1,\ldots,z_{t-1}},\ldots,x_{z_1},w_1,\ldots,w_{t-1},v$$

are correct.

With these assumptions in place, $E[Y_{z_1,\ldots,z_t}-Y_{z'_1,\ldots,z'_t}]$ (note that this estimand is not conditioned on the survival status) is obtained by maximizing the weighted likelihood of

$$\prod_{i=1}^{n} f(Y_{i;z_1,\ldots,z_t}|\theta_{it})^{w_{it}}, \tag{5.3}$$

where i indexes the subjects and θ_{it} are the parameters involved in the model for $Y_{i;z_1,\ldots,z_t}$ and

$$w_{it} = [\prod_{j=1}^{t} P(Z_{ij} = z_{ij}|z_{i1},\ldots,z_{i,j-1},y_{i1},\ldots,y_{i,j-1},x_{i1},\ldots,x_{i,j-1},w_{i1},\ldots,w_{i,j-1},v_i;\tau_j)]^{-1}. \tag{5.4}$$

By weighting using the inverse probability of receiving the observed treatment regime given all covariates and previous treatments, the association between treatment and all observed confounders, including confounding by indication, are broken. Under these four assumptions, inference about the treatment effects under a pseudo-population in which treatment is randomized can then be obtained.

Similarly, this weighting method can be used to remove bias due to dropout. Let $R_i = 1$ indicate that the subject's cognitive score is observed and $R_i = 0$ indicate that the subject's cognitive score is missing. The weight used to account for missing cognitive score is

$$w_{it}^{r} = [\prod_{j=1}^{t} P(R_{ij} = r_{ij}|r_{i1},\ldots,r_{i,j-1},z_{i1},\ldots,z_{i,j-1},y_{i1},\ldots,y_{i,j-1},x_{i1},\ldots,x_{i,j-1},w_{i1},\ldots,w_{i,j-1},v_i;\gamma_j)]^{-1}. \tag{5.5}$$

Finally, death is typically treated as equivalent to dropout in MSM (*Do et al.*, 2013; *Pool et al.*, 2018). Let $D_{it} = 1$ indicate that subject i is dead at time t and $D_{it} = 0$ indicate that the subject survived at time t (thus $D_{it} = 1 - S_{it}$). The weight for death censoring is then

$$w_{it}^{d} = [\prod_{j=1}^{t} P(D_{ij} = d_{ij}|z_{i1},\ldots,z_{i,j-1},y_{i1},\ldots,y_{i,j-1},x_{i1},\ldots,x_{i,j-1},w_{i1},\ldots,w_{i,j-1},v_i;\lambda_j)]^{-1}. \tag{5.6}$$

Assuming that these three weights are independent of each other, the final weight that we used becomes $w_{it}^{f} = w_{it}w_{it}^{d}w_{it}^{r}$. To stabilize the weights, the numerators of Equations 5.4, 5.5, and 5.6 are replaced by the marginal probabilities of treatment, dropout, and death at baseline given by

$$\prod_{j=1}^{t} P(Z_{ij} = z_{ij} | z_{i1}, \ldots, z_{i,j-1}, v_i; \tau_j'),$$

$$\prod_{j=1}^{t} P(R_{ij} = r_{ij} | r_{i1}, \ldots, r_{i,j-1}, v_i; \gamma_j'),$$

and

$$\prod_{j=1}^{t} P(D_{ij} = d_{ij} | v_i; \lambda_j')$$

respectively. We use the stabilized weights in our simulations and analysis.

5.2.3 Penalized Spline of Propensity Methods for Treatment Comparison

PENCOMP uses the same four assumptions made by MSM excluding Equation 5.1 for confounding by indication problems. Full details of PENCOMP can be found in *Zhou et al.* (2018). We briefly describe the algorithm for PENCOMP using multiple imputation (MI) with longitudinal treatment assignments here. Without loss of generality, we assume no time-varying covariates in the data.

1. For $b = 1, \ldots, B$, generate a bootstrap sample $S^{(b)}$ from the original data S by sampling units with replacement, stratified on treatment group. For each sample b, carry out steps 2-7.

2. Estimate a logistic regression model for the distribution of Z_1 given baseline covariates V with regression parameters γ_{z_1}. Estimate the propensity to be assigned treatment $Z_1 = z_1$ as $\hat{P}_{z_1}(V) = Pr(Z_1 = z_1 | V; \hat{\gamma}_{z_1}^b)$, where $\hat{\gamma}_{z_1}^b$ is the maximum likelihood (ML) estimate of γ_{z_1}. Define $\hat{P}_{z_1}^* = \log[\frac{\hat{P}_{z_1}(V)}{1-\hat{P}_{z_1}(V)}]$.

3. Using the cases assigned to treatment group $Z_1 = z_1$, estimate a normal linear

regression of Y_{z_1} on V, with mean

$$E(Y_{z_1}|V, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = s(\hat{P}^*_{z_1}|\theta_{z_1}) + g_{z_1}(\hat{P}^*_{z_1}, V; \beta_{z_1}), \tag{5.7}$$

where $s(\hat{P}*_{z_1}|\theta_{z_1})$ denotes a penalized spline with fixed knots and parameters θ_{z_1} and $g_{z_1}(.)$ represents a parametric function of other predictors of the outcome, indexed by parameters β_{z_1}. One of the covariates might be omitted to avoid collinearity in the covariates in Equation 5.7.

4. For $z_1 = 0, 1$, impute the values of Y_{z_1} for subjects in treatment group $1 - z_1$ in the original data with draws from the predictive distribution of Y_{z_1} given V from the regression in Step 3, with the ML estimates $\hat{\theta}^{(b)}_{z_1}, \hat{\beta}^{(b)}_{z_1}$ substituted for the parameters $\theta^{(b)}_{z_1}, \beta^{(b)}_{z_1}$.

5. Estimate a logistic regression model for the distribution of Z_2 given $V, Z_1, (Y_0, Y_1)$, with regression parameters γ_{z_2} and missing values of (Y_0, Y_1) imputed from Step 4. Estimate the propensity to be assigned treatment $Z_2 = z_2$ given Z_1, Y_{Z_1}, and V as $\hat{P}_{z_2}(Z_1, Y_{Z_1}, V) = Pr(Z_2 = z_2|Z_1 = z_1, Y_{z_1}, V; \hat{\gamma}^{(b)}_{z_2})$, where $\hat{\gamma}^{(b)}_{z_2}$ is the ML estimate of γ_{z_2}. The probability of treatment regimen $(Z_1 = z_1, Z_2 = z_2)$ is denoted as $\hat{P}_{z_1 z_2} = \hat{P}_{z_1}(V)\hat{P}_{z_2}(Z_1, Y_{Z_1}, V)$, and define $\hat{P}^*_{z_1,z_2} = \log[\frac{\hat{P}_{z_1 z_2}}{1-\hat{P}_{z_1 z_2}}]$.

6. Using the cases assigned to treatment group (z_1, z_2), estimate a normal linear regression of Y_{z_1,z_2} on Z_2, Z_1, Y_{Z_1}, and V with mean

$$E(Y_{z_1,z_2}|V, Y_{z_1}, Z_1 = z_1, Z_1 = z_2, \theta_{z_1,z_2}, \beta_{z_1,z_2}) = s(\hat{P}^*_{z_1,z_2}|\theta_{z_1,z_2}) + g_{z_1,z_2}(\hat{P}^*_{z_1,z_2}, Z_2, Z_1, Y_{Z_1}, V; \beta_{z_1,z_2}). \tag{5.8}$$

7. For each combination of (z_1, z_2) impute the values of Y_{z_1,z_2} for subjects not assigned this treatment combination in the original data with draws from the predictive distribution of Y_{z_1,z_2} in Step 6, with ML estimates $\hat{\theta}^{(b)}_{z_1,z_2}, \hat{\beta}^{(b)}_{z_1,z_2}$ substituted for the parameters $\theta^{(b)}_{z_1,z_2}, \beta^{(b)}_{z_1,z_2}$. Let $\hat{\Delta}^{(b)}_{01,00} = E[Y_{01} - Y_{00}]$, $\hat{\Delta}^{(b)}_{11,00} = E[Y_{11} - Y_{00}]$,

and $\hat{\Delta}^{(b)}_{11,01} = E[Y_{11} - Y_{01}]$ denote the average treatment effects, $\hat{\Delta}^{(b)}_{jk,lm}$, with associated pooled variance estimates $W^{(b)}_{jk,lm}$, based on the observed and imputed values of Y for each treatment regimen.

8. The MI estimate of $\Delta_{jk,lm}$ is then $\bar{\Delta}_{jk,lm,B} = \sum_{b=1}^{B} \hat{\Delta}^{(b)}_{jk,lm}$, and the MI estimate of the variance of $\bar{\Delta}_{jk,lm}$ is $T_B = \bar{W}_{jk,lm,B} + (1+1/B)D_{jk,lm,B}$, where $\bar{W}_{jk,lm,B} = \sum_{b=1}^{B} W^{(b)}_{jk,lm}/B$, $D_{jk,lm,B} = \sum_{b=1}^{B} \frac{(\hat{\Delta}^{(b)}_{jk,lm} - \bar{\Delta}_{jk,lm,B})^2}{B-1}$. The estimate $\Delta_{jk,lm}$ follows a t distribution with degree of freedom ν, $\frac{\Delta_{jk,lm} - \bar{\Delta}_{jk,lm,B}}{\sqrt{T_B}} \sim t_\nu$, where $\nu = (B-1)(1 + \frac{\bar{W}_{jk,lm,B}}{D_{jk,lm,B}(B+1)})^2$.

5.2.4 Bayesian additive regression trees

BART (*Chipman et al.*, 2010b) is a flexible estimation technique for any arbitrary function. Suppose we have a continuous outcome Y and corresponding p predictors $X = (X_1, \ldots, X_p)$. Suppose Y is related to X via

$$Y = f(X) + e \tag{5.9}$$

where $f(.)$ is any arbitrary function which could involve complicated non-linear and multiple-way interactions and $e \sim N(0, \sigma^2)$. Formally, BART is written as

$$Y = \sum_{j=1}^{m} g(X, T_j, M_j) + e \tag{5.10}$$

where (T_j, M_j) is the joint distribution of the j^{th} binary tree structure T_j with its corresponding b_j terminal node parameters $M_j = (\mu_{1j}, \ldots, \mu_{b_j j})$. m is the number of regression trees used to estimate $f(X)$ and it is usually fixed at 200.

BART is able to model multiple-way interactions by using regression trees. In essence, a binary regression tree in BART may be viewed as a penalized form of an Analysis of Variance (ANOVA) model. When the binary regression tree only splits on one variable for the whole tree, a main effects model is obtained. When the regression tree involve splits on many different variables, a multiple-way interaction model is obtained. BART combines all m regression trees together in an additive manner to obtain non-linear estimates of the main and interaction effects. This additive procedure is done by first 'breaking' Y into m equal 'pieces' and fitting a regression tree to each piece. Subsequently, the regression tree in each m piece is then estimated by looking at the residual produced by the other $m-1$ most updated regression trees. MCMC procedures are then used to obtain the posterior distribution of $f(X)$. When the default priors of BART suggested by *Chipman et al.* (2010b) are assumed, the MCMC ensures that the eventual distribution of the the sum of regression trees is concentrated around the true distribution of the model (*Rockova and van der Pas*, 2017).

For binary outcomes, BART uses a probit link where

$$P(Y=1|X) = \Phi(\sum_{j=1}^{m} g[X, T_j, M_j]) \tag{5.11}$$

where $\Phi(.)$ is the cdf of a standard normal distribution. Estimation of the posterior distribution is similar to that of continuous outcomes but with the use of data augmentation methods, i.e. draw a continuous latent variable based on whether $Y = 1$ or $Y = 0$ and then run the BART algorithm on the drawn latent variables.

Kapelner and Bleich (2015) suggested a procedure to allow the BART algorithm to include covariates that might contain missing values. In brief, the missingness in the covariates are not imputed but instead, viewed as a 'value level' in the MCMC algorithm. The MCMC algorithm then 'sends' missing data to terminal nodes in the regression trees that would maximize the likelihood. This is termed as "Missing Incorporated in Attributes" (MIA, *Twala et al.*, 2008, Section 2). *Kapelner and Bleich* (2015) showed using simulation examples that incorporating MIA within BART allows the appropriate handling of different types of missing mechanism, MCAR, MAR, and NMAR, for each covariate. We utilize this approach to accommodate the missingness in our covariates for the data analysis.

5.3 Dealing with Censoring by Death

5.3.1 Determining the principal strata

To determine the principal strata definition, we first investigated what the data for our problem could potentially look like. We constructed Table 5.1 for $t = 3$, $p = 1$, and no time-varying covariates without loss of generality. In this table, 'x' indicates an observed value, '?' represent a missing observation which needs to be imputed, and 'NA' indicates a structurally missing observation. For the potential survival outcomes, we did not indicate whether they were missing or observed because we wanted to use Table 5.1 to help us decide how we should be stratifying our subjects once our proposed method imputes the counterfactual survival status.

Table 5.1: Sample example of a censoring by death dataset until $t = 3$ where $Z_t = 1$ indicates a subject having experienced a negative wealth shock and $Z_t = 0$ indicates a subject have not experienced any negative wealth shock till time t

	V	Z_1	Y_1	Y_0	S_1	S_0	Z_2	Y_{00}	Y_{01}	Y_{11}	S_{00}	S_{01}	S_{11}	Z_3	Y_{000}	Y_{001}	Y_{011}	Y_{111}
Subject 1	x	1	x	?	1	1	1	?	?	x	1	1	1	1	?	?	?	x
Subject 2	x	0	?	x	1	1	1	?	x	?	1	1	1	1	?	?	x	?
Subject 3	x	1	x	?	1	1	1	?	?	x	1	1	0	NA	?	?	?	NA
Subject 4	x	0	?	x	1	1	1	?	x	?	1	1	0	1	?	?	x	NA
Subject 5	x	0	?	x	1	1	0	x	?	?	1	0	1	0	x	?	NA	?
Subject 6	x	0	?	x	1	1	0	x	?	?	0	1	1	NA	NA	NA	?	?
Subject 7	x	0	?	x	1	1	0	x	?	?	0	1	1	NA	NA	NA	?	?
Subject 8	x	0	?	x	1	1	0	x	?	?	1	0	0	0	x	?	NA	NA
Subject 9	x	1	x	?	0	1	NA	?	?	NA	1	1	0	NA	?	?	?	NA
Subject 10	x	1	x	?	0	1	NA	?	?	NA	0	1	0	NA	NA	NA	?	NA
Subject 11	x	0	?	x	0	1	1	?	x	NA	0	1	0	1	NA	NA	x	NA
Subject 12	x	0	?	x	0	1	0	x	?	NA	0	1	0	NA	NA	NA	?	NA
Subject 13	x	1	x	?	1	0	1	NA	NA	x	0	0	1	1	NA	NA	NA	x
Subject 14	x	0	?	x	1	0	NA	NA	NA	?	0	0	1	NA	NA	NA	NA	?

From Table 5.1, we can see that the goal of our analysis is to provide inference about $E[Y_{Z_1,\ldots,Z_t} - Y_{Z'_1,\ldots,Z'_t} | S_{Z_1,\ldots,Z_{t-1}} = S_{Z'_1,\ldots,Z'_{t-1}} = 1]$, where $Z_l \neq Z'_l$ for at least one l with $l = 1, \ldots, t$ i.e. we condition on subjects who would potentially survive under two different treatment regimes $Z_1, \ldots, Z_{t-1}$ and $Z'_1, \ldots, Z'_{t-1}$. Thus, the distribution of $(S_{Z_1,\ldots,Z_{t-1}}, S_{Z'_1,\ldots,Z'_{t-1}})$ form our principal strata and meaningful contrasts are defined only in the stratum where $S_{Z_1,\ldots,Z_{t-1}} = S_{Z'_1,\ldots,Z'_{t-1}} = 1$ since the potential outcomes for the two different treatment regimes exist only in this stratum. For example, if we want to estimate the effect for a negative wealth shock at $t = 2$ versus no negative wealth shock by $t = 2$ that is $E[Y_{01} - Y_{00} | S_0 = 1]$, we restrict to subjects who survive if they did not receive a negative wealth shock at $t = 1$ i.e. subjects with $S_0 = 1$ (Subjects 1-12 in Table 5.1). Note that the definition, $E[Y_{Z_1,\ldots,Z_t} - Y_{Z'_1,\ldots,Z'_t} | S_{Z_1,\ldots,Z_{t-1}} = S_{Z'_1,\ldots,Z'_{t-1}} = 1]$, is different from the parameter MSM estimates which is $E[Y_{Z_1,\ldots,Z_t} - Y_{Z'_1,\ldots,Z'_t}]$.

5.3.2 Proposed method

We make the same four assumptions used by MSM (See Section 5.2.2). Our proposed method estimates $E[Y_{Z_1,\ldots,Z_t} - Y_{Z'_1,\ldots,Z'_t} | S_{Z_1,\ldots,Z_{t-1}} = S_{Z'_1,\ldots,Z'_{t-1}} = 1]$ by imputing the survival status of each subject at the current time t and then combine the imputed counterfactual survival status together with the observed survival status to determine which principal stratum a subject belongs to. We then use a slightly modified PENCOMP to impute the counterfactual outcomes among the potentially surviving subjects to account for the bias due to confounding by indication. This approach is doubly robust and reduces the burden of model specification by the researcher. Subsequently, the average difference in the treatment effect within the desired principal strata is calculated. Variance is estimated using Rubin's combine rule to account for the imputation uncertainty (*Heitjan and Little*, 1991). Detailed steps for our method are given below.

1. Generate a bootstrap sample b from the data by sampling the units with replacement.

2. Estimate the model $X^{(b)}_{z_1^{(b)}}|Z_1^{(b)} = z_1^{(b)}, W_1^{(b)}, V^{(b)}$. Use this model to compute the counterfactual of $X^{(b)}_{z_1^{(b)}}$ for bootstrap sample b.

3. Estimate the distribution of $Z_1^{(b)}|W_1^{(b)}, V^{(b)}$. Use this model to estimate the propensity to be assigned treatment $Z_1^{(b)} = z_1^{(b)}$ as $P^*_{z_1^{(b)}} = Pr(Z_1^{(b)} = z_1^{(b)}|W_1^{(b)}, V^{(b)})$. Note that we did not perform a logit transformation to obtain $P^*_{z_1^{(b)}}$ (See PENCOMP Steps 2 and 5). This is because by using PENCOMP modified with BART to predict the outcomes, the non-linear effect of the propensity of assigned treatment will be handled automatically. Hence, any non-linear transformation on the propensity of assigned treatment would not be needed.

4. Estimate the model $Y^{(b)}_{z_1^{(b)}}|P^*_{z_1^{(b)}}, Z_1^{(b)} = z_1^{(b)}, X^{(b)}_{z_1^{(b)}}, W_1^{(b)}, V^{(b)}$. As mentioned, we used PENCOMP modified with BART to estimate this model. The advantage of using BART is the researcher no longer needs to specify the model. BART automatically takes care of any linear or non-linear main effects as well as linear or non-linear interactions. If we observe Equations 5.7 and 5.8, we can see that these two equations are constructed using a non-linear spline specification on the propensity of assigned treatment combined with possible linear interactions between the propensity of assigned treatment and remaining covariates. This fits well with the type of estimation problems that BART was designed to solve. We then use the model produced by BART-modified PENCOMP to compute the counterfactual of $Y^{(b)}_{z_1^{(b)}}$ for bootstrap sample b.

5. Estimate the distribution for $S^{(b)}_{z_1^{(b)}}|Z_1^{(b)} = z_1^{(b)}, Y^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)}}, W_1^{(b)}, V^{(b)}$ at $t = 2$. Use this model to generate a survival status for the counterfactual of $S^{(b)}_{z_1^b}$.

6. Estimate the model $X^{(b)}_{z_1^{(b)},z_2^{(b)}}|Z_1^{(b)} = z_1^{(b)}, Z_2^{(b)} = z_2^{(b)}, Y^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)}}, W_1^{(b)}, W_2^{(b)}, V^{(b)}$.

Use the respective models to impute the counterfactual of $X^{(b)}_{z_1^{(b)},z_2^{(b)}}$, using any previously imputed values for the unobserved treatment regimes and restricting to the subjects that are observed and predicted to survive under the given treatment regimen of interest at $t = 1$.

7. Estimate the distribution of $Z_2^{(b)}|Z_1^{(b)} = z_1^{(b)}, Y^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)}}, W_1^{(b)}, W_2^{(b)}, V^{(b)}$. Use this model to estimate the propensity to be assigned treatment $Z_2^{(b)} = z_2^{(b)}$ as $P_{z_2^{(b)}} = Pr(Z_1^{(b)} = z_1^{(b)}|X^{(b)}_{z_1^{(b)}}, Z_1^{(b)} = z_1^{(b)}, W_1^{(b)}, V^{(b)})$. The probability of treatment regimen $(Z_1^{(b)} = z_1^{(b)}, Z_2^{(b)} = z_2^{(b)})$ is denoted as $P^*_{z_2^{(b)}} = P_{z_2^{(b)}} P^*_{z_1^{(b)}}$.

8. Estimate the model

$$Y^{(b)}_{z_1^{(b)},z_2^{(b)}}|P^*_{z_2^{(b)}}, Z_1^{(b)} = z_1^{(b)}, Z_2^{(b)} = z_2^{(b)}, Y^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)},z_2^{(b)}}, W_1^{(b)}, W_2^{(b)}, V^{(b)}$$

again restricting to subjects that are observed and predicted to survive under the treatment regimes of interest at $t = 2$. Use the respective models to impute the counterfactual of $Y^{(b)}_{z_1^{(b)},z_2^{(b)}}$.

9. Using a similar procedure for steps 5-8 with the restriction determined by $S^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)}} = S^{(b)}_{z_1^{'(b)},\ldots,z_{t-1}^{'(b)}} = 1$ for time t where at least one $z_t^{(b)} \neq z_t^{'(b)}$ and extend the estimation until the desired time point $t = T$.

10. Repeat Steps 1-9 to obtain B bootstrap values for

$$\hat{\Delta}^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)},z_1^{'(b)},\ldots,z_{t-1}^{'(b)}} = E[Y^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)}} - Y^{(b)}_{z_1^{'(b)},\ldots,z_{t-1}^{'(b)}}|S^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)}} = S^{(b)}_{z_1^{'(b)},\ldots,z_{t-1}^{'(b)}} = 1].$$

with associated pooled variance $W^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)},z_1^{'(b)},\ldots,z_{t-1}^{'(b)}}$.

11. The estimate of

$$\Delta_{Z_1,\ldots,Z_t,Z'_1,\ldots,Z'_t} = E[Y_{Z_1,\ldots,Z_t} - Y_{Z'_1,\ldots,Z'_t}|S_{Z_1,\ldots,Z_{t-1}} = S_{Z'_1,\ldots,Z'_{t-1}} = 1]$$

is then

$$\bar{\Delta}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B} = \sum_{b=1}^{B}(\hat{\Delta}^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)},z_1'^{(b)},\ldots,z_{t-1}'^{(b)}})/B,$$

and the estimate of the variance of $\bar{\Delta}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B}$ is

$$T_B = \bar{W}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B} + (1+1/B)D_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B},$$

where

$$\bar{W}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B} = \sum_{b=1}^{B}(W^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)},z_1'^{(b)},\ldots,z_{t-1}'^{(b)}})/B$$

and

$$D_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B} = \sum_{b=1}^{B}\frac{(\hat{\Delta}^{(b)}_{z_1^{(b)},\ldots,z_{t-1}^{(b)},z_1'^{(b)},\ldots,z_{t-1}'^{(b)}} - \bar{\Delta}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B})^2}{B-1}.$$

The estimate $\Delta_{Z_1,\ldots,Z_t,Z'_1,\ldots,Z'_t}$ follows a t distribution with degree of freedom ν,

$$\frac{\Delta_{Z_1,\ldots,Z_t,Z'_1,\ldots,Z'_t} - \bar{\Delta}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B}}{\sqrt{T_B}} \sim t_\nu,$$

where $\nu = (B-1)(1 + \frac{\bar{W}_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B}}{D_{z_1,\ldots,z_t,z'_1,\ldots,z'_t,B}(B+1)})^2$.

Remark. The idea of including the BART estimated propensity score within BART as a predictor in Steps 4 and 8 is not new. *Hahn et al.* (2018) showed that including a BART estimated propensity score as a predictor within BART improved the estimation of heterogenous treatment effects for observational studies. *Tan et al.* (2018) also reported that the inclusion of the BART estimated propensity score as a predictor within BART to impute missing data, under the missing at random assumption, worked well in situations where the non-linear main and interaction effects are complex for the mean and propensity model. For situations with simpler non-linear effects like a quadratic relationship, using BART to estimate the propensity score and imputing the missing values using penalized splines of propensity prediction (*Zhang and*

Little, 2009, PENCOMP version for missing data) worked better. Using PENCOMP with a BART estimated propensity score for Steps 4 and 8 would be an interesting alternative. However, our aim of Steps 4 and 8 was to ease the implementation burden on the researcher. Hence, we suggest the use of PENCOMP with a BART estimated propensity score for Steps 4 and 8 only if the researcher is certain that the non-linear effect has a simple form for example, a quadratic or cubic relationship.

5.4 Simulation

We conducted a simulation study to determine how well our proposed method would perform compared to the naïve method and MSM in three scenarios: 1) where there is low association between treatment allocation and confounder as well as treatment and survival status; 2) where there is a strong association between treatment and confounder as well as treatment and survival status; and finally 3) where there is a strong association between treatment and confounder, treatment and survival status, and an interaction between treatment, confounder, and survival status. We expect all three methods to perform well in the first scenario because there is little to no confounding. For the second scenario, we expect MSM and our proposed method to perform well because there is no difference in the treatment effect between the principal strata, and other stratification groups. The naïve method should not perform well due to the strong association between treatment and confounder as well as treatment and survival status. Finally, for scenario three, we expect only our proposed method to perform well because an association between the treatment effect and principal strata, $S_{Z_1,\ldots,Z_{t-1}} = S_{Z'_1,\ldots,Z'_{t-1}} = 1$, is induced by the stronger interaction effect between treatment, confounder, and survival status. We fit standard linear and logistic regression models rather than BART and PENCOMP with BART since our focus is not on model misspecification but rather, the effect of confounding by indication and censoring by death.

5.4.1 Setup

To set up our simulation study, we set the size of our target population as 1 million. We then generate a single baseline variable V from a normal distribution. We set $T = 3$ and model our treatment allocation, Z_1, as

$$logit[P(Z_1 = 1|V)] = \gamma_0 + \gamma_1 V. \tag{5.12}$$

For the potential outcome at $t = 1$, Y_{Z_1}, we model it as

$$Y_{Z_1} = \beta_0 + \beta_Z I\{Z_1 = 1\} + \beta_V V + \beta_{VZ} V I\{Z_1 = 1\} + e, \tag{5.13}$$

where $e \sim N(0, 1)$.

We model the potential survival status at $t = 2$, S_{Z_1} as

$$\begin{aligned} logit(P[S_{Z_1} = 1|V, Y_{Z_1}]) = \alpha_0 + \alpha_{Y_1} Y_1 I\{Z_1 = 1\} + \alpha_{Y_0} Y_0 [1 - I\{Z_1 = 1\}] \\ + \alpha_Z I\{Z_1 = 1\} + \alpha_V V + \alpha_{VZ} V I\{Z_1 = 1\}. \end{aligned} \tag{5.14}$$

Because a negative wealth shock is an absorbing state, if $Z_1 = 1$, then $Z_2 = 1$. So when $Z_1 = 0$, we have

$$logit(P[Z_2 = 1|V, Y_0]) = \gamma_0 + \gamma_{Y_0,2} Y_0 + \gamma_2 V. \tag{5.15}$$

We model the potential outcome at $t = 2$, Y_{Z_1,Z_2} as

$$\begin{aligned} Y_{Z_1,Z_2} = \beta_0 &+ \beta_{Z_{01}} I\{Z_1 = 0, Z_2 = 1\} + \beta_{Z_{11}} I\{Z_1 = 1, Z_2 = 1\} \\ &+ \beta_{Y_0 Z_{00}} Y_0 I\{Z_1 = 0, Z_2 = 0\} + \beta_{Y_0 Z_{01}} Y_0 I\{Z_1 = 0, Z_2 = 1\} \\ &+ \beta_{Y_1 Z_{11}} Y_1 I\{Z_1 = 1, Z_2 = 1\} + \beta_V V + \beta_{V Z_{01}} V I\{Z_1 = 0, Z_2 = 1\} \\ &+ \beta_{V Z_{11}} V I\{Z_1 = 1, Z_2 = 1\} + e, \end{aligned} \tag{5.16}$$

where $e \sim N(0,1)$.

For the potential survival status at $t = 3$, S_{Z_1,Z_2}, if $S_{Z_1} = 0$, then $S_{Z_1,Z_2} = 0$. When $S_{Z_1} = 1$, we have

$$\begin{aligned} logit(P[S_{Z_1,Z_2} = 1|X, Y_{Z_1,Z_2}, S_{Z_1} = 1]) = {} & \alpha_0 + \alpha_{Z_{01}} I\{Z_1 = 0, Z_2 = 1\} \\ & + \alpha_{Z_{11}} I\{Z_1 = 1, Z_2 = 1\} \\ & + \alpha_{Y_{00}Z_{00}} Y_{00} I\{Z_1 = 0, Z_2 = 0\} \\ & + \alpha_{Y_{01}Z_{01}} Y_{01} I\{Z_1 = 0, Z_2 = 1\} \\ & + \alpha_{Y_{11}Z_{11}} Y_{11} I\{Z_1 = 1, Z_2 = 1\} \\ & + \alpha_V V + \alpha_{VZ_{01}} V I\{Z_1 = 0, Z_2 = 1\} \\ & + \alpha_{VZ_{11}} V I\{Z_1 = 1, Z_2 = 1\}. \end{aligned} \tag{5.17}$$

For the treatment allocation at $t = 3$, Z_3, if $Z_1 = Z_2 = 0$, we have

$$logit(P[Z_3 = 1|X, Y_{00}]) = \gamma_0 + \gamma_{Y_{00}} Y_{00} + \gamma_{Y_{0,3}} Y_0 + \gamma_3 V. \tag{5.18}$$

For the potential outcome at $t = 3$, Y_{Z_1,Z_2,Z_3}, we have

$$\begin{aligned} Y_{Z_1,Z_2,Z_3} = {} & \beta_0 + \beta_{Z_{001}} I\{Z_1 = 0, Z_2 = 0, Z_3 = 1\} + \beta_{Z_{011}} I\{Z_1 = 0, Z_2 = 1, Z_3 = 1\} \\ & + \beta_{Z_{111}} I\{Z_1 = 1, Z_2 = 1, Z_3 = 1\} + \beta_{Y_{00}Z_{000}} Y_{00} I\{Z_1 = 0, Z_2 = 0, Z_3 = 0\} \\ & + \beta_{Y_{00}Z_{001}} Y_{00} I\{Z_1 = 0, Z_2 = 0, Z_3 = 1\} \\ & + \beta_{Y_{01}Z_{011}} Y_{01} I\{Z_1 = 0, Z_2 = 1, Z_3 = 1\} \\ & + \beta_{Y_{11}Z_{111}} Y_{11} I\{Z_1 = 1, Z_2 = 1, Z_3 = 1\} + \beta_{Y_0 Z_0} Y_0 I\{Z_1 = 0\} \\ & + \beta_{Y_1 Z_1} Y_1 I\{Z_1 = 1\} + \beta_V V + \beta_{VZ_{001}} V I\{Z_1 = 0, Z_2 = 0, Z_3 = 1\} \\ & + \beta_{VZ_{011}} V I\{Z_1 = 0, Z_2 = 1, Z_3 = 1\} \\ & + \beta_{VZ_{111}} V I\{Z_1 = 1, Z_2 = 1, Z_3 = 1\} + e. \end{aligned} \tag{5.19}$$

Table 5.2 shows the parameters we used to achieve the three different simulation scenarios. Scenario 1 is achieved by setting γ_1, α_Z, γ_2, $\gamma_{Y_0,2}$, $\alpha_{Z_{01}}$, $\alpha_{Z_{11}}$, γ_3, $\gamma_{Y_0,3}$, and $\gamma_{Y_{00}}$ to be about 10 times smaller than the values in Scenarios 2 and 3. The rest of the differences between Scenario 1 versus 2 and 3 were to ensure the resulting simulated population would have enough deaths and subjects in the various different treatment regimes for the assumptions used by MSM and our proposed method to be valid. The difference between Scenario 2 versus 3 lie in β_{VZ}, α_{Y_1}, α_{Y_0}, $\beta_{Y_0Z_{00}}$, $\beta_{Y_0Z_{01}}$, $\beta_{Y_1Z_{11}}$, $\alpha_{Y_0Z_{00}}$, $\alpha_{Y_0Z_{01}}$, $\alpha_{Y_1Z_{11}}$, $\beta_{Y_{00}Z_{000}}$, $\beta_{Y_{00}Z_{001}}$, $\beta_{Y_{01}Z_{011}}$, and $\beta_{Y_{11}Z_{111}}$ where the values for Scenario 2 is about 10 times smaller compared to Scenario 3.

To calculate the true parameters, we used the generated population data (size 1 million), and then took:

1. $\Delta_{1,0} = \bar{Y}_1 - \bar{Y}_0$;
2. $\Delta_{01,00} = \bar{Y}_{01} - \bar{Y}_{00}$ given $S_0 = 1$;
3. $\Delta_{11,00} = \bar{Y}_{11} - \bar{Y}_{00}$ given $S_0 = S_1 = 1$;
4. $\Delta_{11,01} = \bar{Y}_{11} - \bar{Y}_{01}$ given $S_0 = S_1 = 1$;
5. $\Delta_{001,000} = \bar{Y}_{001} - \bar{Y}_{000}$ given $S_{00} = 1$;
6. $\Delta_{011,000} = \bar{Y}_{011} - \bar{Y}_{000}$ given $S_{00} = S_{01} = 1$;
7. $\Delta_{111,000} = \bar{Y}_{111} - \bar{Y}_{000}$ given $S_{00} = S_{11} = 1$;
8. $\Delta_{011,001} = \bar{Y}_{011} - \bar{Y}_{001}$ given $S_{00} = S_{01} = 1$;
9. $\Delta_{111,001} = \bar{Y}_{111} - \bar{Y}_{001}$ given $S_{00} = S_{11} = 1$; and
10. $\Delta_{111,011} = \bar{Y}_{111} - \bar{Y}_{011}$ given $S_{01} = S_{11} = 1$.

We measured performance using the empirical bias, root mean squared error (RMSE), 95% coverage, and the average 95% Confidence Interval (CI) length (AIL).

Table 5.2: Table of parameters for simulation

	Scenario 1	Scenario 2	Scenario 3
V	$N(0, 2^2)$	$N(17, 2^2)$	$N(17, 2^2)$
γ_0	0	2	2
γ_1	-0.02	-0.2	-0.2
β_0	0	5.3	5.3
β_Z	-1.5	-1.5	-1.5
β_V	0.015	0.15	0.2
β_{VZ}	-0.005	-0.11	-0.05
α_0	0	1	0
α_{Y_1}	0.005	0.00625	0.0625
α_{Y_0}	0.01	0.0125	0.125
α_Z	-0.01	-0.2	-0.2
α_V	0.002	0.02	0.02
α_{VZ}	-0.002	-0.02	-0.02
γ_2	-0.002	-0.02	-0.02
$\gamma_{Y_0,2}$	-0.02	-0.2	-0.2
$\beta_{Z_{01}}$	-1.5	-1.5	-1.5
$\beta_{Z_{11}}$	-1	-1	-1
$\beta_{Y_0 Z_{00}}$	0.015	0.02	0.3
$\beta_{Y_0 Z_{01}}$	0.01	0.015	0.2
$\beta_{Y_1 Z_{11}}$	0.005	0.01	0.1
$\beta_{VZ_{01}}$	-0.00011	-0.011	-0.011
$\beta_{VZ_{11}}$	-0.00005	-0.005	-0.005
$\alpha_{Z_{01}}$	-0.01	-0.2	-0.2
$\alpha_{Z_{11}}$	-0.015	-0.1	-0.1
$\alpha_{Y_0 Z_{00}}$	0.01	0.0125	0.125
$\alpha_{Y_0 Z_{01}}$	0.005	0.00625	0.0625
$\alpha_{Y_1 Z_{11}}$	0.0025	0.003125	0.03125
$\alpha_{VZ_{01}}$	-0.0001	-0.02	-0.02
$\alpha_{VZ_{11}}$	-0.0005	-0.05	-0.05
γ_3	-0.0002	-0.002	-0.002
$\gamma_{Y_0,3}$	-0.002	-0.02	-0.02
$\gamma_{Y_{00}}$	-0.02	-0.2	-0.2
$\beta_{Z_{001}}$	-1.5	-1.5	-1.5
$\beta_{Z_{011}}$	-1	-1	-1
$\beta_{Z_{111}}$	-0.5	-0.5	-0.5
$\beta_{Y_{00} Z_{000}}$	0.015	0.02	0.3
$\beta_{Y_{00} Z_{001}}$	0.01	0.015	0.2
$\beta_{Y_{01} Z_{011}}$	0.005	0.01	0.1
$\beta_{Y_{11} Z_{111}}$	0.0025	0.005	0.05
$\beta_{Y_0 Z_0}$	0.0008	0.08	0.08
$\beta_{Y_1 Z_1}$	0.0003	0.03	0.03
$\beta_{VZ_{001}}$	-0.00011	-0.011	-0.011
$\beta_{VZ_{011}}$	-0.00005	-0.005	-0.005
$\beta_{VZ_{111}}$	-0.00003	-0.003	-0.003

1000 simulations were used to estimate these quantities. Under each simulation, a simple random sample of 4,000 or 8,000 subjects was drawn from the target population data. All methods were then implemented on the sampled data to obtain the effect estimates. For MSM and our proposed method, the models were specified using Equations 5.12 to 5.19 respectively. For our proposed method, because our focus is not on model misspecification but rather, confounding by indication and censoring by death, we chose to implement a simpler version of our method by skipping Steps 3 and 7 of our algorithm and using $Y^{(b)}_{z_1^{(b)}}|Z_1^{(b)} = z_1^{(b)}, X^{(b)}_{z_1^{(b)}}, W_1^{(b)}, V^{(b)}$ and $Y^{(b)}_{z_1^{(b)},z_2^{(b)}}|Z_1^{(b)} = z_1^{(b)}, Z_2^{(b)} = z_2^{(b)}, Y^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)}}, X^{(b)}_{z_1^{(b)},z_2^{(b)}}, W_1^{(b)}, W_2^{(b)}, V^{(b)}$ for Steps 4 and 8 respectively. We also simplified the prediction of the potential outcomes and survival status by using linear and logistic regression instead of BART.

5.4.2 Results

Table 5.3 shows the simulation results for sample size of 4,000. As expected, under Scenario 1, all three methods were relatively unbiased with all three methods achieving similar RMSE. MSM and our proposed method reported slightly greater than nominal coverage due to the wider AIL. Under Scenario 2, the absolute bias of the naïve method was always larger than MSM and our proposed method. RMSE was larger as well in comparison and coverage was often far below the nominal 95% value. For this scenario MSM produced the less conservative coverage while our proposed method suggested better bias performance and reduced RMSE. Finally, under Scenario 3, the naïve method was clearly biased with poor RMSE and coverage. MSM performed slightly better compared to the naïve method but absolute bias clearly increased compared to Scenario 2. Coverage for some treatment effects were poor as well. Our proposed method remained unbiased, produced a lower RMSE compared to the other two methods, and reached nominal coverage under Scenario 3. All methods behaved as expected under these three scenarios.

Table 5.4 shows the results with the sample size increased to 8,000, approximately the sample size in our application. The simulation results for all three methods under Scenario 1 remained relatively similar. Under Scenario 2, an increase in sample size did not affect the absolute bias of all three methods but, the coverage of the naïve method was clearly affected with huge decreases in the coverage for all parameters. Coverage for MSM and our proposed method remained fairly similar. Finally, under Scenario 3, we observe once again that the amount of bias for the three methods remained the same but, coverage for the naïve method and MSM decreased for most of the treatment effects when the sample size increased to 8,000. Coverage for our proposed method remained relatively similar to the results observed for the sample size of 4,000. In summary, bias for the three methods was rather stable when the sample size changed. However, if the method is poor in the estimation of the particular treatment effect, increasing the sample size can cause large decreases in coverage.

Table 5.3: Simulation results for sample size 4,000

Scenario 1		Naïve				MSM				Proposed			
Parameter	True value	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL
$\Delta_{1,0}$	-1.497	-0.001	0.032	95.4	0.123	-0.0002	0.032	95.1	0.123	-0.0003	0.032	96.8	0.143
$\Delta_{01,00}$	-1.499	-0.004	0.062	95.0	0.247	-0.004	0.062	95.2	0.247	-0.003	0.063	96.3	0.264
$\Delta_{11,00}$	-1.005	-0.005	0.055	95.3	0.214	-0.004	0.055	95.2	0.214	-0.003	0.056	99.3	0.334
$\Delta_{11,01}$	0.493	0.001	0.055	93.8	0.214	0.002	0.055	94.3	0.214	0.001	0.057	99.7	0.333
$\Delta_{001,000}$	-1.502	-0.013	0.124	94.6	0.494	-0.013	0.124	99.2	0.713	-0.013	0.126	96.5	0.531
$\Delta_{011,000}$	-1.006	-0.007	0.110	94.2	0.428	-0.007	0.111	99.2	0.616	-0.006	0.113	99.4	0.669
$\Delta_{111,000}$	-0.494	-0.014	0.101	94.0	0.392	-0.013	0.102	99.2	0.566	-0.012	0.105	100.0	0.893
$\Delta_{011,001}$	0.498	0.004	0.111	94.2	0.428	0.004	0.112	99.5	0.614	0.005	0.115	99.8	0.662
$\Delta_{011,001}$	1.002	0.004	0.100	94.7	0.390	0.005	0.100	99.8	0.562	0.006	0.104	100.0	0.885
$\Delta_{111,011}$	0.511	-0.007	0.080	94.6	0.304	-0.005	0.080	99.5	0.435	-0.005	0.083	100.0	0.695
Scenario 2		Naïve				MSM				Proposed			
Parameter	True value	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL
$\Delta_{1,0}$	-3.367	-0.047	0.061	78.5	0.154	0.002	0.041	93.8	0.160	0.002	0.041	96.0	0.177
$\Delta_{01,00}$	-1.727	-0.039	0.056	83.9	0.161	-0.034	0.053	87.6	0.162	-0.003	0.040	97.3	0.172
$\Delta_{11,00}$	-1.201	-0.137	0.147	25.5	0.206	-0.019	0.058	92.8	0.209	-0.002	0.054	97.3	0.252
$\Delta_{11,01}$	0.527	-0.098	0.112	50.9	0.203	0.015	0.056	94.4	0.205	0.001	0.054	97.5	0.248
$\Delta_{001,000}$	-1.728	-0.027	0.072	93.4	0.259	-0.020	0.069	97.5	0.299	0.002	0.064	96.8	0.266
$\Delta_{011,000}$	-1.184	-0.060	0.085	82.7	0.233	-0.042	0.073	94.3	0.269	0.005	0.058	97.7	0.281
$\Delta_{111,000}$	-1.168	-0.159	0.176	41.9	0.291	-0.029	0.081	96.8	0.343	0.010	0.074	99.3	0.418
$\Delta_{011,001}$	0.545	-0.034	0.066	89.9	0.218	-0.022	0.061	95.7	0.251	0.002	0.054	98.7	0.264
$\Delta_{011,001}$	0.558	-0.130	0.149	54.8	0.280	-0.006	0.075	97.4	0.329	0.009	0.074	99.3	0.397
$\Delta_{111,011}$	0.016	-0.099	0.118	66.8	0.256	0.013	0.069	97.4	0.302	0.006	0.067	99.4	0.379
Scenario 3		Naïve				MSM				Proposed			
Parameter	True value	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL
$\Delta_{1,0}$	-2.347	-0.123	0.130	14.5	0.160	0.002	0.042	94.0	0.160	0.002	0.041	96.5	0.176
$\Delta_{01,00}$	-2.561	-0.115	0.123	26.8	0.176	-0.060	0.074	73.0	0.175	-0.002	0.040	97.5	0.174
$\Delta_{11,00}$	-3.073	-0.230	0.238	4.0	0.236	-0.023	0.065	92.5	0.230	0.0009	0.059	98.0	0.285
$\Delta_{11,01}$	-0.508	-0.118	0.132	50.6	0.236	0.034	0.069	91.1	0.231	0.0006	0.060	97.9	0.286
$\Delta_{001,000}$	-2.822	-0.128	0.144	53.4	0.264	-0.065	0.092	91.1	0.313	-0.002	0.058	96.2	0.245
$\Delta_{011,000}$	-3.611	-0.142	0.153	28.1	0.217	-0.086	0.103	77.4	0.260	-0.005	0.052	96.9	0.228
$\Delta_{111,000}$	-4.046	-0.285	0.297	5.7	0.326	-0.069	0.110	93.1	0.405	0.003	0.081	99.0	0.450
$\Delta_{011,001}$	-0.787	-0.017	0.066	93.1	0.246	-0.023	0.067	96.3	0.294	-0.005	0.058	96.2	0.258
$\Delta_{011,001}$	-1.224	-0.158	0.180	58.3	0.347	-0.005	0.088	98.3	0.429	-0.0003	0.083	99.5	0.494
$\Delta_{111,011}$	-0.445	-0.133	0.156	62.6	0.313	0.026	0.086	97.6	0.390	0.009	0.078	99.6	0.445

Table 5.4: Simulation results for sample size 8,000

Scenario 1		Naïve				MSM				Proposed			
Parameter	True value	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL
$\Delta_{1,0}$	-1.497	-0.0007	0.023	94.2	0.087	0.0003	0.023	94.0	0.087	0.0003	0.023	96.6	0.100
$\Delta_{01,00}$	-1.499	-0.004	0.044	95.9	0.174	-0.003	0.044	95.9	0.174	-0.003	0.045	96.0	0.186
$\Delta_{11,00}$	-1.005	-0.005	0.039	94.5	0.151	-0.004	0.039	94.2	0.151	-0.003	0.039	99.5	0.236
$\Delta_{11,01}$	0.493	0.00004	0.039	95.0	0.152	0.001	0.039	95.0	0.152	0.0007	0.040	99.5	0.235
$\Delta_{001,000}$	-1.502	-0.010	0.087	95.0	0.349	-0.010	0.087	99.6	0.499	-0.010	0.088	96.2	0.374
$\Delta_{011,000}$	-1.006	-0.004	0.076	95.3	0.303	-0.004	0.076	99.6	0.433	-0.004	0.077	99.6	0.466
$\Delta_{111,000}$	-0.494	-0.012	0.071	94.4	0.277	-0.011	0.071	99.4	0.396	-0.009	0.073	100.0	0.617
$\Delta_{011,001}$	0.498	0.004	0.076	94.8	0.302	0.004	0.076	99.7	0.432	0.005	0.078	99.7	0.466
$\Delta_{011,001}$	1.002	0.004	0.068	96.2	0.276	0.005	0.068	99.7	0.394	0.006	0.071	100.0	0.613
$\Delta_{111,011}$	0.511	-0.007	0.056	94.4	0.215	-0.006	0.056	99.3	0.306	-0.006	0.058	100.0	0.489
Scenario 2		Naïve				MSM				Proposed			
Parameter	True value	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL
$\Delta_{1,0}$	-3.367	-0.047	0.055	59.6	0.109	0.002	0.029	94.0	0.113	0.003	0.029	96.3	0.125
$\Delta_{01,00}$	-1.727	-0.037	0.047	73.9	0.114	-0.032	0.043	79.9	0.115	-0.002	0.029	96.7	0.123
$\Delta_{11,00}$	-1.201	-0.135	0.140	4.1	0.146	-0.017	0.042	93.2	0.147	-0.0008	0.037	98.2	0.177
$\Delta_{11,01}$	0.527	-0.098	0.105	22.4	0.144	0.015	0.040	93.6	0.145	0.001	0.037	97.8	0.174
$\Delta_{001,000}$	-1.728	-0.026	0.053	90.1	0.183	-0.019	0.051	96.0	0.211	0.004	0.046	95.9	0.189
$\Delta_{011,000}$	-1.184	-0.060	0.074	69.0	0.165	-0.041	0.060	87.9	0.189	0.006	0.042	97.9	0.198
$\Delta_{111,000}$	-1.168	-0.160	0.169	14.1	0.206	-0.029	0.061	94.7	0.241	0.010	0.054	98.9	0.296
$\Delta_{011,001}$	0.545	-0.035	0.053	85.6	0.154	-0.023	0.046	94.8	0.177	0.001	0.039	98.1	0.185
$\Delta_{011,001}$	0.558	-0.132	0.141	26.2	0.198	-0.008	0.053	96.7	0.232	0.007	0.052	99.1	0.282
$\Delta_{111,011}$	0.016	-0.099	0.110	41.9	0.181	0.013	0.051	96.1	0.212	0.005	0.048	99.1	0.266
Scenario 3		Naïve				MSM				Proposed			
Parameter	True value	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL	Bias	RMSE	95% Coverage	AIL
$\Delta_{1,0}$	-2.347	-0.123	0.126	1.5	0.133	0.002	0.029	94.5	0.113	0.003	0.029	95.8	0.124
$\Delta_{01,00}$	-2.561	-0.115	0.119	4.8	0.124	-0.059	0.067	55.2	0.123	-0.002	0.030	95.7	0.123
$\Delta_{11,00}$	-3.073	-0.229	0.233	0.1	0.167	-0.022	0.047	91.6	0.162	0.002	0.041	98.1	0.201
$\Delta_{11,01}$	-0.508	-0.117	0.125	21.0	0.167	0.034	0.053	88.6	0.163	0.001	0.041	98.1	0.201
$\Delta_{001,000}$	-2.822	-0.129	0.138	21.8	0.186	-0.066	0.081	82.9	0.220	-0.003	0.041	95.1	0.173
$\Delta_{011,000}$	-3.611	-0.141	0.147	6.3	0.154	-0.084	0.094	55.2	0.183	-0.004	0.038	96.5	0.162
$\Delta_{111,000}$	-4.046	-0.288	0.294	0.2	0.230	-0.071	0.094	86.6	0.285	-0.0003	0.057	99.0	0.316
$\Delta_{011,001}$	-0.787	-0.015	0.047	94.0	0.174	-0.021	0.049	96.7	0.207	-0.003	0.041	97.2	0.182
$\Delta_{011,001}$	-1.224	-0.159	0.171	28.1	0.245	-0.005	0.064	97.6	0.301	-0.002	0.060	99.6	0.345
$\Delta_{111,011}$	-0.445	-0.137	0.148	33.3	0.221	0.023	0.063	96.8	0.274	0.006	0.055	99.3	0.314

5.5 Determining the effect of a negative wealth shock on cognitive score for Health and Retirement Study subjects

5.5.1 Health and Retirement Study

To investigate the association between negative wealth shock and cognitive ability in late middle aged US adults, we used data from the Health and Retirement Study (HRS). HRS is a longitudinal study of US adults, enrolled at age 50 and older. These individuals have been surveyed biennially since 1992 with detailed modules on financial status and health (*Sonnega et al.*, 2014).

We use HRS data collected from 1996 to 2002 for our analysis. Subjects were obtained from the original HRS cohort, born in the years 1931-1941. Although data collection began in 1992, consistent collection of a subject's cognitive ability only began in 1996. Hence, we excluded the data collected before 1996 and treated the variables collected in 1996 as the baseline for our analysis. We excluded subjects who did not have longitudinal measurements for net worth because we were unable to distinguish whether they have already experienced a negative wealth shock. Subjects with zero or negative net worth at baseline were excluded since we did not know if these subjects have lifelong asset poverty or experienced a negative wealth shock prior to study entry. We also removed subjects who experienced a negative wealth shock and death between 1992 to 1996. These subjects were removed because they were no longer at risk for a negative wealth shock or death. There were 9,750 participants in the original HRS cohort, and of these, 7,106 participants (72.9%) were eligible for this analysis. These participants consists of a representative sample of the 1996 US population aged 55 to 65 who had not experienced a negative wealth shock in the previous five years.

5.5.1.1 Determining negative wealth shock

To determine whether a subject experienced a negative wealth shock from the previous follow-up period to the current follow-up period, we first obtained data from the module assessing net worth administered at every wave of HRS. Measured assets include housing value, net value of businesses, individual retirement accounts, checking/savings accounts, certificates of deposits and savings bonds, investment holdings, net value of vehicles, and the value of any other substantial assets. From this asset total, debts were subtracted, including home mortgages, other home equity loans, and unsecured debt values, like credit card balances, student loans, and medical debts. Missing values for wealth were imputed at the level of each asset or debt, using an unfolding bracket imputation method (*Juster and Smith*, 1997). Wealth data were not imputed for those who do not participate in a given wave. Negative wealth shock was measured and then dichotomized (yes or no) for each time point. Loss of 75% or more of total wealth between two consecutive waves was used as the cut-point for negative wealth shock (*Pool et al.*, 2018). Subjects were considered at risk for negative wealth shock until they have experienced a negative wealth shock or reached age 65.

5.5.1.2 Cognitive ability

The cognitive ability of a subject is assessed in HRS using the Telephone Interview for Cognitive Status (TICS). Unfortunately, the full HRS cognitive battery is not available for participants under 65. Hence, we used an abbreviated measure that included questions about episodic memory (Immediate Word recall [10 points] and Delayed Word recall [10 points]) and mental status (Serial 7's [5 points], backwards counting from 20 [2 points]) (*Crimmins et al.*, 2011). All responses were combined to create a composite score ranging from 0 to 27, with a higher score indicating higher cognitive ability. Some of these measures may be imputed implying that the cognitive summary score may include one or more imputed scores (*Fisher, G.G. and Hassan, H. and Faul, J.D. and Rodgers, W.L. and Weir, D.R.*, 2018). We treated this measure as continuous and normally distributed.

5.5.1.3 Descriptive statistics at baseline

Tables 5.5 to 5.6 show the descriptive statistics of the subjects at baseline by whether or not they experienced a negative wealth shock over the next six years regardless of survival status. At baseline, aside from whether the subject eventually survived until 2002 and health conditions like whether the subject ever had heart problems, high blood pressure, and stroke, all the other variables in Tables 5.5 to 5.6 were significantly associated with experiencing a negative wealth shock. A typical subject who would eventually experience a wealth shock would have a lower cognitive score at baseline; slightly higher BMI; lower opinion about his or her health; lower word recall score; likely still smoking; not insured; have depression; slightly lower income; either working, unemployed, or disabled; divorced or never married; lower wealth rank; have diabetes and/or psychological problems; younger; lesser years of education; and likely non-White.

Table 5.7 shows the change in unadjusted mean cognitive score between consecutive waves for subjects who did not receive a wealth shock versus those who ever received a negative wealth shock. Follow-up surveys occurred at years 2, 4, and 6. We can see that for a subject who ever got shocked, the largest observed decline in cognitive score occurs from Baseline to Wave 1. Subsequently, the decline in cognitive score is no longer as large between waves. Similarly, the bulk of our subjects were shocked at Wave 1 (second year of follow up). In later waves, the proportion of new subjects who received a negative wealth shock decreases.

Table 5.5: Descriptive statistics of 1996 Health and Retirement Study (baseline), part 1

Variables	No wealth shock Mean/Frequency (S.E./%)	Ever wealth shock Mean/Frequency (S.E./%)	*p*-value
Eventually survived?:			0.57
Yes	6,207 (94.7)	516 (94.0)	
No	350 (5.3)	33 (6.0)	
Cognitive score	17.07 (4.07)	16.26 (4.35)	< 0.01
BMI	27.21 (4.84)	27.73 (5.40)	0.03
Self-reported health			< 0.01
Excellent	1,207 (19.9)	83 (15.7)	
Very Good	2,126 (35.0)	128 (24.3)	
Good	1,715 (28.2)	163 (30.9)	
Fair	763 (12.6)	103 (19.5)	
Poor	261 (4.3)	50 (9.5)	
Current Smoking status:			< 0.01
Never	2,353 (40.0)	166 (32.4)	
Former	2,410 (41.0)	187 (36.5)	
Current	1,116 (19.0)	159 (31.1)	
Alcohol consumption:			< 0.01
Never	3,799 (62.9)	347 (66.1)	
Moderate	1,686 (27.9)	116 (22.1)	
Heavy	555 (9.2)	62 (11.8)	
Insured?:			< 0.01
No	1,014 (15.5)	120 (21.9)	
Yes	5,543 (84.5)	429 (78.1)	
Depression?:			< 0.01
No	4,922 (85.5)	361 (73.1)	
Yes	832 (14.5)	133 (26.9)	
Income (log transformed)	10.48 (1.21)	10.18 (1.45)	< 0.01
Labor force status:			< 0.01
Working	3,111 (51.2)	314 (59.6)	
Unemployed	96 (1.6)	13 (2.5)	
Retired	2,178 (35.9)	104 (19.7)	
Disabled	143 (2.4)	43 (8.2)	
Not in labor force	547 (9.0)	53 (10.1)	
Martial status:			< 0.01
Married	4,897 (80.8)	373 (70.8)	
Divorced	591 (9.7)	90 (17.1)	
Widowed	426 (7.0)	42 (8.0)	
Never Married	149 (2.5)	22 (4.2)	
Wealth rank in tertiles:			< 0.01
0	1,728 (26.4)	326 (59.4)	
1	2,360 (36.0)	124 (22.6)	
2	2,469 (37.7)	99 (18.0)	
Gender:			0.08
Male	3,113 (47.5)	239 (43.5)	
Female	3,444 (52.5)	310 (56.5)	

Table 5.6: Descriptive statistics of 1996 Health and Retirement Study (baseline), part 2

Variables	No wealth shock Mean/Frequency (S.E./%)	Ever wealth shock Mean/Frequency (S.E./%)	p-value
Ever had diabetes?:			< 0.01
No	5,474 (90.2)	451 (85.6)	
Yes	596 (9.8)	76 (14.4)	
Ever had heart problems?:			0.43
No	5,343 (88.0)	457 (86.7)	
Yes	730 (12.0)	70 (13.3)	
Ever had HBP?:			0.07
No	3,888 (64.0)	316 (60.0)	
Yes	2,183 (36.0)	211 (40.0)	
Ever had psych problems?:			< 0.01
No	5,691 (93.7)	469 (89.2)	
Yes	380 (6.3)	57 (10.8)	
Ever had stroke?:			0.1
No	5,912 (97.3)	506 (96.0)	
Yes	161 (2.7)	21 (4.0)	
Age	59.73 (3.19)	57.26 (2.18)	< 0.01
Number of education years centered	0.52 (2.93)	-0.17 (3.32)	< 0.01
Race:			< 0.01
Non-hispanic White	5,236 (79.9)	342 (62.3)	
Non-hispanic Black	759 (11.6)	120 (21.9)	
Hispanic	449 (6.8)	70 (12.8)	
Other	113 (1.7)	17 (3.1)	

Table 5.7: Change in unadjusted cognitive score between consecutive waves stratified by negative wealth shock status

	Never shocked	Ever shocked	Change in proportion shocked
Baseline to Wave 1	0.19	-1.61	3.5%
Wave 1 to Wave 2	-0.55	0.06	2.1%
Wave 2 to Wave 3	-0.05	-0.10	1.3%

5.5.2 Analysis

We were interested in how a negative wealth shock would affect the cognitive ability of late middle aged adults in the HRS during the six years of follow-up as well as how the duration of a negative wealth shock affects cognitive ability accounting for missingness in the cognitive outcome as well as censoring by death. We employed four different methods to estimate this effect and make inference. The four methods were the naïve method, where all subjects who died under their observed negative wealth shock status were removed from analysis; baseline adjusted method, where similar to the naïve method, all subjects who died were removed from analysis but the mean cognitive score was adjusted using a model that included all baseline covariates; MSM, where negative wealth shock allocation, missingness, and censoring by death were accounted for by inverse probability weighting; and our proposed method including the PENCOMP modification described in Subsection 5.3.2. We assumed that depression was the time-varying covariate that depends on the negative wealth shock status ($X_{Z_1,\ldots,Z_t}$ in Section 5.2) and the rest of the time-varying covariates are: self-reported health status, whether subject was insured, labor force status of subject, income, level of alcohol consumption, current smoking status, and number of health conditions (W_t in Section 5.3). We also assumed that the cognitive score is missing at random given the baseline variables presented in Tables 5.5 to 5.6, past negative wealth shock status, time-varying covariates, and cognitive score. For MSM, we accounted for this missingness by modeling the propensity of response while for our proposed method, we imputed the missing cognitive score by using the modified version of PENCOMP discussed in Subsection 5.3.2. All our models (baseline adjusted, MSM, and our proposed method) were specified using BART. For the naïve, baseline adjusted, and MSM method, we employed 1,000 bootstrap samples to calculate the mean and the 95% Confidence Interval (CI). The 95% CI was determined by taking the 2.5 and 97.5 percentile. For our proposed method, we estimated the effect and accounted for our uncertainty using our algorithm described in Subsection 5.3.2.

5.5.3 Results

Table 5.8 shows the adjusted effect estimate of a negative wealth shock on cognitive score depending on the duration of the shock for late middle aged adults in the original HRS cohort from 1996 to 2002. In general, the naïve and baseline adjusted method suggests that experiencing a negative wealth shock has a much larger negative effect on the cognitive score of subjects in our sample compared to the adjusted estimates reported by MSM and our proposed method. The naïve and baseline adjusted method produced very similar results suggesting low association between cognitive score and the baseline covariates. The effect for subjects who experienced a negative wealth shock within the first 2 years of follow up versus no shock (6 years vs. no shock), subjects who experienced a negative wealth shock within the first 2 years of follow up versus subjects who experienced a negative wealth shock between the second and fourth year of follow up (6 years vs. 2 years), and subjects who experienced a negative wealth shock within the first 2 years of follow up versus subjects who experienced a negative wealth shock between the fourth and sixth year of follow up (6 years vs. no shock), were significantly larger than 0 under the naïve and baseline adjusted method. For MSM and our proposed method all effects were reported to be not significant.

Table 5.8: Effect estimate of negative wealth shock on cognitive score for late middle aged adults in original Health Retirment Study cohort from 1996 to 2002.

	Naïve		Baseline adjusted†		MSM*		Proposed*	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
2 years vs. no shock	-0.51	(-1.45, 0.35)	-0.51	(-1.37, 0.3)	-0.01	(-1.18, 1.07)	-0.16	(-0.84, 0.52)
4 years vs. no shock	-0.69	(-1.45, 0.05)	-0.7	(-1.4, 0.03)	-0.31	(-1.23, 0.58)	0.21	(-0.66, 1.08)
6 years vs. no shock	-1.95	(-2.62, -1.25)	-1.94	(-2.6, -1.26)	-0.12	(-1.12, 0.89)	-0.28	(-1.23, 0.67)
4 years vs. 2 years	-0.18	(-1.33, 1.04)	-0.19	(-1.26, 0.94)	-0.3	(-1.78, 1.15)	0.34	(-0.65, 1.32)
6 years vs. 2 years	-1.45	(-2.54, -0.38)	-1.43	(-2.46, -0.4)	-0.1	(-1.61, 1.36)	-0.14	(-1.02, 0.75)
6 years vs. 4 years	-1.26	(-2.27, -0.2)	-1.24	(-2.2, -0.24)	0.19	(-1.11, 1.61)	-0.45	(-1.34, 0.44)

*Adjusted by gender, education category, race, cognitive score, BMI, self-reported health status, alcohol consumption, insurance status, depression status, income, labor force status, marital status, age, smoking status, diabetes status, heart condition, HBP status, psychological problem status, and stroke status at baseline as well as time-varying self-reported health status, alcohol consumption, insurance status, income, labor force status, smoking status, number of health conditions, and depression.

†Adjusted by gender, education category, race, cognitive score, BMI, self-reported health status, alcohol consumption, insurance status, depression status, income, labor force status, marital status, age, smoking status, diabetes status, heart condition, HBP status, psychological problem status, and stroke status at baseline.

5.6 Discussion

In this paper, we were interested in how a negative wealth shock affects the cognitive ability of late middle aged Americans participating in the HRS from 1996 to 2002. The main difficulty we faced was the presence of death in some subjects causing their cognitive score to be censored. Under situations where we believe death does not depend on the cognitive ability or whether a subject received a negative wealth shock, removing subjects who have died from our analysis would yield an unbiased estimate of the effect of negative wealth shock on cognitive ability as our simulation results suggest. Unfortunately, it is very possible that subjects with lower cognitive ability and/or have experienced a negative wealth shock would have a higher risk of death. In this situation, accounting for the censoring by death would be needed. This is because without randomization, there is a high likelihood that the proportion of deaths between subjects who did not receive a negative wealth shock versus those who received a wealth shock, would be imbalanced. In addition, subjects who die are more likely to have a lower cognition score. As a result, if we remove the subjects who died from our analysis, the effect of the negative wealth shock on cognitive ability that we measure would be confounded by death. Although MSM is commonly employed to weight the subjects who survived, this approach is arguably not sensible and would likely produce biased estimates when the effect depends on the principal strata as well as when adjustments on the weights have to be employed in order to stabilize the MSM estimate. To overcome these issues, we propose a new method to estimate the effect by imputing the counterfactual survival status of each subject in order to compare outcomes among individuals who would survive only under both sets of treatments being considered. Our method remained unbiased for all the simulation scenarios we tried and produced reasonable coverage. When applied to the HRS dataset, our method suggested that the effect of a negative wealth shock on the cognitive ability is close to null whereas the naïve method and MSM suggested an estimate with a slightly larger effect.

One shortcoming of our approach is our failure to incorporate the HRS sample design, in particular the sampling weights, in our inference. Given that a key use of weights in regression-type analysis is to reduce the effect of model misspecification (*Korn and Graubard*, 1995), we hope that our use of BART will minimize the degree of model misspecification. We leave the incorporation of such features in a general approach to future work. Another aspect of our method which could be improved is to allow our method to be applicable to studies where the follow-up time is not fixed. In such a situation, Cox based survival models would have to be employed and time would have to be included as a covariate in the survival and outcome models. The difficulty in this extension would be how to develop a systematic way, applicable to all subjects, to determine the relation in time between the allocation of the treatment, measuring the outcome, and death.

CHAPTER VI

Future work

6.1 Joint BART models

Although riBART was very successful in improving the prediction performance of BART under correlated outcomes, the model formulation of riBART was restrictive in the sense that it can only be applied to correlated continuous and binary outcomes. However, there may be various situations where the general idea of allowing a portion of the model to be specified using BART while the rest of the model is specified using other methods, for example, linear regression. Hence, it is worthwhile to investigate how we may generalize this idea and provide guidance for future researchers on the properties and important assumptions that such "joint" BART models require in order to provide valid estimation and inference. In the next few paragraphs, I shall briefly provide a brief sketch of my idea.

Suppose the usual BART model with Y_k being the outcomes, $\mathbf{X}_k$ being the covariates, T_j and M_j being the j^{th} tress structure and terminal nodes, and σ being the uncertainty parameter. We now add an additional model $H|\mathbf{Z}_k \sim P(\mathbf{Z}_k|\theta)$ where $P(\mathbf{Z}_k)$ is a distribution with parameter θ possibly depending on $\mathbf{Z}_k$, another set of q covariates $\mathbf{Z}_k = (Z_{k1}, \ldots, Z_{kq})^T$. Note that θ could be either a vector or scalar. The important feature of this modeling framework is that θ is independent of all T_js, $\mathbf{M}_j$s, and σ.

The whole modeling framework can thus be written as

$$Y_k = \sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j) + H(\mathbf{Z}_k; \theta) + \epsilon_k \tag{6.1}$$

where $\epsilon_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. This framework encompasses a few types of models discussed in literature thus far. The spatial BART (*Zhang et al.*, 2007) and survival BART model of (*Bonato et al.*, 2011) can be considered special cases of equation (6.1). Similarly, our riBART model in Chapter III can be considered a special case. More interesting special cases include setting the model $H(\mathbf{Z}_k; \theta)$ as BART.

6.2 Generalizing the censoring by death imputation method

In Chapter V we proposed a method which successfully allowed us to estimate the unbiased effect of a negative wealth shock on the cognitive score of late middle aged US adults under the presence of censoring due to subject death in our longitudinal dataset. Because our follow-up time was the same and at fixed intervals for each subject, we were able to estimate the survival model by using less complex models which need not consider the time of death. However, many studies may have follow-up times that are irregular and at different times for example clinical trials where subjects are not recruited at the same time and hence treatment allocation times and follow-up times would be different. In such scenarios, our proposed method is no longer applicable and some modifications would be required. Although a straight forward way to handle this problem is to include the time of death into the survival model and perhaps employ a Cox proportional hazard model to estimate the probability of death, the treatment allocation may be time-dependent especially when more than one treatment type is allocated to a subject during the study. This makes the problem more complex as we would need to decide how to handle the modeling of both the treatment allocation and event of death. Hence, a possible future work is to investigate how to extend our proposed model in Chapter V to irregular and varying follow-up times.

A second improvement for our proposed method in Chapter V is to investigate how to generalize our method to the target population of the data. In studies where the weights for the target population are provided, this is not as straightforward as re-weighting by the weights provided in the study. This is because these study reported weights are derived under the assumption that the sample is drawn from the study's target population. However, under principal stratification, the target population is no longer the study's population. The target population is now the principal strata population. Therefore, some form of adjustments would be needed in order for our proposed method to be generalized to the principal strata population of interest. I suggest two approaches here.

First, I propose to re-weight by the estimated probability of belonging to the principal strata. For example, let V be the baseline covariates, Z_t be the treatment allocation at time $t = 1, \ldots, T$ where $Z_t = 1$ indicates a subject receiving treatment at t. For simplicity, I assume that there are no time-varying covariates. Let $Y_{Z_1,\ldots,Z_t}$ be the potential outcome under treatments $Z_1, \ldots, Z_t$. To simplify notation, we write $Y_{Z_1=0} = Y_0$ and $Y_{Z_1=1} = Y_1$ for the potential outcomes at $t = 1$. For the potential survival outcome $S_{Z_1,\ldots,Z_{t-1}}$, $S_{Z_1,\ldots,Z_{t-1}} = 1$ indicates survival under treatments $Z_1, \ldots, Z_{t-1}$ at time t while $S_{Z_1,\ldots,Z_{t-1}} = 0$ indicates death. We simplify our notation for the potential survival outcomes as $S_{Z_1=0} = S_0$ and $S_{Z_1=1} = S_1$ at $t = 1$. Under the estimated effect $E[Y_{11}] - E[Y_{00}]$, the principal stratification is $S_1 = S_0 = 1$. Let PS be the principal stratification status where $PS = 1$ indicate that the subject was estimated to belong to the principal strata $S_1 = S_0 = 1$ and $PS = 0$ indicate that the subject does not belong to $S_1 = S_0 = 1$. The probability of the subjects belonging to principal strata $S_1 = S_0 = 1$ can then be estimated as $P(PS = 1|Z_1, Y_{Z_1}, V)$. The inverse estimated probability weight can be calculated as

$$w_{PS} = \frac{1}{P(PS = 1|Z_1, Y_{Z_1}, V)}. \tag{6.2}$$

The multiplication of the weight provided by the study and w_{PS} can then be used to re-weight our outcomes to obtain the estimated effect $E[Y_{11}] - E[Y_{00}]$ for the principal strata population of interest.

The second approach utilizes ideas discussed by *Zhou et al.* (2016) where they proposed to use Bayesian Finite bootstrapping methods to create a synthetic population and conduct analysis on the re-created synthetic population to obtain the estimated effect for the target population. The *Zhou et al.* (2016) method can be used to re-create the synthetic population and our method proposed in Chapter V can then be used to obtain the estimated effect and uncertainty. Results from such a method should be reflective of the target principal strata population.

For both approaches, work needs to be done to verify that these proposals are valid and if not valid, investigate whether there are ways to modify the approach to make them valid.

6.3 Bayesian Dynamic Treatment Regime

It is commonly known that variance estimates for commonly used DTR methods are difficult to obtain. However, if we setup the framework of DTR methods from a Bayesian perspective, variance estimates can be easily obtained from the MCMC draws making such an endeavor worthwhile. Briefly, DTR randomizes subjects to different treatment regimes at each follow-up time based on the subject's baseline and past covariates. At the start of the study, subjects would be randomized based on their baseline. At the next time point, depending on the collected covariates and possibly outcome in addition to the baseline information, the subject is once again randomized to another or the same treatment. Subsequent time points will then randomize subjects to the same or different treatment given current and past measured covariates and possibly outcome of interest. At the end of the study, the final outcome of interest is then recorded and the analysis proceeds to identify the treatment regime that maximizes the desirable outcome for each subject. Commonly, DTRs have been solved using algorithms such as Q-learning and reinforcement learning. Although a maximization technique is used here, we observe that some form of uncertainty is still involved in the sense that the value of the desired outcome is random and hence the maximum value observed involves some form of variability. Thus, the natural way to view this problem would be to think of each regime providing a probability that would maximize the desirable outcome for the subject. This suggests that formulating DTR in a Bayesian framework would also aid the understanding this problem and provide new insights.

6.3.1 Setup

Let $t = 0, 1, \ldots, T$ denote the different time points with 0 indicating baseline and $i = 1, \ldots, n$ denoting the subjects. Suppose we have q covariates denoted as $X_{1it}, \ldots, X_{qit}$ where $X_{1i0}, \ldots, X_{qi0}$ are the baseline covariates. X_{1it} means covariate 1 for subject i at time t and so on. Let $Z_{it} = 1, \ldots, z$ be the treatment for subject i at time t. This means that we have z possible treatments available to each subject although at time t depending on the values of the covariates and outcome, only some treatments are possible. Let Y_{it} be the outcome of interest for subject i at time t and let $p = 1, \ldots, P$ be the possible paths generated by the combination of treatments from $t = 0, \ldots, T-1$. Then $R_{ip} = Y_{iT}(p)$ would be the reward for path p which is just the potential outcome of treatment path p or regime p for individual i at the end of the study. Note that if path p was observed, under the usual causal inference assumptions, $R_{ip} = Y_{iT}$. Assume that a desirable outcome is to have R_{ip} be as large as possible we are then looking at a problem of

$$P(R_{ia} > R_{ip} | X_{1i0}, \ldots, X_{1iT}, \ldots, X_{qi0}, \ldots, X_{qiT}, Z_{i0}, \ldots, Z_{i,T-1}, Y_{i0}, \ldots, Y_{iT}) \tag{6.3}$$

$\forall\, p \neq a$, $a = 1, \ldots, P$.

6.3.2 Method

In the Bayesian context, this can be easily solved by examining the joint distribution of

$$P(R_{i1}, \ldots, R_{iP} | X_{1i0}, \ldots, X_{1iT}, \ldots, X_{qi0}, \ldots, X_{qiT}, Z_{i0}, \ldots, Z_{i,T-1}, Y_{i0}, \ldots, Y_{iT}), \tag{6.4}$$

The posterior joint distribution of the rewards. The main essential work will be then to tease out how the conditional distribution in equation (6.4) can be decomposed so that the usual MCMC or Bayesian draws can be made. For example, given certain combinations or values of the covariates at certain time points t, a certain path p may not be available to the subject. The decomposition would have to take this into consideration among many others like sequential randomization, etc. Assuming these issues have been taken care of and the MCMC algorithm is valid and can produce results in a timely manner, tackling equation (6.3) just reduces to investigating the proportion of $R_{ia} > R_{ip}\ \forall\, p \neq a$, $a = 1, \ldots, P$ in the MCMC draws. Moreover, equations (6.3) and (6.4) are written in a way that is very general so different types of DTR problems could be solved e.g.

$$P(a \leq R_{ip} \leq b | X_{1i0}, \ldots, X_{1iT}, \ldots, X_{qi0}, \ldots, X_{qiT}, Z_{i0}, \ldots, Z_{i,T-1}, Y_{i0}, \ldots, Y_{iT}) \tag{6.5}$$

i.e. we are interested in which regime would give us a potential outcome which would lie within a certain range of values $[a, b]$.

APPENDICES

APPENDIX A

Derivations of the conditional draws and Metropolis-Hastings ratio in BART

A.1 Posterior distributions for μ_{ij} and σ^2 in BART

A.1.1 $P(\mu_{ij}|T_j, \sigma, \mathbf{R}_j)$

Let $\mathbf{R}_{ij} = (r_{1j}, \ldots, r_{n_ij})^T$ be a subset from $\mathbf{R}_j$ where n_i is the number of r_{ij}s allocated to the terminal node with parameter μ_{ij}. We note that $\mathbf{R}_{ij}|g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j), \sigma \sim N(\mu_{ij}, \sigma^2)$ and $\mu_{ij}|T_j \sim N(\mu_\mu, \sigma_\mu^2)$. Then the posterior distribution of μ_{ij} is given by

$$
\begin{aligned}
P(\mu_{ij}|T_j, \sigma, \mathbf{R}_j) &\propto P(R_{ij}|T_j, \mu_{ij}, \sigma)P(\mu_{ij}|T_j) \\
&\propto \exp[-\frac{\sum_i (r_{ij} - \mu_{ij})^2}{2\sigma^2}] \exp[-\frac{(\mu_{ij} - \mu_\mu)^2}{2\sigma_\mu^2}] \\
&\propto \exp[-\frac{(n_i\sigma_\mu^2 + \sigma^2)\mu_{ij}^2 - 2(\sigma_\mu^2 \sum_i r_{ij} + \sigma^2\mu_\mu)\mu_{ij}}{2\sigma^2\sigma_\mu^2}] \\
&\propto \exp[-\frac{(\mu_{ij} - \frac{\sigma_\mu^2 \sum_i r_{ij} + \sigma^2\mu_\mu}{n_i\sigma_\mu^2 + \sigma^2})^2}{2\frac{\sigma^2\sigma_\mu^2}{n_i\sigma_\mu^2 + \sigma^2}}]
\end{aligned}
$$

where $\sum_i (r_{ij} - \mu_{ij})^2$ is the summation of the squared difference between the parameter μ_{ij} and the r_{ij}s allocated to the terminal node with parameter μ_{ij}.

A.1.2 $P(\sigma^2|(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \mathbf{Y})$

Let $\mathbf{Y} = (y_1, \ldots, y_n)^T$ and k index the subjects $k = 1, \ldots, n$. With $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$, we obtain the posterior draw of σ as follows

$$
\begin{aligned}
P(\sigma^2|(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \mathbf{Y}) &\propto P(\mathbf{Y}|(T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma)P(\sigma^2) \\
&= P(Y|\sum_{j=1}^{m} g(\mathbf{X}_k, T_j, \mathbf{M}_j), \sigma)P(\sigma^2) \\
&= \{\prod_{k=1}^{n} (\sigma^2)^{-\frac{1}{2}} \exp[-\frac{(y_k - \sum_{j=1}^{m} g_k(\mathbf{X}_k, T_j, \mathbf{M}_j))^2}{2\sigma^2}]\} \\
&\quad (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp(-\frac{\nu\lambda}{2\sigma^2}) \\
&= (\sigma^2)^{-(\frac{\nu+n}{2}+1)} \\
&\quad \exp[-\frac{\nu\lambda + \sum_{k=1}^{n}(y_k - \sum_{j=1}^{m} g_k(X_k, T_j, M_j))^2}{2\sigma^2}]
\end{aligned}
$$

where $\sum_j^m g_k(X_k, T_j, M_j)$ is the predicted value of BART assigned to observed outcome y_k.

A.2 Metropolis-Hastings ratio for the grow and prune step

This section is modified from Appendix A of *Kapelner and Bleich* (2016). Note that

$$
\alpha(T_j, T_j^*) = \min\{1, \frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} \frac{P(\mathbf{R}_j|X, T_j^*, \mathbf{M}_j)}{P(\mathbf{R}_j|X, T_j, \mathbf{M}_j)} \frac{P(T_j^*)}{P(T_j)}\}.
$$

where $\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)}$ is the transition ratio, $\frac{P(\mathbf{R}_j|X, T_j^*, \mathbf{M}_j)}{P(\mathbf{R}_j|X, T_j, \mathbf{M}_j)}$ is the likelihood ratio, and $\frac{P(T_j^*)}{P(T_j)}$ is the tree structure ratio of Kapelner and Bleich, Appendix A. We now present the explicit formula for each ratio under the grow and prune proposal.

A.2.1 Grow proposal

A.2.1.1 Transition ratio

$q(T_j^*, T_j)$ indicates the probability of moving from T_j to T_j^* i.e. selecting and terminal node and growing two children from T_j. Hence,

$$\begin{aligned}P(T_j^*|T_j) &= P(grow)P(\text{selecting terminal node to grow from})\times \\ &\quad P(\text{selecting covariate to split from})\times \\ &\quad P(\text{selecting value to split on}) \\ &= P(grow)\frac{1}{b_j}\frac{1}{p}\frac{1}{\eta}.\end{aligned}$$

In the above equation, $P(grow)$ is fixed at 0.5 in our codes, b_j is the number of available terminal nodes to split on in T_j, p is the number of variables left in the partition of the chosen terminal node, and η is the number of unique values left in the chosen variable after adjusting for the parents' splits.

$q(T_j, T_j^*)$ on the other hand indicates a pruning move which involves the probability of selecting the correct internal node to prune on such T_j^* becomes T_j. This is given as

$$\begin{aligned}P(T_j|T_j^*) &= P(prune)P(\text{selecting the correct internal node to prune}) \\ &= P(prune)\frac{1}{w_2^*}\end{aligned}$$

where w_2^* denotes the number of internal nodes which have only two children terminal nodes.

This gives a transition ratio of

$$\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} = \frac{P(T_j^*|T_j)}{P(T_j|T_j^*)} = \frac{P(prune)}{P(grow)}\frac{b_j p \eta}{w_2^*}.$$

If there are no variables with two or more unique values, this transition ratio will be set to 0.

A.2.1.2 Likelihood ratio

Since the rest of the tree structure will be the same between T_j^* and T_j except for the terminal node where the two children are grown, we need only concentrate on this terminal node. Let l be the selected node and l_L and l_R be the two children of the grow step. Then

$$\begin{aligned}\frac{P(\mathbf{R}_j|X,T_j^*,\mathbf{M}_j)}{P(\mathbf{R}_j|X,T_j,\mathbf{M}_j)} &= \frac{P(\mathbf{R}_{l_{(L,1)},j},\ldots,\mathbf{R}_{l_{(L,n_L)},j}|\sigma^2)P(\mathbf{R}_{l_{(R,1)},j},\ldots,\mathbf{R}_{l_{(R,n_R)},j}|\sigma^2)}{P(\mathbf{R}_{1,j},\ldots,\mathbf{R}_{n_l,j}|\sigma^2)}\\ &= \sqrt{\frac{\sigma^2(\sigma^2+n_l\sigma_\mu^2)}{(\sigma^2+n_L\sigma_\mu^2)(\sigma^2+n_R\sigma_\mu^2)}}\exp[\frac{\sigma_\mu^2}{2\sigma^2}(\frac{(\sum_{i=1}^{n_L}\mathbf{R}_{l_{(L,i)},j})^2}{\sigma^2+n_L\sigma_\mu^2}\\ &\quad+\frac{(\sum_{i=1}^{n_R}\mathbf{R}_{l_{(R,i)},j})^2}{\sigma^2+n_R\sigma_\mu^2}-\frac{(\sum_{i=1}^{n_l}\mathbf{R}_{l_{(l,i)},j})^2}{\sigma^2+n_l\sigma_\mu^2})].\end{aligned}$$

A.2.1.3 Tree structure ratio

Because the T can be specified using 3 aspects, we let $P_{SPLIT}(\theta)$ denote the probability that a selected node θ will split and $P_{RULE}(\theta)$ denote the probability that which variable and value is selected. Then based on $P_{SPLIT}(\theta) \propto \frac{\alpha}{(1+d_\theta)^\beta}$ and because T_j and T_j^* only differs at the children nodes, we have

$$\begin{aligned}\frac{P(T_j^*)}{P(T_j)} &= \frac{\prod_{\theta\in H^*_{terminals}}(1-P_{SPLIT}(\theta))\prod_{\theta\in H^*_{internals}}P_{SPLIT}(\theta)\prod_{\theta\in H^*_{internals}}P_{RULE}(\theta)}{\prod_{\theta\in H_{terminals}}(1-P_{SPLIT}(\theta))\prod_{\theta\in H_{internals}}P_{SPLIT}(\theta)\prod_{\theta\in H_{internals}}P_{RULE}(\theta)}\\ &= \frac{[1-P_{SPLIT}(\theta_L)][1-P_{SPLIT}(\theta_R)]P_{SPLIT}(\theta)P_{RULE}(\theta)}{1-P_{SPLIT}(\theta)}\\ &= \frac{(1-\frac{\alpha}{(1+d_{\theta_L})^\beta})(1-\frac{\alpha}{(1+d_{\theta_R})^\beta})\frac{\alpha}{(1+d_\theta)^\beta}\frac{1}{p}\frac{1}{\eta}}{\frac{\alpha}{(1+d_\theta)^\beta}}\\ &= \alpha\frac{(1-\frac{\alpha}{(2+d_\theta)^\beta})^2}{[(1+d_\theta)^\beta-\alpha]p\eta}\end{aligned}$$

because $d_{\theta_L}=d_{\theta_R}=d_\theta+1$.

A.2.2 Prune proposal

Since prune is the direct opposite of the grow proposal, the explicit formula of $\alpha(T_j, T_j^*)$ will just be the inverse of the grow proposal.

APPENDIX B

Derivations of the conditional draws for riBART MCMC algorithm

B.1 Posterior distributions of a_k and σ^2 for riBART

In this section, k still indexes the subjects and while i now indexes the number of repeated measures for each subject i.e. $i = 1, \ldots, n_k$. Let
$\mathbf{Y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{K1}, \ldots, y_{Kn_K})^T$ and $\hat{y}_{ik} = \sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j)$.

B.1.1 $P(a_k|\mathbf{Y}, (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma, \tau)$

Since $a_k \sim N(0, \tau^2)$, we have

$$\begin{aligned}
P(a_k|\mathbf{Y}, (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), \sigma, \tau) &\propto P(\mathbf{Y}|\sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j), \sigma, a_k)P(a_k|\tau^2) \\
&\propto \{\prod_{i=1}^{n_k} \exp[-\frac{(y_{ik} - \hat{y}_{ik} - a_k)^2}{2\sigma^2}]\} \exp[-\frac{a_k^2}{2\tau^2}] \\
&\propto \exp[-\frac{\sum_{i=1}^{n_k}(y_{ik} - \hat{y}_{ik} - a_k)^2}{2\sigma^2}] \exp[-\frac{a_k^2}{2\tau^2}] \\
&\propto \exp[-\frac{(n_k\tau^2 + \sigma^2)a_k^2 - 2\tau^2 a_k \sum_{i=1}^{n_k}(y_{ik} - \hat{y}_{ik})}{2\sigma^2\tau^2}] \\
&= \exp[-\frac{(a_k - \frac{\tau^2\sum_{i=1}^{n_k}(y_{ik}-\hat{y}_{ik})}{n_k\tau^2+\sigma^2})^2}{2\frac{\sigma^2\tau^2}{n_k\tau^2+\sigma^2}}].
\end{aligned}$$

B.1.2 $P(\sigma^2|\mathbf{Y}, (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), a_k, \tau)$

For the posterior of σ^2, since we have $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$, we obtain

$$\begin{aligned}
P(\sigma^2|\mathbf{Y}, (T_1, \mathbf{M}_1), \ldots, (T_m, \mathbf{M}_m), a_k, \tau) &\propto P(\mathbf{Y}|\sum_{j=1}^{m} g(\mathbf{X}_{ik}, T_j, \mathbf{M}_j), \sigma, a_k)P(\sigma^2) \\
&\propto \{\prod_{k=1}^{K}\prod_{i=1}^{n_k}(\sigma^2)^{-\frac{1}{2}} \exp[-\frac{(y_{ik} - \hat{y}_{ik} - a_k)^2}{2\sigma^2}]\} \\
&\quad (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp[-\frac{\nu\lambda}{2\sigma^2}] \\
&\propto (\sigma^2)^{-(\frac{N+\nu}{2}+1)} \\
&\quad \exp[-\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik} - \hat{y}_{ik} - a_k)^2 + \nu\lambda}{2\sigma^2}]
\end{aligned}$$

where $\sum_{k=1}^{K} n_k = N$.

B.2 Posterior distribution of τ under $P(\tau^2) \propto 1$ and $\tau^2 \sim IG(1,1)$

B.2.1 $\tau^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),a_k,\sigma$ for $P(\tau^2) \propto 1$

$$\begin{aligned}
P(\tau^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),a_k,\sigma) &\propto \{\prod_{k=1}^{K} P(a_k|\tau^2)\}P(\tau) \\
&\propto (\tau^2)^{-\frac{K}{2}} \exp[-\frac{\sum_{k=1}^{K} a_k^2}{2\tau^2}].
\end{aligned}$$

B.2.2 $\tau^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),a_k,\sigma$ for $\tau^2 \sim IG(1,1)$

$$\begin{aligned}
P(\tau^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),a_k,\sigma) &\propto \{\prod_{k=1}^{K} P(a_k|\tau^2)\}P(\tau) \\
&\propto (\tau^2)^{-\frac{K}{2}} \exp[-\frac{\sum_{k=1}^{K} a_k^2}{2\tau^2}](\tau^2)^{-(1+1)} \exp[-\frac{1}{\tau^2}] \\
&\propto (\tau^2)^{-(\frac{K}{2}+1+1)} \exp[-\frac{\sum_{k=1}^{K} a_k^2 + 2}{2\tau^2}].
\end{aligned}$$

B.3 Posterior distributions for ξ, η_k, θ and σ^2 for riBART with half-Cauchy prior on τ^2

B.3.1 $P(\xi|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\eta_k,\theta,\sigma)$

We note that $\xi \sim N(0,B^2)$, $\eta_k \sim N(0,\theta^2)$, $\sigma^2 \sim \nu\lambda\chi^2_\nu$, and $\theta^2 \sim IG(e,f)$. Now for

$$\begin{aligned}
P(\xi|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\eta_k,\theta,\sigma) &\propto P(Y|\sum_{j=1}^{m} g(\mathbf{X}_{ik},T_j,\mathbf{M}_j),\sigma,\eta_k,\xi)P(\xi) \\
&\propto \{\prod_{k=1}^{K}\prod_{i=1}^{n_k}(\sigma^2)^{-\frac{1}{2}}\exp[-\frac{(y_{ik}-\hat{y}_{ik}-\xi\eta_k)^2}{2\sigma^2}]\} \\
&\quad \exp[-\frac{\xi^2}{2B^2}] \\
&\propto \exp[-\frac{(\xi-\frac{B^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\eta_k(y_{ik}-\hat{y}_{ik})}{B^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\eta_k^2+\sigma^2})^2}{2\frac{\sigma^2B^2}{B^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\eta_k^2+\sigma^2}}].
\end{aligned}$$

is the kernel of a $N(\frac{B^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\eta_k(y_{ik}-\hat{y}_{ik})}{B^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\eta_k^2+\sigma^2},\frac{\sigma^2B^2}{B^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\eta_k^2+\sigma^2})$. Set $e=f=0.5$ and $B=25$ to obtain a half-Cauchy prior on τ^2.

B.3.2 $P(\eta_k|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\xi,\theta,\sigma)$

$$\begin{aligned}
P(\eta_k|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\xi,\theta,\sigma) &\propto P(Y|\sum_{j=1}^{m} g(X_{ik},T_j,\mathbf{M}_j),\sigma,\eta_k,\xi)P(\eta_k) \\
&\propto \{\prod_{i=1}^{n_k}(\sigma^2)^{-\frac{1}{2}}\exp[-\frac{(y_{ik}-\hat{y}_{ik}-\xi\eta_k)^2}{2\sigma^2}]\} \\
&\quad \exp[-\frac{\eta_k^2}{2\theta^2}] \\
&\propto \exp[-\frac{(\eta_k-\frac{\theta^2\xi\sum_{i=1}^{n_k}(y_{ik}-\hat{y}_{ik})}{\theta^2\xi^2n_k+\sigma^2})^2}{2\frac{\sigma^2\theta^2}{\theta^2\xi^2n_k+\sigma^2}}].
\end{aligned}$$

B.3.3 $P(\theta^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\xi,\eta_k,\sigma)$

$$\begin{aligned}
P(\theta^2|Y,(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\xi,\eta_k,\sigma) &\propto \{\prod_{k=1}^{K} p(\eta_k|\theta^2)\}p(\theta^2) \\
&\propto (\theta^2)^{-\frac{K}{2}} \exp[-\frac{\sum_{k=1}^{K}\eta_k^2}{2\theta^2}](\theta^2)^{-(\frac{e}{2}-1)} \exp[-\frac{ef}{2\theta^2}] \\
&\propto (\theta^2)^{-(\frac{e+K}{2}-1)} \exp[-\frac{\sum_{k=1}^{K}\eta_k^2+ef}{2\theta^2}].
\end{aligned}$$

B.3.4 $P(\sigma^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\xi,\eta_k,\theta)$

$$\begin{aligned}
P(\sigma^2|\mathbf{Y},(T_1,\mathbf{M}_1),\ldots,(T_m,\mathbf{M}_m),\xi,\eta_k,\theta) &\propto P(\mathbf{Y}|\sum_{j=1}^{m} g(\mathbf{X}_{ik},T_j,\mathbf{M}_j),\sigma,\xi,\eta_k,\theta)P(\sigma^2) \\
&\propto \{\prod_{k=1}^{K}\prod_{i=1}^{n_k}(\sigma^2)^{-\frac{1}{2}} \exp[-\frac{(y_{ik}-\hat{y}_{ik}-\xi\eta_k)^2}{2\sigma^2}]\} \\
&\quad (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp[-\frac{\nu\lambda}{2\sigma^2}] \\
&\propto (\sigma^2)^{-(\frac{N+\nu}{2}+1)} \\
&\quad \exp[-\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik}-\hat{y}_{ik}-\xi\eta_k)^2+\nu\lambda}{2\sigma^2}]
\end{aligned}$$

APPENDIX C

Data preparation for Chapter III

Our original data contains the time series of speed for the vehicle every 10 milliseconds starting from 100 meters away from the center of an intersection. We rescale the original time series predictors to measure distance-series of vehicle speed from the intersection because, in a turn that is not complete, only the distance from the intersection will be known in advance. We recorded the distance series at every single meter i.e. $d = -100, \ldots, -1$ where 0 is the center of an intersection and -100 is 100 meters from the center of an intersection. To determine the vehicle speed at a certain meter, we searched for the vehicle speed recorded that was closet to the meter mark. In the situation where more than one speed sample point was closest to the meter, we took their average as the speed at that meter.

Because vehicles can stop and restart before reaching the center of the intersection, we define "stopping" as a distance-varying outcome. Let $i = 1, \ldots, n_k$ index the i^{th} turn made by the k^{th} driver where $k = 1, \ldots, K$ index the driver. Let s_{ikd} be the distance series of vehicle speed and y_{ikd} be the distance-varying outcome (1=stopped in future, 0=will not stop in future). We defined y_{ikd} as follows:

1. If $s_{ikd} > 1m/s \, \forall \, d = -100, \ldots, -1$, then set $y_{ikd} = 0$ for all d.

2. If $s_{ikd} \leq 1m/s$ for some $d \in \{-100, \ldots, -1\}$, let $c \in \{-100, \ldots, -1\}$ be the index such that for every $d > c$, $s_{ikd} > 1m/s$. We set $y_{ik,-100} = y_{ik,-99} = \ldots = y_{ik,c} = 1$ and $y_{ik,c+1} = y_{ik,c+2} = \ldots = y_{ik,-1} = 0$.

Next, for every d^{th} meter, we defined the moving window of speeds as,

$$M_{ikd} = \{s_{ik,d-w+1}, s_{ik,d-w+2}, \ldots, s_{ikd}\}$$

where w is the size of the moving window. We the implemented PCA on these M_{ikd}s to reduce the number of covariates in our prediction model. Before reduction, the covariates are $s_{..,k-w+1}, s_{..,k-w+2}, \ldots, s_{..d}$. We let

$$M_d = \begin{bmatrix} s_{11,d-w+1} & s_{11,d-w+2} & \cdots & s_{11j} \\ \vdots & \vdots & \vdots & \vdots \\ s_{1n_1,j-w+1} & s_{1n_1,j-w+2} & \cdots & s_{1n_1j} \\ \vdots & \vdots & \vdots & \vdots \\ s_{Kn_K,j-w+1} & s_{Kn_K,j-w+2} & \cdots & s_{Kn_Kj} \end{bmatrix}$$

and

$$u(d) = \begin{bmatrix} u_{d-w+1} \\ u_{d-w+2} \\ \vdots \\ u_d \end{bmatrix}$$

where M_d is the matrix of moving windows with the first row being M_{11d}, n_1^{th} row being M_{1n_1d},..., and the last row being M_{Kn_Kd}. There are w (number of columns in M_d) orthogonal vectors $u(d)$ that decompose the variance of M_d into w parts under the condition that for each $u(d)$, $||u(d)|| = 1$. To obtain the w decomposed variances, we used the formula: $PC_d = Var[M_d u(d)]$. If we let $PC_{d(q)}$ be the ordered statistic where $q = 1, \ldots, w$ and $u(d)_{(q)}$ be the ordered vector corresponding to $PC_{d(q)}$, then the first PC is $\mathbf{X}_{d1} = M_d u(d)_{(w)}$, the second PC is $\mathbf{X}_{d2} = M_d u(d)_{(w-1)}$, and so on.

We used the first two PCs in our analysis for reasons already covered in our main paper. We then added a third predictor, the number of stops made by the vehicle until distance d to obtain Table 3.3.

APPENDIX D

Consistency of the AIPWT estimator

The AIPWT estimator is a consistent estimator for the population mean parameter μ when either the propensity model or mean model in equation (1) is correctly specified. To see this, we first assume that $\hat{\beta} \xrightarrow{p} \beta^*$ and $\hat{\theta} \xrightarrow{p} \theta^*$ i.e. the parameters in equations (2) and (3) are consistent. This is valid since the models we used to estimate these parameters were multiple linear regression and multiple logistic regression which under the usual maximum likelihood assumptions, will converge asymptotically to their true values. From equation (1), this implies that

$$\begin{aligned}
\hat{\mu}_{AIPWT} &= \frac{1}{n}\sum_{k=1}^{n}\{\frac{R_kY_k}{Z_k} - \frac{R_k - Z_k}{Z_k}m(\mathbf{X}_k, \hat{\beta})\} \\
&\xrightarrow{p} E[\frac{R_kY_k}{Z_k} - \frac{R_k - Z_k}{Z_k}m(\mathbf{X}_k, \hat{\beta})] \\
&= E[Y_k - Y_k + \frac{R_kY_k}{Z_k} - \frac{R_k - Z_k}{Z_k}m(\mathbf{X}_k, \hat{\beta})] \\
&= \mu + E[\frac{R_kY_k}{Z_k} - \frac{R_k - Z_k}{Z_k}m(\mathbf{X}_k, \hat{\beta}) - Y_k] \\
&= \mu + E[\frac{R_k}{Z_k}Y_k - \frac{R_k}{Z_k}m(\mathbf{X}_k, \hat{\beta}) - \{Y_k - m(\mathbf{X}_k, \hat{\beta})\}] \\
&= \mu + E[\{\frac{R_k}{Z_k} - 1\}\{Y_k - m(\mathbf{X}_k, \hat{\beta})\}].
\end{aligned}$$

Under the MAR assumption, we have $Y \perp R|X$. Hence, we have

$$\begin{aligned}\hat{\mu}_{AIPWT} &\xrightarrow{p} \mu + E[\{\frac{R_k}{Z_k} - 1\}\{Y_k - m(\mathbf{X}_k, \hat{\beta})\}] \\ &= \mu + E[E[\{\frac{R_k}{Z_k} - 1\}\{Y_k - m(\mathbf{X}_k, \hat{\beta})\}|\mathbf{X}_k]] \\ &= \mu + E[E[(\frac{R_k}{Z_k} - 1)|\mathbf{X}_k]E[(Y_k - m(\mathbf{X}_k, \hat{\beta}))|\mathbf{X}_k]].\end{aligned}$$

Suppose that the true propensity model is $\pi_0(X)$ and the propensity model in equation (1) is correctly specified. Then $Z_k \xrightarrow{p} \pi_0(\mathbf{X}_k)$ and

$$\begin{aligned}E[(\frac{R_k}{Z_k} - 1)|\mathbf{X}_k] &\xrightarrow{p} E[(\frac{R_k}{\pi_0(\mathbf{X}_k)} - 1)|\mathbf{X}_k] \\ &= \frac{\pi_0(\mathbf{X}_k)}{\pi_0(\mathbf{X}_k)} - 1 \\ &= 0.\end{aligned}$$

This implies that $\hat{\mu}_{AIPWT} \xrightarrow{p} \mu$ if the propensity model is correctly specified regardless of whether the mean model is correctly specified. Now suppose that the true mean model is $m_0(\mathbf{X}_k)$ and the mean model in equation (1) is correctly specified. Then $m(\mathbf{X}_k, \hat{\beta}) \xrightarrow{p} m_0(\mathbf{X}_k)$ and

$$\begin{aligned}E[(Y_k - m(\mathbf{X}_k, \hat{\beta}))|\mathbf{X}_k] &\xrightarrow{p} E[(Y_k - m_0(\mathbf{X}_k))|\mathbf{X}_k] \\ &= \mu - \mu \\ &= 0.\end{aligned}$$

Hence, $\hat{\mu}_{AIPWT} \xrightarrow{p} \mu$ if the mean model is correctly specified.

APPENDIX E

Consistency of the PSPP estimator

We show that the PSPP model is doubly robust closely following Zhang and Little (2009)'s arguments in the first corollary of their supplementary materials. We first rewrite equation (4) as

$$(Y_k|Z_k, X_{k1}, \ldots, X_{kp}; \phi, \eta) \sim N(s(Z_k; \phi) + f(X_{k1}, \ldots, X_{kp}, \eta), \sigma^2), \tag{E.1}$$

where $s[Z_k; \phi] = \phi_0 + \sum_{l=1}^{L} \phi_l Z_k^L + \sum_{h=1}^{H} \phi_{L+h}(Z_k - \tau_h)_+^L$. Suppose we specified the mean function $f(X_{k1}, \ldots, X_{kp}, \eta)$ correctly, then $s(Z_k; \phi)$ is absorbed into the error term and hence $s(Z_k; \phi) + f(X_{k1}, \ldots, X_{kp}, \eta) \xrightarrow{p} \mu$.

Now suppose instead that equation (3) was specified correctly. We consider two scenarios, one where we omit $f(X_{k1}, \ldots, X_{kp}, \eta)$ in equation (4) and the other where $f(X_{k1}, \ldots, X_{kp}, \eta)$ is specified. Let

$$\mathbf{Z} = [1, Z_k, (Z_k - \tau_1)_+, \ldots, (Z_k - \tau_L)_+],$$

the truncated linear basis of the propensity score and

$$\mathbf{X} = [f_1(X_{k1}, \ldots, X_{kp}), \ldots, f_T(X_{k1}, \ldots, X_{kp})] = [V_{k1}, \ldots, V_{kT}]$$

be the elements in the function f. Let T be the total number of elements in f. For the scenario where we omit $f(X_{k1}, \ldots, X_{kp}, \eta)$, $E[Y_k|Z_k] = \phi_0 + \phi_1 Z_k + \sum_{h=1}^{H} \phi_{1+h}(Z_k - \tau_h)_+$ and we obtain ϕ by minimizing $||\mathbf{Y} - \mathbf{Z}\phi||^2 + \lambda^2 \phi^T \mathbf{D} \phi$ where $\phi = (\phi_0, \phi_1, \ldots, \phi_{1+H})^T$, λ is the penalty, and $\mathbf{D} = \text{diag}(1_H)$. Using the mixed model representation and by restricted maximum likelihood estimation, $\hat{\mathbf{Y}}(Z_k, \hat{\lambda}, \mathbf{D}) = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \hat{\lambda}^2\mathbf{D})^{-1}\mathbf{Z}^T\mathbf{Y}$. As $n \to \infty$, $\hat{\lambda} \to 0$ and hence the predicted value of $\mathbf{Y}$ converges to

$$\hat{\mathbf{Y}}(\mathbf{Z}, 0, \mathbf{D}) = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y} = \mathbf{Z}\hat{\phi}. \tag{E.2}$$

Equation (E.2) estimates the marginal mean of $\mathbf{Y}$ consistently because of the balancing property of propensity score, Z_k, that is, missingness is completely at random conditional on Z_k, so predicted values of Y_k using a smooth function of Z should yield consistent estimation of the missing values.

If $f(X_{k1}, \ldots, X_{kp}), \eta)$ was specified but incorrect, then the conditional mean of $\mathbf{Y}$ is

$$\begin{aligned} E[Y_k|Z_k, X_{k1}, \ldots, X_{kp}] &= s(Z_k; \phi) + f(X_{k1}, \ldots, X_{kp}, \eta) \\ &= \phi_0 + \phi_1 Z_k + \sum_{h=1}^{H} \phi_{1+h}(Z_k - \tau_h)_+ + \mathbf{X}\eta. \end{aligned}$$

$(\phi, \eta)^T$ is obtained by minimizing $||\mathbf{Y} - [\mathbf{Z}, \mathbf{X}](\phi, \eta)^T||^2 + \lambda^2(\phi, \eta)\mathbf{D}(\phi, \eta)^T$ where λ is the penalty and $\mathbf{D} = \text{diag}(1_H, 0_{2+T})$. Using the mixed model representation and by restricted maximum likelihood estimation,

$$\hat{\mathbf{Y}}(Z_k, X_{k1}, \ldots, X_{kp}, \hat{\lambda}, \mathbf{D}) = [\mathbf{Z}, \mathbf{X}]([\mathbf{Z}, \mathbf{X}]^T[\mathbf{Z}, \mathbf{X}] + \hat{\lambda}^2\mathbf{D})^{-1}[\mathbf{Z}, \mathbf{X}]^T\mathbf{Y}.$$

When $n \to \infty$, $\hat{\lambda} \to 0$ and

$$\hat{\mathbf{Y}}(Z_k, X_{k1}, \ldots, X_{kp}, \hat{\lambda}, \mathbf{D}) \to [\mathbf{Z}, \mathbf{X}]([\mathbf{Z}, \mathbf{X}]^T[\mathbf{Z}, \mathbf{X}])^{-1}[\mathbf{Z}, \mathbf{X}]^T\mathbf{Y},$$

the predicted value of $\mathbf{Y}$ can then be written as

$$\hat{\mathbf{Y}}(Z_k, X_{k1}, \ldots, X_{kp}, 0, \mathbf{D}) = \mathbf{Z}\hat{\phi} + \mathbf{X}\hat{\eta}. \tag{E.3}$$

Now we regress each term in f on the propensity score i.e. V_i on $\mathbf{Z}$ for all $i = 1, \ldots, T$ where $\mathbf{Z}$ is the predictor and each V_i are the outcome. As $n \to \infty$, the predicted value of each element in f, $\hat{\mathbf{V}}_i(\mathbf{Z}; \hat{\lambda}) \to \hat{\mathbf{V}}_i(\mathbf{Z}; 0) \to \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}\mathbf{V}_i$. Let $\hat{\mathbf{X}} = [\mathbf{V}_1, \ldots, \mathbf{V}_T]$ and substitute $\hat{\mathbf{X}}$ into equation (E.3). Then

$$E[\hat{\mathbf{Y}}(Z_k, X_{k1}, \ldots, X_{kp})|Z_k] = \mathbf{Z}\hat{\phi} + \hat{\mathbf{X}}\hat{\eta}. \tag{E.4}$$

By lemma 1 in Zhang and Little (2009)'s supplementary materials, equation (E.4) converges to equation (E.2) as $n \to \infty$ and hence equation (4) is consistent for the marginal mean of Y if the propensity model is correctly specified but the mean model is incorrectly specified.

APPENDIX F

Simulation Results for Sample Sizes 500, 1,000, and 5,000

F.1 Linear interaction in mean model

Table F.1: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 500 using bootstrap.

$n = 500$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	-0.01	0.13	94	0.48	-0.01	0.13	94	0.48
CC	0.51	0.53	9.8	0.59	0.51	0.53	9.8	0.59
MLR	0	0.17	98.6	0.87	0.44	0.47	40.6	0.81
PSPP	-0.01	0.21	99	1.29	0.05	0.18	97.4	0.86
AIPWT	0	0.18	93.4	0.67	0.03	0.36	89.2	1.04
PSBPP	0	0.18	98.8	0.95	-0.06	0.21	98.6	1.1
AIPWT with BART	0.15	0.23	78.4	0.61	0.15	0.23	78.4	0.61
BART	0.19	0.26	86.8	0.79	0.19	0.26	86.8	0.79
BARTps	0.11	0.21	93.6	0.83	0.11	0.21	93.6	0.83
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	-0.01	0.13	94	0.48	-0.01	0.13	94	0.48
CC	0.51	0.53	9.8	0.59	0.51	0.53	9.8	0.59
MLR	0	0.17	98.6	0.87	0.44	0.47	40.6	0.81
PSPP	0	0.18	98.8	0.9	0.25	0.31	90.4	1.04
AIPWT	0	0.18	93	0.65	0.42	0.46	25.8	0.61
PSBPP	0	0.18	98.8	0.95	-0.06	0.21	98.6	1.1
AIPWT with BART	0.15	0.23	78.4	0.61	0.15	0.23	78.4	0.61
BART	0.19	0.26	86.8	0.79	0.19	0.26	86.8	0.79
BARTps	0.11	0.21	93.6	0.83	0.11	0.21	93.6	0.83

Table F.2: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 1,000 using bootstrap.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.2	0.34	0	0.09	94.2	0.34
CC	0.51	0.53	0.6	0.42	0.51	0.53	0.6	0.42
MLR	0	0.12	99	0.62	0.45	0.46	10	0.57
PSPP	0.01	0.14	99.8	0.78	0.05	0.13	97.4	0.61
AIPWT	0	0.12	94.4	0.47	0.04	0.18	87.2	0.6
PSBPP	0	0.13	99.2	0.64	-0.06	0.15	98.4	0.71
AIPWT with BART	0.11	0.17	78.2	0.44	0.11	0.17	78.2	0.44
BART	0.14	0.19	87.4	0.57	0.14	0.19	87.4	0.57
BARTps	0.07	0.14	95.8	0.6	0.07	0.14	95.8	0.6
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.2	0.34	0	0.09	94.2	0.34
CC	0.51	0.53	0.6	0.42	0.51	0.53	0.6	0.42
MLR	0	0.12	99	0.62	0.45	0.46	10	0.57
PSPP	0	0.12	99	0.63	0.22	0.26	85	0.78
AIPWT	0	0.12	92.8	0.46	0.43	0.45	5	0.43
PSBPP	0	0.13	99.2	0.64	-0.06	0.15	98.4	0.71
AIPWT with BART	0.11	0.17	78.2	0.44	0.11	0.17	78.2	0.44
BART	0.14	0.19	87.4	0.57	0.14	0.19	87.4	0.57
BARTps	0.07	0.14	95.8	0.6	0.07	0.14	95.8	0.6

Table F.3: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 5,000 using bootstrap.

$n = 5,000$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.04	96	0.15	0	0.04	96	0.15
CC	0.51	0.52	0	0.19	0.51	0.52	0	0.19
MLR	0	0.05	99.4	0.27	0.44	0.45	0	0.26
PSPP	0	0.05	99.2	0.29	0.03	0.06	97.6	0.27
AIPWT	0	0.05	94.4	0.21	0.01	0.1	88.4	0.32
PSBPP	0	0.05	99.4	0.28	-0.04	0.07	97.2	0.29
AIPWT with BART	0.05	0.07	80.6	0.2	0.05	0.07	80.6	0.2
BART	0.06	0.08	88.8	0.27	0.06	0.08	88.8	0.27
BARTps	0.02	0.06	97.6	0.27	0.02	0.06	97.6	0.27
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.04	96	0.15	0	0.04	96	0.15
CC	0.51	0.52	0	0.19	0.51	0.52	0	0.19
MLR	0	0.05	99.4	0.27	0.44	0.45	0	0.26
PSPP	0	0.05	99.4	0.28	0.16	0.19	69.4	0.4
AIPWT	0	0.05	94.8	0.2	0.43	0.43	0	0.19
PSBPP	0	0.05	99.4	0.28	-0.04	0.07	97.2	0.29
AIPWT with BART	0.05	0.07	80.6	0.2	0.05	0.07	80.6	0.2
BART	0.06	0.08	88.8	0.27	0.06	0.08	88.8	0.27
BARTps	0.02	0.06	97.6	0.27	0.02	0.06	97.6	0.27

Table F.4: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 500 using MI with posterior mean of propensity scores.

$n = 500$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.13	95	0.49	0	0.13	95	0.49
CC	0.51	0.53	4	0.41	0.51	0.53	4	0.41
MLR	0	0.18	94.6	0.7	0.44	0.47	27.4	0.66
PSPP	0	0.21	96.4	0.9	0.06	0.19	92.2	0.69
PSBPP	0	0.18	94.6	0.76	0.01	0.2	93.8	0.79
BART	0.18	0.26	80.6	0.68	0.18	0.26	80.6	0.68
BARTps	0.1	0.21	90.6	0.73	0.1	0.21	90.6	0.73
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.13	95	0.49	0	0.13	95	0.49
CC	0.51	0.53	4	0.41	0.51	0.53	4	0.41
MLR	0	0.18	94.6	0.7	0.44	0.47	27.4	0.66
PSPP	0	0.18	94.2	0.73	0.22	0.32	69.2	0.73
PSBPP	0	0.18	94.6	0.76	0.01	0.2	93.8	0.79
BART	0.18	0.26	80.6	0.68	0.18	0.26	80.6	0.68
BARTps	0.1	0.21	90.6	0.73	0.1	0.21	90.6	0.73

Table F.5: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 1,000 using MI with posterior mean of propensity scores.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.6	0.35	0	0.09	94.6	0.35
CC	0.51	0.52	0	0.29	0.51	0.52	0	0.29
MLR	0	0.12	95	0.5	0.44	0.46	5.2	0.47
PSPP	0	0.14	96.6	0.58	0.06	0.14	93.2	0.49
PSBPP	0	0.13	95.6	0.52	0	0.14	94.4	0.55
BART	0.14	0.19	78.6	0.49	0.14	0.19	78.6	0.49
BARTps	0.06	0.15	90.6	0.52	0.06	0.15	90.6	0.52
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.6	0.35	0	0.09	94.6	0.35
CC	0.51	0.52	0	0.29	0.51	0.52	0	0.29
MLR	0	0.12	95	0.5	0.44	0.46	5.2	0.47
PSPP	0	0.13	95.2	0.51	0.2	0.27	63.8	0.51
PSBPP	0	0.13	95.6	0.52	0	0.14	94.4	0.55
BART	0.14	0.19	78.6	0.49	0.14	0.19	78.6	0.49
BARTps	0.06	0.15	90.6	0.52	0.06	0.15	90.6	0.52

Table F.6: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 5,000 using MI with posterior mean of propensity scores.

$n = 5,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.04	97	0.15	0	0.04	97	0.15
CC	0.51	0.52	0	0.13	0.51	0.52	0	0.13
MLR	0	0.05	96.8	0.22	0.44	0.45	0	0.21
PSPP	0	0.05	96.4	0.24	0.03	0.07	90	0.22
PSBPP	0	0.05	96.6	0.22	0	0.06	95.4	0.23
BART	0.06	0.08	80.2	0.22	0.06	0.08	80.2	0.22
BARTps	0.03	0.06	94.8	0.23	0.03	0.06	94.8	0.23
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.04	97	0.15	0	0.04	97	0.15
CC	0.51	0.52	0	0.13	0.51	0.52	0	0.13
MLR	0	0.05	96.8	0.22	0.44	0.45	0	0.21
PSPP	0	0.05	96	0.22	0.16	0.19	33.6	0.22
PSBPP	0	0.05	96.6	0.22	0	0.06	95.4	0.23
BART	0.06	0.08	80.2	0.22	0.06	0.08	80.2	0.22
BARTps	0.03	0.06	94.8	0.23	0.03	0.06	94.8	0.23

Table F.7: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 500 using MI with posterior draw of propensity scores.

$n = 500$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.13	95.2	0.49	0	0.13	95.2	0.49
CC	0.51	0.53	4	0.41	0.51	0.53	4	0.41
MLR	0	0.18	94.6	0.7	0.44	0.47	27.6	0.66
PSPP	0.01	0.2	96.6	0.92	0.06	0.18	93.6	0.69
PSBPP	0	0.18	94.6	0.75	0.2	0.27	86.2	0.84
BART	0.18	0.26	80.8	0.68	0.18	0.26	80.8	0.68
BARTps	0.17	0.25	84.2	0.71	0.17	0.25	84.2	0.71
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.13	95.2	0.49	0	0.13	95.2	0.49
CC	0.51	0.53	4	0.41	0.51	0.53	4	0.41
MLR	0	0.18	94.6	0.7	0.44	0.47	27.6	0.66
PSPP	0	0.18	94.2	0.73	0.25	0.32	80.6	0.91
PSBPP	0	0.18	94.6	0.75	0.2	0.27	86.2	0.84
BART	0.18	0.26	80.8	0.68	0.18	0.26	80.8	0.68
BARTps	0.17	0.25	84.2	0.71	0.17	0.25	84.2	0.71

Table F.8: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 1,000 using MI with posterior draw of propensity scores.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.6	0.35	0	0.09	94.6	0.35
CC	0.51	0.53	0	0.29	0.51	0.53	0	0.29
MLR	0	0.12	95	0.5	0.44	0.46	5.2	0.47
PSPP	0	0.14	96.4	0.58	0.06	0.14	91.6	0.49
PSBPP	0	0.13	95.4	0.52	0.16	0.2	83.8	0.6
BART	0.14	0.19	78.4	0.49	0.14	0.19	78.4	0.49
BARTps	0.12	0.18	84	0.51	0.12	0.18	84	0.51
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.09	94.6	0.35	0	0.09	94.6	0.35
CC	0.51	0.53	0	0.29	0.51	0.53	0	0.29
MLR	0	0.12	95	0.5	0.44	0.46	5.2	0.47
PSPP	0	0.13	95.6	0.51	0.23	0.27	79.6	0.72
PSBPP	0	0.13	95.4	0.52	0.16	0.2	83.8	0.6
BART	0.14	0.19	78.4	0.49	0.14	0.19	78.4	0.49
BARTps	0.12	0.18	84	0.51	0.12	0.18	84	0.51

Table F.9: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the linear interaction in mean model scenario with sample size 5,000 using MI with posterior draw of propensity scores.

$n = 5,000$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.04	97	0.15	0	0.04	97	0.15
CC	0.51	0.52	0	0.13	0.51	0.52	0	0.13
MLR	0	0.05	96.8	0.22	0.44	0.45	0	0.21
PSPP	0	0.05	97.8	0.23	0.04	0.06	92	0.23
PSBPP	0	0.05	95.4	0.23	0.08	0.09	82	0.26
BART	0.06	0.08	80.2	0.22	0.06	0.08	80.2	0.22
BARTps	0.05	0.07	87.2	0.23	0.05	0.07	87.2	0.23
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.04	97	0.15	0	0.04	97	0.15
CC	0.51	0.52	0	0.13	0.51	0.52	0	0.13
MLR	0	0.05	96.8	0.22	0.44	0.45	0	0.21
PSPP	0	0.05	97.4	0.22	0.17	0.19	63.6	0.38
PSBPP	0	0.05	95.4	0.23	0.08	0.09	82	0.26
BART	0.06	0.08	80.2	0.22	0.06	0.08	80.2	0.22
BARTps	0.05	0.07	87.2	0.23	0.05	0.07	87.2	0.23

F.2 Quadratic interaction in mean model

Table F.10: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 500 using bootstrap.

$n = 500$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.02	0.34	90.4	1.19	0.02	0.34	90.4	1.19
CC	1.21	1.23	2.8	0.89	1.21	1.23	2.8	0.89
MLR	0.01	0.36	96.8	1.84	1.24	1.26	3.4	1.15
PSPP	0.01	0.37	97.2	1.87	0.31	0.54	82.6	2.39
AIPWT	0.02	0.36	91.2	1.31	0.16	1.75	63.8	2.75
PSBPP	0.01	0.37	97.4	1.87	0.21	0.49	92.4	3.03
AIPWT with BART	0.57	0.67	35	1.02	0.57	0.67	35	1.02
BART	0.64	0.71	46.8	1.28	0.64	0.71	46.8	1.28
BARTps	0.54	0.63	60.4	1.41	0.54	0.63	60.4	1.41
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.02	0.34	90.4	1.19	0.02	0.34	90.4	1.19
CC	1.21	1.23	2.8	0.89	1.21	1.23	2.8	0.89
MLR	0.01	0.36	96.8	1.84	1.24	1.26	3.4	1.15
PSPP	0.01	0.36	97	1.87	0.83	0.89	61.4	2.05
AIPWT	0.02	0.36	91.4	1.3	1.21	1.23	2	0.84
PSBPP	0.01	0.37	97.4	1.87	0.21	0.49	92.4	3.03
AIPWT with BART	0.57	0.67	35	1.02	0.57	0.67	35	1.02
BART	0.64	0.71	46.8	1.28	0.64	0.71	46.8	1.28
BARTps	0.54	0.63	60.4	1.41	0.54	0.63	60.4	1.41

Table F.11: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 1,000 using bootstrap.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	91.8	0.86	0	0.24	91.8	0.86
CC	1.21	1.23	0.2	0.63	1.21	1.23	0.2	0.63
MLR	0	0.26	99	1.32	1.24	1.25	0	0.8
PSPP	0	0.26	98.8	1.33	0.21	0.44	81.2	2
AIPWT	0	0.26	91.2	0.93	0.22	0.72	67	1.68
PSBPP	0	0.26	98.6	1.33	0.13	0.35	94	2.16
AIPWT with BART	0.45	0.51	29.8	0.77	0.45	0.51	29.8	0.77
BART	0.52	0.57	42	0.97	0.52	0.57	42	0.97
BARTps	0.41	0.47	63.4	1.07	0.41	0.47	63.4	1.07
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	91.8	0.86	0	0.24	91.8	0.86
CC	1.21	1.23	0.2	0.63	1.21	1.23	0.2	0.63
MLR	0	0.26	99	1.32	1.24	1.25	0	0.8
PSPP	0	0.26	98.6	1.33	0.72	0.77	61.8	1.69
AIPWT	0	0.25	91	0.92	1.21	1.22	0	0.59
PSBPP	0	0.26	98.6	1.33	0.13	0.35	94	2.16
AIPWT with BART	0.45	0.51	29.8	0.77	0.45	0.51	29.8	0.77
BART	0.52	0.57	42	0.97	0.52	0.57	42	0.97
BARTps	0.41	0.47	63.4	1.07	0.41	0.47	63.4	1.07

Table F.12: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 5,000 using bootstrap.

$n = 5,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.1	94	0.39	0.01	0.1	94	0.39
CC	1.21	1.21	0	0.29	1.21	1.21	0	0.29
MLR	0.01	0.11	98.6	0.59	1.24	1.24	0	0.37
PSPP	0.01	0.11	98.6	0.59	0.12	0.23	80.8	0.87
AIPWT	0.01	0.11	95	0.42	0.09	0.45	71.8	1.16
PSBPP	0.01	0.11	98.6	0.59	0.09	0.17	91.2	0.9
AIPWT with BART	0.24	0.26	26	0.39	0.24	0.26	26	0.39
BART	0.28	0.3	40.6	0.5	0.28	0.3	40.6	0.5
BARTps	0.2	0.23	67.8	0.54	0.2	0.23	67.8	0.54
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.1	94	0.39	0.01	0.1	94	0.39
CC	1.21	1.21	0	0.29	1.21	1.21	0	0.29
MLR	0.01	0.11	98.6	0.59	1.24	1.24	0	0.37
PSPP	0.01	0.11	98.6	0.59	0.52	0.56	50.6	0.99
AIPWT	0.01	0.11	93.8	0.42	1.21	1.21	0	0.27
PSBPP	0.01	0.11	98.6	0.59	0.09	0.17	91.2	0.9
AIPWT with BART	0.24	0.26	26	0.39	0.24	0.26	26	0.39
BART	0.28	0.3	40.6	0.5	0.28	0.3	40.6	0.5
BARTps	0.2	0.23	67.8	0.54	0.2	0.23	67.8	0.54

Table F.13: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 500 using MI with posterior mean of propensity scores.

$n = 500$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.02	0.34	91	1.23	0.02	0.34	91	1.23
CC	1.2	1.23	0.8	0.62	1.2	1.23	0.8	0.62
MLR	0.01	0.37	93.2	1.37	1.24	1.26	2	1
PSPP	0.01	0.36	93	1.39	0.28	0.65	61.2	1.18
PSBPP	0.01	0.37	93.8	1.4	0.25	0.66	68.4	1.31
BART	0.59	0.68	35	0.94	0.59	0.68	35	0.94
BARTps	0.47	0.59	49.8	1.04	0.47	0.59	49.8	1.04
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.02	0.34	91	1.23	0.02	0.34	91	1.23
CC	1.2	1.23	0.8	0.62	1.2	1.23	0.8	0.62
MLR	0.01	0.37	93.2	1.37	1.24	1.26	2	1
PSPP	0.02	0.37	93.2	1.39	0.75	0.87	36.4	1.16
PSBPP	0.01	0.37	93.8	1.4	0.25	0.66	68.4	1.31
BART	0.59	0.68	35	0.94	0.59	0.68	35	0.94
BARTps	0.47	0.59	49.8	1.04	0.47	0.59	49.8	1.04

Table F.14: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 1,000 using MI with posterior mean of propensity scores.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	92.6	0.89	0	0.24	92.6	0.89
CC	1.21	1.23	0	0.44	1.21	1.23	0	0.44
MLR	0	0.26	94.2	0.98	1.24	1.25	0	0.7
PSPP	0	0.26	94.4	0.99	0.16	0.56	54.4	0.88
PSBPP	0	0.26	93.8	0.99	0.21	0.41	68.2	0.91
BART	0.47	0.53	31.4	0.71	0.47	0.53	31.4	0.71
BARTps	0.35	0.44	51	0.77	0.35	0.44	51	0.77
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	92.6	0.89	0	0.24	92.6	0.89
CC	1.21	1.23	0	0.44	1.21	1.23	0	0.44
MLR	0	0.26	94.2	0.98	1.24	1.25	0	0.7
PSPP	0	0.26	94	0.99	0.66	0.75	28.4	0.84
PSBPP	0	0.26	93.8	0.99	0.21	0.41	68.2	0.91
BART	0.47	0.53	31.4	0.71	0.47	0.53	31.4	0.71
BARTps	0.35	0.44	51	0.77	0.35	0.44	51	0.77

Table F.15: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 5,000 using MI with posterior mean of propensity scores.

$n = 5,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.1	94.8	0.4	0.01	0.1	94.8	0.4
CC	1.21	1.21	0	0.2	1.21	1.21	0	0.2
MLR	0.01	0.11	95	0.44	1.24	1.24	0	0.32
PSPP	0.01	0.11	95	0.44	0.1	0.24	52	0.4
PSBPP	0.01	0.11	95.2	0.44	0.15	0.22	56.4	0.39
BART	0.25	0.27	29.8	0.36	0.25	0.27	29.8	0.36
BARTps	0.17	0.21	56.8	0.39	0.17	0.21	56.8	0.39
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.1	94.8	0.4	0.01	0.1	94.8	0.4
CC	1.21	1.21	0	0.2	1.21	1.21	0	0.2
MLR	0.01	0.11	95	0.44	1.24	1.24	0	0.32
PSPP	0.01	0.11	95.4	0.44	0.5	0.56	15	0.39
PSBPP	0.01	0.11	95.2	0.44	0.15	0.22	56.4	0.39
BART	0.25	0.27	29.8	0.36	0.25	0.27	29.8	0.36
BARTps	0.17	0.21	56.8	0.39	0.17	0.21	56.8	0.39

Table F.16: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 500 using MI with posterior draw of propensity scores.

$n = 500$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.34	90.8	1.24	0.01	0.34	90.8	1.24
CC	1.2	1.23	0.8	0.62	1.2	1.23	0.8	0.62
MLR	0.01	0.37	93	1.38	1.23	1.26	2	1
PSPP	0.01	0.37	92.4	1.39	0.3	0.52	72	1.46
PSBPP	0.01	0.37	93.6	1.41	0.67	0.79	54	2.18
BART	0.59	0.68	35.4	0.94	0.59	0.68	35.4	0.94
BARTps	0.57	0.66	40.4	1.03	0.57	0.66	40.4	1.03
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.34	90.8	1.24	0.01	0.34	90.8	1.24
CC	1.2	1.23	0.8	0.62	1.2	1.23	0.8	0.62
MLR	0.01	0.37	93	1.38	1.23	1.26	2	1
PSPP	0.01	0.37	93.6	1.41	0.79	0.86	50.4	1.72
PSBPP	0.01	0.37	93.6	1.41	0.67	0.79	54	2.18
BART	0.59	0.68	35.4	0.94	0.59	0.68	35.4	0.94
BARTps	0.57	0.66	40.4	1.03	0.57	0.66	40.4	1.03

Table F.17: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 1,000 using MI with posterior draw of propensity scores.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	92.2	0.89	0	0.24	92.2	0.89
CC	1.21	1.23	0	0.44	1.21	1.23	0	0.44
MLR	0	0.26	94	0.98	1.24	1.25	0	0.7
PSPP	0	0.26	93.6	0.99	0.19	0.45	71.4	1.24
PSBPP	0	0.26	93	0.99	0.53	0.64	58.6	2.07
BART	0.47	0.53	30.6	0.71	0.47	0.53	30.6	0.71
BARTps	0.44	0.51	39	0.78	0.44	0.51	39	0.78
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0	0.24	92.2	0.89	0	0.24	92.2	0.89
CC	1.21	1.23	0	0.44	1.21	1.23	0	0.44
MLR	0	0.26	94	0.98	1.24	1.25	0	0.7
PSPP	0	0.26	93.6	0.99	0.71	0.77	52.8	1.53
PSBPP	0	0.26	93	0.99	0.53	0.64	58.6	2.07
BART	0.47	0.53	30.6	0.71	0.47	0.53	30.6	0.71
BARTps	0.44	0.51	39	0.78	0.44	0.51	39	0.78

Table F.18: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) of the eight estimators under the quadratic interaction in mean model scenario with sample size 5,000 using MI with posterior draw of propensity scores.

$n = 5,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.1	94.8	0.4	0.01	0.1	94.8	0.4
CC	1.21	1.21	0	0.2	1.21	1.21	0	0.2
MLR	0.01	0.11	95	0.44	1.24	1.24	0	0.32
PSPP	0.01	0.11	95.4	0.44	0.1	0.23	65.6	0.54
PSBPP	0.01	0.11	95	0.44	0.32	0.36	65.4	1.31
BART	0.25	0.27	29.8	0.36	0.25	0.27	29.8	0.36
BARTps	0.23	0.25	43.4	0.41	0.23	0.25	43.4	0.41
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.1	94.8	0.4	0.01	0.1	94.8	0.4
CC	1.21	1.21	0	0.2	1.21	1.21	0	0.2
MLR	0.01	0.11	95	0.44	1.24	1.24	0	0.32
PSPP	0.01	0.11	95.8	0.44	0.53	0.58	41.6	0.91
PSBPP	0.01	0.11	95	0.44	0.32	0.36	65.4	1.31
BART	0.25	0.27	29.8	0.36	0.25	0.27	29.8	0.36
BARTps	0.23	0.25	43.4	0.41	0.23	0.25	43.4	0.41

F.3 Kang and Schafer (2007) example

Table F.19: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 500 using bootstrap.

$n = 500$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.11	1.58	95.6	6.17	0.11	1.58	95.6	6.17
CC	-9.96	10.2	0.2	8.42	-9.96	10.2	0.2	8.42
MLR	0.03	1.55	99.2	9.04	-0.66	2.04	99.4	10.99
PSPP	0.03	1.55	99.2	9.04	-0.11	1.71	99.6	9.66
AIPWT	0.1	1.58	95.6	6.18	0.25	2.17	94	8.19
PSBPP	0.03	1.55	99.2	9.04	1.72	2.54	97.4	11.18
AIPWT with BART	-0.13	1.6	94.4	6.37	-0.6	1.75	91	6.74
BART	-0.32	1.59	99.6	9.12	-1.05	1.94	98.6	9.36
BARTps	-0.06	1.57	99.4	9.39	0.49	1.7	99	9.9
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.11	1.58	95.6	6.17	0.11	1.58	95.6	6.17
CC	-9.96	10.2	0.2	8.42	-9.96	10.2	0.2	8.42
MLR	0.03	1.55	99.2	9.04	-0.66	2.04	99.4	10.99
PSPP	0.03	1.55	99.2	9.04	-1.99	2.77	92	9.42
AIPWT	0.32	5.09	95.6	8	-46.4	858.18	68.6	326.68
PSBPP	0.03	1.55	99.2	9.04	-1.39	2.43	98.8	11.89
AIPWT with BART	-0.13	1.6	94.4	6.37	-0.75	1.81	90.8	6.74
BART	-0.32	1.59	99.6	9.12	-1.05	1.94	98.6	9.36
BARTps	-0.15	1.58	99.8	9.39	-0.89	1.87	99.2	9.75

Table F.20: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 1,000 using bootstrap.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.07	1.11	95.2	4.38	0.07	1.11	95.2	4.38
CC	-9.96	10.09	0	5.97	-9.96	10.09	0	5.97
MLR	0.07	1.11	99.4	6.38	-0.74	1.63	98	7.78
PSPP	0.06	1.11	99.4	6.38	-0.07	1.21	99.2	6.66
AIPWT	0.06	1.11	95.6	4.38	0.07	1.66	94.2	6.01
PSBPP	0.07	1.11	99.4	6.38	1.46	1.95	96.8	7.4
AIPWT with BART	-0.05	1.12	95.2	4.42	-0.31	1.19	93.8	4.61
BART	-0.13	1.12	99.6	6.38	-0.59	1.29	99.2	6.5
BARTps	0	1.11	99.4	6.46	0.39	1.23	99.2	6.8
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.07	1.11	95.2	4.38	0.07	1.11	95.2	4.38
CC	-9.96	10.09	0	5.97	-9.96	10.09	0	5.97
MLR	0.07	1.11	99.4	6.38	-0.74	1.63	98	7.78
PSPP	0.07	1.11	99.4	6.38	-2.12	2.52	77.2	6.29
AIPWT	-0.08	2.28	95.6	5.1	-35.69	477.13	41.2	196.51
PSBPP	0.07	1.11	99.4	6.38	-1.13	1.73	99	7.84
AIPWT with BART	-0.06	1.12	95.2	4.42	-0.45	1.24	93.2	4.62
BART	-0.13	1.12	99.6	6.38	-0.59	1.29	99.2	6.5
BARTps	-0.05	1.12	99.6	6.46	-0.52	1.27	99.2	6.7

Table F.21: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 5,000 using bootstrap.

$n = 5,000$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.02	0.49	96.4	1.96	0.02	0.49	96.4	1.96
CC	-9.94	9.97	0	2.65	-9.94	9.97	0	2.65
MLR	0.02	0.5	99.4	2.87	-0.84	1.06	92	3.49
PSPP	0.02	0.5	99.4	2.87	-0.04	0.53	99.2	2.95
AIPWT	0.01	0.49	96.2	1.96	0.05	0.7	94.4	2.76
PSBPP	0.02	0.5	99.4	2.87	0.86	1.01	88.8	3.07
AIPWT with BART	-0.02	0.49	95.8	1.97	-0.08	0.51	95.6	2
BART	-0.04	0.5	99.4	2.87	-0.24	0.56	99.2	2.89
BARTps	-0.02	0.5	99.4	2.87	0.16	0.53	99.6	2.9
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.02	0.49	96.4	1.96	0.02	0.49	96.4	1.96
CC	-9.94	9.97	0	2.65	-9.94	9.97	0	2.65
MLR	0.02	0.5	99.4	2.87	-0.84	1.06	92	3.49
PSPP	0.02	0.5	99.4	2.87	-2.28	2.36	5.2	2.62
AIPWT	-0.01	0.59	96	2.07	-19.29	91.29	0.2	44.76
PSBPP	0.02	0.5	99.4	2.87	-0.26	0.6	99.2	3.17
AIPWT with BART	-0.02	0.49	95.8	1.97	-0.21	0.54	94.2	2
BART	-0.04	0.5	99.4	2.87	-0.24	0.56	99.2	2.89
BARTps	-0.02	0.5	99.4	2.87	-0.23	0.56	99.2	2.9

Table F.22: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 500 using MI with posterior mean of propensity scores.

$n = 500$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.09	1.56	96.8	6.36	0.09	1.56	96.8	6.36
CC	-10.02	10.25	0	6.11	-10.02	10.25	0	6.11
MLR	0.08	1.56	96.4	6.34	-0.74	2.13	95	8.08
PSPP	0.08	1.56	96.4	6.34	-0.06	1.74	95	6.69
PSBPP	0.09	1.56	96.4	6.35	1.39	2.28	91.2	7.7
BART	-0.15	1.58	96.4	6.36	-0.74	1.8	93.4	6.74
BARTps	-0.05	1.58	97	6.47	0.35	1.7	97.4	6.97
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.09	1.56	96.8	6.36	0.09	1.56	96.8	6.36
CC	-10.02	10.25	0	6.11	-10.02	10.25	0	6.11
MLR	0.08	1.56	96.4	6.34	-0.74	2.13	95	8.08
PSPP	0.08	1.56	96.4	6.34	-1.99	2.82	80.4	7.79
PSBPP	0.08	1.56	96.4	6.35	-1.4	2.46	91.4	8.17
BART	-0.15	1.58	96.4	6.36	-0.74	1.8	93.4	6.74
BARTps	-0.08	1.58	96.8	6.48	-0.61	1.76	94.2	6.96

Table F.23: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 1,000 using MI with posterior mean of propensity scores.

$n = 1,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.05	1.11	96.2	4.49	0.05	1.11	96.2	4.49
CC	-9.97	10.11	0	4.32	-9.97	10.11	0	4.32
MLR	0.04	1.11	96.4	4.49	-0.82	1.68	90.8	5.69
PSPP	0.05	1.11	96.4	4.49	-0.07	1.22	94.8	4.71
PSBPP	0.05	1.11	96.6	4.49	0.99	1.61	91.2	5.27
BART	-0.08	1.13	96.2	4.5	-0.46	1.24	95	4.71
BARTps	-0.04	1.12	96.2	4.54	0.26	1.21	95.4	4.83
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.05	1.11	96.2	4.49	0.05	1.11	96.2	4.49
CC	-9.97	10.11	0	4.32	-9.97	10.11	0	4.32
MLR	0.04	1.11	96.4	4.49	-0.82	1.68	90.8	5.69
PSPP	0.05	1.11	96.4	4.49	-2.17	2.57	65.4	5.34
PSBPP	0.05	1.11	96.4	4.49	-1.41	1.92	83.8	5.55
BART	-0.08	1.13	96.2	4.5	-0.46	1.24	95	4.71
BARTps	-0.03	1.13	96.4	4.54	-0.4	1.24	95.4	4.85

Table F.24: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 5,000 using MI with posterior mean of propensity scores.

$n = 5,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.01	0.49	96.4	2.01	0.01	0.49	96.4	2.01
CC	-9.94	9.96	0	1.93	-9.94	9.96	0	1.93
MLR	0.01	0.49	96.4	2.01	-0.86	1.08	70.8	2.54
PSPP	0.01	0.49	96.4	2.01	-0.03	0.53	97	2.1
PSBPP	0.01	0.49	96	2.01	0.44	0.69	89	2.22
BART	-0.03	0.49	96.6	2.01	-0.19	0.54	95.8	2.08
BARTps	-0.02	0.5	96.8	2.01	0.1	0.51	96	2.08
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.49	96.4	2.01	0.01	0.49	96.4	2.01
CC	-9.94	9.96	0	1.93	-9.94	9.96	0	1.93
MLR	0.01	0.49	96.4	2.01	-0.86	1.08	70.8	2.54
PSPP	0.01	0.49	96.4	2.01	-2.28	2.36	1.8	2.32
PSBPP	0.01	0.49	96.4	2.01	-0.49	0.73	89.2	2.33
BART	-0.03	0.49	96.6	2.01	-0.19	0.54	95.8	2.08
BARTps	-0.02	0.5	97	2.01	-0.21	0.55	95.4	2.1

Table F.25: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 500 using MI with posterior draw of propensity scores.

$n = 500$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.09	1.56	96.8	6.36	0.09	1.56	96.8	6.36
CC	-10.02	10.25	0	6.11	-10.02	10.25	0	6.11
MLR	0.08	1.56	96.4	6.34	-0.74	2.13	95	8.08
PSPP	0.08	1.56	96.4	6.34	-0.11	1.72	96.6	7.22
PSBPP	0.09	1.56	96.4	6.35	0.3	1.81	98	8.68
BART	-0.15	1.58	96.4	6.36	-0.74	1.8	93.4	6.74
BARTps	-0.12	1.58	97	6.5	-0.39	1.67	96.4	7.08
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.09	1.56	96.8	6.36	0.09	1.56	96.8	6.36
CC	-10.02	10.25	0	6.11	-10.02	10.25	0	6.11
MLR	0.08	1.56	96.4	6.34	-0.74	2.13	95	8.08
PSPP	0.09	1.56	96.6	6.35	-1.94	2.73	84.2	8.24
PSBPP	0.08	1.56	96.4	6.35	-1.22	2.26	96	8.88
BART	-0.15	1.58	96.4	6.36	-0.74	1.8	93.4	6.74
BARTps	-0.13	1.58	97	6.47	-0.76	1.81	93.8	6.95

Table F.26: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 1,000 using MI with posterior draw of propensity scores.

$n = 1,000$		Both correct				Propensity correct		
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.05	1.11	96.2	4.49	0.05	1.11	96.2	4.49
CC	-9.97	10.11	0	4.32	-9.97	10.11	0	4.32
MLR	0.04	1.11	96.4	4.49	-0.82	1.68	90.8	5.69
PSPP	0.05	1.11	96.4	4.49	-0.12	1.21	96.4	5.03
PSBPP	0.05	1.11	96.4	4.49	0.07	1.26	98.2	6.14
BART	-0.08	1.13	96.2	4.5	-0.46	1.24	95	4.71
BARTps	-0.07	1.12	96.4	4.54	-0.22	1.18	96.6	4.91
		Mean correct				Both wrong		
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.05	1.11	96.2	4.49	0.05	1.11	96.2	4.49
CC	-9.97	10.11	0	4.32	-9.97	10.11	0	4.32
MLR	0.04	1.11	96.4	4.49	-0.82	1.68	90.8	5.69
PSPP	0.05	1.11	96.4	4.49	-2.12	2.52	68.6	5.47
PSBPP	0.05	1.11	96.6	4.49	-1.42	1.92	89.4	6.25
BART	-0.08	1.13	96.2	4.5	-0.46	1.24	95	4.71
BARTps	-0.07	1.12	96.4	4.54	-0.47	1.25	96.2	4.85

Table F.27: Bias, RMSE, 95% coverage, and average 95% confidence interval length (AIL) under the Kang and Schafer (2007) example with sample size 5,000 using MI with posterior draw of propensity scores.

$n = 5,000$	Both correct				Propensity correct			
Method	Bias	RMSE	Coverage	AIL*	Bias	RMSE	Coverage	AIL
BD	0.01	0.49	96.4	2.01	0.01	0.49	96.4	2.01
CC	-9.94	9.96	0	1.93	-9.94	9.96	0	1.93
MLR	0.01	0.49	96.4	2.01	-0.86	1.08	70.8	2.54
PSPP	0.01	0.49	96.4	2.01	-0.04	0.53	97.6	2.21
PSBPP	0.01	0.49	96.2	2.01	-0.17	0.55	98.6	2.53
BART	-0.03	0.49	96.6	2.01	-0.19	0.54	95.8	2.08
BARTps	-0.02	0.5	96.8	2.02	-0.08	0.51	96.8	2.1
	Mean correct				Both wrong			
Method	Bias	RMSE	Coverage	AIL	Bias	RMSE	Coverage	AIL
BD	0.01	0.49	96.4	2.01	0.01	0.49	96.4	2.01
CC	-9.94	9.96	0	1.93	-9.94	9.96	0	1.93
MLR	0.01	0.49	96.4	2.01	-0.86	1.08	70.8	2.54
PSPP	0.01	0.49	96.4	2.01	-2.27	2.34	2.8	2.35
PSBPP	0.01	0.49	96	2.01	-1	1.13	74.6	2.7
BART	-0.03	0.49	96.6	2.01	-0.19	0.54	95.8	2.08
BARTps	-0.02	0.5	96.8	2.02	-0.2	0.54	95.8	2.11

APPENDIX G

Web Appendix D: Simple descriptive statistics for NASS-CDS 2014 and FARS 2015

G.1 NASS-CDS 2014

Table G.1: Summary statistics stratified by missingness in total delta-v.

Variables	Missing (%)	Non-missing (%)	*p*-value
Crash type			< 0.01
Change traffic-way, vehicle turning	18.4	32.3	
Same traffic-way	29.3	32.8	
Single driver	31.1	11.5	
Others or missing	21.3	23.4	
Heading angle			< 0.01
Frontal	17.8	24.6	
Back	16.2	21.3	
Left	18.2	21.3	
Right	16.6	21.7	
Missing	31.2	11.1	
Climate			0.09
Clear	76.1	72.9	
Cloudy	11.1	12.7	
Others or Missing	12.8	14.4	
Bodytype			0.23
Automobiles	66.9	67.8	
SUV	17.1	18.1	
Trucks	16	14	
Curb weight			< 0.01
< 1500kg	38.7	42.3	
1500-2000kg	40.5	44	
≥ 2000kg or Missing	20.8	13.7	
Documentation of trajectory?			0.75
Yes	22.4	22.9	
No	77.6	77.1	
Driver distracted?			< 0.01
Attentive	23.7	29.6	
Distracted	10.7	10.4	
Missing	65.6	60	
Police reported alcohol presence			< 0.01
Yes	9.3	6.5	
No	84.2	88.1	
Missing	6.4	5.4	
Pre-impact location			< 0.01
Stayed on roadway	69.1	86.5	
Did not stay on roadway or missing	30.9	13.5	

Table G.2: Summary statistics stratified by missingness in total delta-v, continued.

Variables	Missing (%)	Non-missing (%)	*p*-value
No. of lanes			< 0.01
≤ 2 or Missing	46	42.5	
3	17.5	17.6	
4	14.2	19.4	
5	15.5	15.3	
≥ 6	6.9	5.2	
Light condition			< 0.01
Dark	10.6	6.7	
Dark but lighted	25	23.9	
Daylight	60.3	65.7	
Dusk, Dawn, or Missing	4.1	3.7	
Vehicle make			0.11
American	47.2	50.6	
Japanese	39.6	36.3	
Europe or other foreign	13.2	13.1	
Avoidance maneuver?			< 0.01
Yes	18.8	23.6	
No	35.8	36.7	
Missing	45.4	39.8	
Model year			0.16
< 2003 or Missing	33.7	31.5	
≥ 2003	66.3	68.5	
No. of occupants			0.81
1	71.3	70.6	
2	19.1	19.9	
≥ 3	9.6	9.5	
Pre-crash event			< 0.01
Traveling	42.1	35.1	
Loss control	9.3	5.7	
Other or Missing	48.6	59.3	
Pre-event movement			0.59
Going straight	55.6	54.6	
Other or Missing	44.4	45.4	
Pre-impact stability			< 0.01
Skidding	10.6	9.5	
Tracking	74.2	79	
Other or Missing	15.3	11.5	
Road alignment			0.31
Straight	79.5	80.7	
Curve left	10.6	9.1	
Curve right	9.9	10.2	

Table G.3: Summary statistics stratified by missingness in total delta-v, continued.

Variables	Missing (%)	Non-missing (%)	*p*-value
Surface condition			0.03
Dry	81.4	79.6	
Wet	12.6	15.5	
Other or Missing	5.9	4.9	
Surface type			0.07
Concrete	13.4	11.4	
Asphalt and Others	86.6	88.6	
Race			< 0.01
White	34.8	38.4	
Black	10.3	12.3	
Other or Missing	54.9	49.3	
Relation to interchange			< 0.01
Interchange area related	12.7	11.5	
Intersection related	41.4	57.5	
Non-interchange area and non-junction	45.9	31	
Other drug test results			< 0.01
No test given	80.9	87.2	
Drugs found	1.7	2.6	
Drugs not found	3	2	
Results not known	3.2	1.9	
Missing	11.1	6.2	
Traffic control device			< 0.01
No traffic control	64.7	54.8	
Traffic control signal	25.3	35.8	
Other or Missing	10.1	9.4	
Travel speed			< 0.01
$\leq$ 40km/h	13.3	15.9	
40-80km/h	10.6	14.2	
$>$ 80km/h	7.4	4.2	
Missing	68.7	65.7	
Traffic flow			< 0.01
Not Divided or One way	66.4	66.6	
Divided with barrier	18.9	12.7	
Divided/no barrier	14.8	20.8	
Other drug present?			< 0.01
Yes	2.3	1.7	
No	75.4	82.6	
Missing	22.4	15.7	
Vehicle has roof?			< 0.01
Yes	81.1	86.9	
No or missing	18.9	13.1	

Table G.4: Summary statistics stratified by missingness in total delta-v, continued.

Variables	Missing (%)	Non-missing (%)	*p*-value
Antilock brakes			< 0.01
Not available	3.4	3.5	
Standard	71.6	75	
Optional	16.9	19.2	
Missing	8.1	2.4	
Daytime running lights			< 0.01
Not available	34.9	34.3	
Standard	39.5	43.6	
Optional	14.1	16.3	
Missing	11.5	5.7	
Other vehicle body type			< 0.01
Automobiles	23.5	57.6	
SUV	9.9	20.9	
Trucks	10.1	15.9	
Other or Missing	56.6	5.6	
Direct damage width			< 0.01
$<$ 50cm	10.1	14.4	
50-100cm	7.3	21.6	
100-150cm	8.1	26.7	
$\geq$ 150cm	8.5	21.8	
Missing	66	15.6	
Highest deformation extent			< 0.01
1	12.8	26.3	
≥ 2	32.5	60.9	
Missing	54.7	12.8	
Second highest deformation extent			< 0.01
1	8.8	15	
≥ 2	8.8	10.6	
Missing	82.4	74.4	
Second highest object contacted			< 0.01
Vehicle	11.1	18.1	
Other	15.1	12.6	
Missing	73.8	69.3	
Principal direction of force			< 0.01
Frontal	40	63	
Back	5.4	9.3	
Left	5.1	9.8	
Right	4.9	8	
Other or Missing	44.6	9.9	

Table G.5: Summary statistics stratified by missingness in total delta-v, continued.

Variables	Missing (%)	Non-missing (%)	*p*-value
No. of seriously injured occupants			0.15
0	29.6	28	
$\geq$>= 1	4	5.3	
Missing	66.4	66.7	
Age			0.49
< 21 or Missing	12.7	12.9	
21-30	26.4	25	
30-40	18.9	19.2	
40-50	14.9	13.9	
50-60	12.9	12.7	
$\geq$ 60	14.2	16.4	
Police reported airbag use			< 0.01
Not deployed	33.4	28.9	
Deployed	38.9	52.5	
Not reported	22.1	12	
Other or Missing	5.6	6.6	
Driver's height			0.63
< 160cm	5.4	6.3	
160-170cm	14.6	15.4	
170-180cm	16.8	15.7	
$\geq$ 180cm	12.6	11.9	
Missing	50.6	50.8	
Police reported injury severity			< 0.01
No injury (O)	44.6	39.2	
Possible injury (C)	18.7	21.3	
Nonincapaciting injury (B)	10.8	15.5	
Incapacitating injury (A)	16.1	18.3	
Killed (K)	5.6	2.8	
Unknown injury or Missing	4.2	2.8	
Police reported belt use			0.01
None used	8.1	6.6	
Used	82.1	85.7	
Not reported or Missing	9.8	7.6	
Sex			< 0.01
Female	40.4	46	
Male or Missing	59.6	54	

Table G.6: Summary statistics stratified by missingness in total delta-v, continued.

Variables	Missing (%)	Non-missing (%)	*p*-value
Driver's weight			0.84
< 60kg	6.4	6.9	
60-70kg	9	9.6	
70-80kg	11.7	10.5	
80-90kg	8.2	9	
90-100kg	6	6.2	
≥ 100kg	8.4	8.2	
Missing	50.3	49.6	

G.2 2015 FARS

Table G.7: Summary statistics stratified by missingness in blood alcohol concentration (BAC)

Variables	Missing (%)	Non-missing (%)	*p*-value
Hour of crash			< 0.01
12-6am	15.5	23.1	
6-10am	13.1	13.4	
10am-4pm	26.8	23.5	
4-8pm	24.3	19.9	
8pm-12am	20	19.4	
Unknown	0.4	0.7	
Day of crash			< 0.01
Mon-Thu	52.9	50.1	
Fri	15.9	15.1	
Sat	16.1	17.7	
Sun	15.1	17.1	
Intersection type			< 0.01
4-way	23.2	15.7	
Other	11	9.1	
Not an intersection, Not reported, or Unknown	65.8	75.2	
Work zone?			0.06
Yes	2.3	1.9	
No	97.7	98.1	
Relation to road			< 0.01
On roadside	14.5	31.3	
On roadway	80.7	60.2	
Other, Not reported, or Unknown	4.8	8.5	
Climate			0.01
Clear	71.6	69.4	
Cloudy	16.7	17.7	
Rain	7.9	8.9	
Other, Not reported, or Unknown1	3.7	4	
No. of fatalities			< 0.01
1	92.5	87.8	
2	6.3	9.9	
≥ 3	1.2	2.3	
Number of motor vehicles in transport			< 0.01
≤ 2	78	90.2	
≥ 3	22	9.8	
Functional system			< 0.01
Arterial	56.7	53.5	
Collector	10.7	17.1	
Interstate	12.8	11.5	
Local, not in state inventory, not reported, or unknown	19.8	17.9	
Manner of collision			< 0.01
Front to front	13.3	18.1	
Front to rear	12.4	7.4	
Angle	29.1	22.8	
Non-collision, other, not reported, or unknown	45.1	51.7	
Month of crash			< 0.01
Jan	7.5	8.5	
Feb	6.4	6.6	
Mar	7.5	8.1	
Apr	7	8.1	
May	8.3	8.9	
Jun	8.3	8.3	
Jul	9.1	8.4	
Aug	8.5	9	
Sep	8.7	8.6	
Oct	9.7	8.8	
Nov	9.4	8.3	
Dec	9.6	8.5	
Vehicle make			< 0.01
American	43.6	50.1	
Japanese	41.2	36.2	
Other	15.2	13.6	

Table G.8: Summary statistics stratified by missingness in blood alcohol concentration (BAC), continued

Variables	Missing (%)	Non-missing (%)	p-value
Model year			< 0.01
Before 1995	4.6	7.5	
1995-2005	33.8	41	
2005-2015	55.7	48.1	
Beyond 2015 or unknown	5.9	3.3	
Fire?			< 0.01
Yes	2	4.4	
No	98	95.6	
Age			< 0.01
Younger than 21	12.3	12.8	
21-30	24.6	28.7	
30-40	16.1	17	
40-50	11.8	10.7	
50-60	12.2	11.7	
Older than 60	21.4	19.1	
Not reported or unknown	1.6	0.1	
Sex			< 0.01
Male	58.7	66.1	
Female, not reported or unknown	41.3	33.9	
Police reported injury severity			< 0.01
No injury (O), Not reported or unknown	37.6	11.8	
Possible injury (C)	12.4	4.1	
Minor injury (B)	12.1	6.6	
Serious injury (A)	11	6.7	
Fatal injury (K)	27	70.8	
Restraint used			< 0.01
None used	12.6	33.1	
Lap and shoulder belt use	75	56.1	
Other, not applicable, not reported, or unknown	12.4	10.8	
Air bag deployed?			< 0.01
Not deployed or switched off	43.6	27.8	
Deployed	49.1	64.3	
Not applicable, not reported, or unknown	7.3	7.9	
Driver extricated?			< 0.01
Extricated	9.5	24.1	
Not extricated	88.1	71.4	
Unknown	2.4	4.4	
Police reported alcohol involvement			< 0.01
Yes	4.8	24.3	
No	67.1	47.9	
Not reported	17.2	6.9	
Unknown	10.9	20.9	
Method of alcohol determination			< 0.01
Evidential Test	0.8	25.3	
Other	7.6	10.4	
Not reported	91.6	64.3	

Table G.9: Summary statistics stratified by missingness in blood alcohol concentration (BAC), continued

Variables	Missing (%)	Non-missing (%)	*p*-value
Alcohol test type			< 0.01
Blood test	2.3	91.9	
Other	0.1	6.4	
Not given, not reported, or unknown	97.6	1.7	
Police reported drug involvement			< 0.01
Yes	2.2	12	
No	61.4	49.1	
Not reported	26.2	19.3	
Unknown	10.2	19.6	
Method of drug determination			< 0.01
Evidential Test	0.9	18.4	
Other	7.9	17.8	
Not reported	91.1	63.9	
No. of occupants			< 0.01
1	62.2	68.2	
2	24.1	21.2	
3	8.3	6.6	
≥ 4 or unknown	5.4	4	
Hit and run?			< 0.01
Yes	4.1	1.3	
No or unknown	95.9	98.7	
Owner of vehicle			< 0.01
Driver	59.5	59.9	
Not driver	32.7	34.7	
Company or Rental	4.7	3.6	
Not applicable or unknown	3.2	1.8	
Travel speed			< 0.01
Stopped	6.3	2.1	
1-50 mph	19.5	12	
≥ 50 mph	17.1	23.2	
Not reported or unknown	57.1	62.6	
Underride?			< 0.01
Yes or unknown	0.7	1.7	
No	99.3	98.3	
Rollover location			< 0.01
No rollover	91.7	80.1	
On roadside	5.6	14.2	
Other or unknown	2.6	5.8	
Vehicle towed?			< 0.01
Not towed	14.7	2.7	
Towed due to disabling damage	72.8	90.5	
Towed not due to disabling damage	11.1	5.6	
Not reported or unknown	1.4	1.2	
Most harmful event			< 0.01
Non collision	5.7	13.8	
Collision with vehicle	60.3	51.5	
Collision with non-vehicle	21	10.1	
Collision with fixed object	10.1	23.2	
Other or unknown	2.9	1.4	

Table G.10: Summary statistics stratified by missingness in blood alcohol concentration (BAC), continued

Variables	Missing (%)	Non-missing (%)	p-value
Any vehicle related factors?			0.04
Yes or unknown	1.4	1	
No	98.6	99	
License status			< 0.01
Licensed	86.2	84.2	
Not licensed or no driver	11.3	15.1	
Unknown	2.4	0.7	
Any license restrictions?			< 0.01
Yes	31.5	29.5	
No	65.8	69.5	
No driver or unknown	2.7	1	
Driver height			< 0.01
< 65 inches	27.4	24.8	
65-70 inches	35	39.6	
> 75 inches	22.8	28	
No driver or unknown	14.9	7.7	
Driver weight			< 0.01
< 150 pounds	20.8	22.9	
150-200 pounds	25.6	33.9	
> 200 pounds	11.4	15.7	
No driver or unknown	42.1	27.5	
No. of previous accidents			< 0.01
0	73.3	73	
1	12	13.4	
≥ 2	3.3	4.2	
No driver, not reported, or unknown	11.4	9.4	
Speed related crash?			< 0.01
Yes	12.8	26.8	
No	84.1	68.4	
No driver or unknown	3	4.8	
Trafficway description			< 0.01
One way	1.6	1.1	
Two way, divided	37.4	29	
Two way, not divided	58.5	67.5	
Entrance/exit ramp	1.4	1.5	
Non trafficway, not reported, or unknown	1.1	0.9	
No. of lanes			< 0.01
1	1.4	1.3	
2	58.4	72	
3	14.1	9.8	
4	13	8.5	
≥ 5	11.8	7.4	
Non trafficway, not reported, or unknown	1.4	1	

Table G.11: Summary statistics stratified by missingness in blood alcohol concentration (BAC), continued

Variables	Missing (%)	Non-missing (%)	p-value
Speed limit			< 0.01
≤ 25 mph	5	4.5	
30 mph	5.4	4.6	
35 mph	11.4	10	
40 mph	8.7	7.3	
45 mph	17.2	15	
50 mph	4.9	5.4	
55 mph	21.9	29.4	
60 mph	3.7	3.8	
65 mph	8.7	8.5	
≥ 70 mph	7.5	7.3	
No limit, not reported or unknown	5.7	4.1	
Road alignment			< 0.01
Straight	81.8	72.4	
Curve left	6.8	12.2	
Curve right	6.4	11.1	
Curve unknown direction	1.1	1.8	
Non trafficway, not reported, or unknown	4	2.5	
Profile			< 0.01
Uphill	3.5	4.5	
Downhill	4.2	6.5	
Grade, unknown slope	9.4	11.7	
Hillcrest or sag	2.6	3.4	
Level	71.2	67.1	
Non trafficway, not reported, or unknown	9.1	6.8	
Surface type			< 0.01
Blacktop, bituminous, or asphalt	63.2	75.7	
Concrete	7.3	8	
Other, non trafficway, not reported, or unknown	29.4	16.3	
Surface condition			< 0.01
Dry	84	81.1	
Wet	12	14.3	
Other	2.3	3	
Non trafficway, not reported, or unknown	1.7	1.6	
Traffic control device			< 0.01
Traffic signals	14.7	8.5	
Regulatory signs	9.4	12.2	
No controls, not reported, or unknown	75.9	79.3	
Pre-event movement			< 0.01
Going straight	64	60.9	
Other	35.4	38.5	
Unknown	0.6	0.6	
Pre-crash event			< 0.01
Traveling	55.4	63	
Loss of control	5.4	12.2	
Other vehicle in lane	39.3	24.7	
Attempt avoidance?			< 0.01
Yes	14.8	16.8	
No	36.2	37.3	
No driver or unknown	49	45.8	

Table G.12: Summary statistics stratified by missingness in blood alcohol concentration (BAC), continued

Variables	Missing (%)	Non-missing (%)	*p*-value
Pre-impact stability			< 0.01
Tracking	83.4	72.9	
Other	8.1	17	
No driver or unknown	8.5	10.1	
Pre-impact location			< 0.01
Stayed in original travel lane	68	40.5	
Stayed on roadway, but left original travel lane	10.3	16.4	
Stayed on roadway, not known if left original travel lane	1.7	1.1	
Departed roadway	17.2	38.5	
Other, no driver or unknown	2.8	3.5	
Crash type			< 0.01
Changing trafficway, vehicle turning	12.8	9	
Intersecting paths	9.4	8.3	
Same trafficway, opposite direction	14.9	22.4	
Same trafficway, same direction	10.9	8.8	
Single driver, Misc or no impact	52	51.5	
Driver drinking			< 0.01
Yes	4.8	36.5	
No	95.2	63.5	
Drug test results			< 0.01
Positive	3.2	32.1	
Negative	0.5	42.6	
Not tested	96.3	25.3	
Any crash factors?			< 0.01
Yes	9.9	6.2	
No	90.1	93.8	
Any driver factors?			< 0.01
Yes	35.1	53.4	
No	64.9	46.6	

BIBLIOGRAPHY

BIBLIOGRAPHY

Agarwal, R., P. Ranjan, and H. Chipman (2013), A new Bayesian ensemble of trees approach for land cover classification of satellite imagery, *Canadian Journal of Remote Sensing*, *39*(6), 507–520.

Albert, J., and S. Chib (1993), Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association*, *88*(422), 669–679.

Albert, J., and S. Chib (1996), *Bayesian modeling of binary repeated measures data with application to crossover trials*, 577-599 pp., In Bayesian Biostatistics, D. A. Berry and D. K. Stangl, eds. New York: Marcel Dekker.

Bonato, V., V. Baladandayuthapani, B. Broom, E. Sulman, K. Aldape, and K. Do (2011), Bayesian ensemble methods for survival prediction in gene expression data, *Bioinformatics*, *27*(3), 359–367.

Box, G., and D. Cox (1964), An analysis of transformations, *Journal of the Royal Statistical Society Series B*, *26*, 211–252.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984), *Classification and regression Trees*, Wadsworth, Belmont, CA.

Butrica, B., K. Smith, and E. Toder (2010), What the 2008 stock market crash means for retirement security, *Journal of Aging & Social Policy*, *22*(4), 339–359.

Cao, W., A. Tsiatis, and M. Davidian (2009), Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika*, *96*, 723–734.

Chaix, B., D. Evans, J. Merlo, and E. Suzuki (2012), Commentary: Weighing up the dead and missing: reflections on inverse probability weighting and principal stratification to address truncation by death, *Epidemiology*, *23*(1), 129–131.

Chipman, H., R. McCulloch, and V. Dorie (2015), Discrete Bayesian Additive Regression Trees Sampler, Retrieved June 26, 2016, from `https://github.com/vdorie/dbarts`

Chipman, H., E. George, and R. McCulloch (1998), Bayesian CART Model Search, *Journal of the American Statistical Association*, *93*(433), 935–948.

Chipman, H., E. George, L. Lemp, and R. McCulloch (2010a), Bayesian flexible modeling of trip durations, *Transportation Research Part B*, *44*(5), 686–698.

Chipman, H., E. George, and R. McCulloch (2010b), BART: Bayesian Additive Regression Trees, *The Annals of Applied Statistics*, *4*(1), 266–298.

Crimmins, E., J. Kim, K. Langa, and D. Weir (2011), Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the Aging, Demographics, and Memory Study, *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *66*(Suppl 1), 162–171.

Davies, A. (2015), GM Has 'Aggressive' Plans for Self-Driving Cars, Retrieved May 15, 2016, from `https://www.wired.com/2015/10/gm-has-aggressive-plans-for-self-driving-cars/`

Ding, J., et al. (2012), Feature based classifiers for somatic mutation detection in tumour-normal paired sequencing data, *Bioinformatics*, *28*(2), 167–175.

Do, D., L. Wang, and M. Elliott (2013), Investigating the Relationship between Neighborhood Poverty and Mortality Risk: A Marginal Structural Modeling Approach, *Social Science and Medicine*, *91*, 58–66.

Dorie, V., M. Harada, N. Carnegie, and J. Hill (2016), A flexible, interpretable framework for assessing sensitivity to unmeasured confounding, *Statistics in Medicine*, p. doi:10.1002/sim.6973.

Elliott, M., and R. Little (2015), Discussion of "On Bayesian Estimation of Marginal Structural Models", *Biometrics*, *71*(2), 288–291.

Elliott, M., M. Joffe, and Z. Chen (2006), A Potential Outcomes Approach to Developmental Toxicity Analyses, *Biometrics*, *62*(2), 352–360.

Fisher, G.G. and Hassan, H. and Faul, J.D. and Rodgers, W.L. and Weir, D.R. (2018), Health and Retirement Study Imputation of Cognitive Functioning Measures: 1992-2014, Retrieved July 17, 2018, from `http://hrsonline.isr.umich.edu/modules/meta/xyear/cogimp/desc/COGIMPdd.pdf`

Frangakis, C., and D. Rubin (2002), Principal Stratification in Causal Inference, *Biometrics*, *58*(1), 21–29.

Franke, R. (1982), Smooth interpolation of scattered data by local thin plate splines, *Computers and Mathematics with Applications*, *8*, 273–281.

Friedman, J. (1991), Multivariate Adaptive Regression Splines (with discussion and a rejoinder by the author), *The Annals of Statistics*, *19*(1), 1–67.

Friedman, J., T. Hastie, and R. Tibshirani (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, *33*, 1–22.

Friedman, M. (1956), *A Theory of the Consumption Function*, Princeton University Press, Princeton, NJ.

Gammermann, A. (2000), Support vector machine learning algorithm and transduction, *Computational Statistics*, *5*, 31–39.

Google (2015), What were up to, Retrieved August 26, 2015, from `http://www.google.com/selfdrivingcar/`

Green, D., and H. Kern (2012), Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees, *Public Opinion Quarterly*, *76*(3), 491–511.

Hahn, P., J. Murray, and C. Carvalho (2018), Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects, *arXiv*, p. 1706.09523v2.

Han, P., and L. Wang (2013), Estimation with missing data: beyond double robustness, *Biometrika*, *100*, 417–430.

Hanley, J., and B. McNeil (1982), The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve, *Radiology*, *143*, 29–36.

Hastie, T., and R. Tibshirani (1990), *Generalized additive models*, CRC Press: Boca Raton, FL.

Hastie, T., and R. Tibshirani (2000), Bayesian backfitting (with comments and a rejoinder by the authors), *Statistical Science*, *15*, 196–223.

Hedlund, J. (2008), Traffic safety performance measures for states and federal agencies, *Tech. rep.*, Report DOT HS 811 025, National Highway Traffic Safety Administration, Department of Transportation.

Heitjan, D., and R. Little (1991), Multiple imputation for the fatal accident reporting system, *Applied Statistician*, *40*, 13–29.

Hernández, B., S. Pennington, and A. Parnell (2015), Bayesian methods for proteomic biomarker development, *EuPA Open Proteomics*, *9*, 54–64.

Imai, K., and M. Ratkovic (2014), Covariate balancing propensity score, *Journal of the Royal Statistical Society Series B*, *76*, 243–263.

Imbens, G., and D. Rubin (2015), *Causal Inference for Statistical, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, New York, NY.

Juster, F., and J. Smith (1997), Improving the Quality of Economic Data: Lessons from the HRS and AHEAD, *Journal of the American Statistical Association*, *92*(440), 1268–1278.

Kang, J., and J. Schafer (2007), Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data, *Statistical Science*, *22*(4), 523539.

Kapelner, A., and J. Bleich (2015), Prediction with missing data via Bayesian Additive Regression Trees, *The Canadian Journal of Statistics*, *43*(2), 224–239.

Kapelner, A., and J. Bleich (2016), bartMachine: Machine Learning with Bayesian Additive Regression Trees, *Journal of Statistical Software*, *70*(4), 1–40.

Kindo, B., H. Wang, and E. Pena (2016), Multinomial probit Bayesian additive regression trees, *Stat*, *5*(1), 119–131.

Klein, T. (1986), A method for estimating posterior bac distributions for persons involved in fatal traffic accidents, *Tech. rep.*, Report DOT-HS-807-094, National Highway Traffic Safety Administration, Department of Transportation.

Korn, E., and B. Graubard (1995), Examples of Differing Weighted and Unweighted Estimates from a Sample Survey, *the American Statistician*, *49*(3), 291–295.

Krieger, N. (2001), Theories for social epidemiology in the 21st century: an eco-social perspective, *International Journal of Epidemiology*, *30*(4), 668–677.

Kropat, G., F. Bochud, M. Jaboyedoff, J. Laedermann, C. Murith, M. Palacios (Gruson), and S. Baechler (2015), Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units, *Journal of Environmental Radioactivity*, *147*, 51–62.

Lee, Y., J. Back, J. Kim, S. Kim, D. Na, H. Cheong, C. Hong, and Y. Kim (2010), Systematic review of health behavioral risks and cognitive health in older adults, *International Psychogeriatrics*, *22*(2), 174–187.

Leonti, M., S. Cabras, C. Weckerle, M. Solinas, and L. Casu (2010), The causal dependence of present plant knowledge on herbals – Contemporary medicinal plant use in Campania (Italy) compared to Matthioli (1568), *Journal of Ethnopharmacology*, *130*(2), 379–391.

Little, R., and D. Rubin (2002), *Statistical Analysis with Missing Data, 2nd ed.*, John Wiley & Sons, Inc.

Liu, C. (2004), Robit regression a simple robust alternative to logistic and probit regression. in: Gelman, a. and meng, x.l. (eds), *Applied Bayesian modelling and causal inference from incomplete-data perspectives*, *New York: Wiley*, 227–238.

Liu, Y., Z. Shao, and G. Yuan (2010), Prediction of Polycomb target genes in mouse embryonic stem cells, *Genomics*, *96*(1), 17–26.

Liu, Y., M. Traskin, S. Lorch, E. George, and D. Small (2015), Ensemble of trees approaches to risk adjustment for evaluating a hospitals performance, *Health Care Management Science*, *18*(1), 58–66.

Low-Kam, C., D. Telesca, Z. Ji, H. Zhang, T. Xia, J. Zink, and A. Nel (2015), A Bayesian regression tree approach to identify the effect of nanoparticles' properties on toxicity profiles, *The Annals of Applied Statistics*, *9*(1), 383–401.

Mchugh, M. (2015), Teslas Cars Now Drive Themselves, Kinda, Retrieved May 15, 2016, from `http://www.wired.com/2015/10/tesla-self-driving-over-air-update-live/`

Nateghi, R., S. Guikema, and S. Quiring (2011), Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes, *Risk analysis*, *31*(12), 1897–1906.

Neyman, J. (1934), On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Plassman, B., J. J. Williams, J. Burke, T. Holsinger, and S. Benjamin (2010), Systematic review: factors associated with risk for and possible prevention of cognitive decline in later life, *Annals of Internal Medicine*, *153*(3), 182–193.

Pool, L., S. Burgard, B. Needham, M. Elliott, K. Langa, and C. Mendes de Leon (2018), Association of a NegativeWealth Shock With All-Cause Mortality in Middle-aged and Older Adults in the United States, *Journal of the American Medical Association*, *319*(13), 1341–1350.

Pratola, M. (2016), Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models, *Bayesian Analysis*, *11*, 885–911.

Pratola, M., H. Chipman, J. Gattiker, D. Higdon, R. McCulloch, and W. Rust (2014), Parallel Bayesian Additive Regression Trees, *Journal of Computational and Graphical Statistics*, *23*(3), 830–852.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rässler, S. (2002), *Statistical matching: A frequentist theory, practical applications and alternative bayesian approaches.* , Lecture Notes in Statistics, Springer Verlag, New York.

Robins, J., A. Rotnitzky, and L. Zhao (1994), Estimation of Regression Coefficients When Some Regressors are not Always Observed, *Journal of the American Statistical Association*, *89*(427), 846–866.

Robins, J., M. Hernán, and B. Brumback (2000), Marginal structural models and causal inference in epidemiology, *Epidemiology*, *11*(5), 550–560.

Rockova, V., and S. van der Pas (2017), Posterior Concentration for Bayesian Regression Trees and their Ensembles, *arXiv*, p. 1708.08734.

Rosenbaum, P., and D. Rubin (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, *70*(1), 41–55.

Rubin, D. (1974), Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D., J. Schafer, and R. Subramaniam (1998), Multiple imputation of missing blood alcohol concentration (bac) values in fars, *Tech. rep.*, Report DOT-HS-808-816, National Highway Traffic Safety Administration, Department of Transportation.

Ruppert, D., M. Wand, and R. Carrol (2003), *Semiparametric regression*, Cambridge University Press: Cambridge, UK.

Sayer, J., S. Bogard, M. Buonarosa, D. LeBlanc, D. Funkhouser, S. Bao, A. Blankespoor, and C. Winkler (2011), Integrated Vehicle-Based Safety Systems Light-Vehicle Field Operational Test Key Findings Report DOT HS 811 416, Retrieved August 26, 2015, from `http://www.nhtsa.gov/DOT/NHTSA/NVS/Crash%20Avoidance/Technical%20Publications/2011/811416.pdf`

Shrira, A., Y. Palgi, M. Ben-Ezra, T. Spalter, G. Kavé, and D. Shmotkin (2011), For better and for worse: the relationship between future expectations and functioning in the second half of life, *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *66*(2), 195–203.

Smith, D., T. C. Bailey, and A. Munford (1993), Robust classification of artificial neural networks, *Statistics and Computing*, *3*, 71–81.

Sonnega, A., J. Faul, M. Ofstedal, K. Langa, J. Phillips, and D. Weir (2014), Cohort Profile: the Health and Retirement Study (HRS), *International Journal of Epidemiology*, *43*(2), 576–585.

Sparapani, R., B. Logan, R. McCulloch, and P. Laud (2016), Nonparametric survival analysis using Bayesian Additive Regression Trees (BART), *Statistics in Medicine*, *35*(16), 2741–2753.

Stuck, A., J. Walthert, T. Nikolaus, C. Büla, C. Hohmann, and J. Beck (1999), Risk factors for functional status decline in community-living elderly people: a systematic literature review, *Social Science & Medicine*, *48*(4), 445–469.

Subramaniam, R. (2002), Transitioning to multiple imputation – a new method to estimate missing blood alcohol concentration (bac) values in fars, *Tech. rep.*, Report DOT-HS-809-403, National Highway Traffic Safety Administration, Department of Transportation.

Tan, Y., M. Elliott, and C. Flannagan (2017), Development of a real-time prediction model of driver behavior at intersections using kinematic time series data, *Accident Analysis and Prevention*, *106*, 428–436.

Tan, Y., C. Flannagan, and M. Elliott (2018), Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects, *arXiv*, p. 1801.03147.

Tanner, M., and W. Wong (1987), The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, *82*(398), 528–540.

Twala, B., M. Jones, and D. Hand (2008), Good Methods for Coping with Missing Data in Decision Trees, *Pattern Recognition Letters*, *29*(7), 950–956.

van der Laan, M., and E. C. Polley (2010), Super Learner in Prediction, *U.C. Berkeley Division of Biostatistics Working Paper Series*, *Working Paper 266*, `http://biostats.bepress.com/ucbbiostat/paper266`

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

Zhang, G., and R. Little (2009), Extensions of the Penalized Spline of Propensity Prediction Method of Imputation, *Biometrics*, *65*, 911–918.

Zhang, J., and W. Härdle (2010), The Bayesian Additive Classification Tree applied to credit risk modelling, *Computational Statistics and Data Analysis*, *54*(5), 1197–1205.

Zhang, S., Y. Shih, and P. Müller (2007), A Spatially-adjusted Bayesian Additive Regression Tree Model to Merge Two Datasets, *Bayesian Analysis*, *2*(3), 611–634.

Zhou, H., M. Elliott, and T. Raghunathan (2016), Multiple imputation in two-stage cluster samples using weighted finite population bayesian bootstrap, *Journal of Survey Statistics and Methodology*, *4*, 139–170.

Zhou, T., M. Elliott, and R. Little (2018), Penalized Spline of Propensity Methods for Treatment Comparison, *Journal of the American Statistical Association*, p. In press.