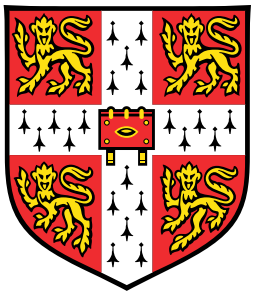
Essays on Tree-based Methods for Prediction and Causal Inference

Eoghan Patrick O'Neill



Faculty of Economics
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Christ's College July 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Summary

Essays on Tree-based Methods for Prediction and Causal Inference

Eoghan Patrick O'Neill

The first chapter of this thesis contains an application of causal forests to a residential electricity smart meter trial dataset. Household specific estimates are obtained for the effect of a Time-of-Use pricing scheme on peak demand. The most and least responsive households differ across education, age, employment status, and past electricity consumption. The results suggest that past consumption information is more useful than pre-trial survey information, which includes building characteristics, household characteristics, and responses to appliance usage questions.

The second chapter explores new variations of Bayesian tree-based machine learning algorithms. Bayesian Additive Regression Trees (BART) (Chipman et al. 2010) and Bayesian Causal Forests (BCF) (Hahn et al. 2020) are state-of-the-art machine learning methods for prediction and causal inference. A number of existing implementations of BART make use of Markov Chain Monte Carlo algorithms, which can be computationally expensive when applied to high-dimensional datasets, do not always perform well in terms of mixing of chains, and have limited parallelizability.

The second chapter introduces four variations of BART that do not rely on MCMC:

1. An improved implementation of the existing method BART-BMA (Hernandez et al. 2018), which averages over sum-of-tree models found by a model search algorithm, performs well on high-dimensional datasets, and produces more interpretable output than other BART implementations because the output includes a comparatively small number of sum-of-tree models. Improvements are made to the model search algorithm, calculation of predictions, and credible intervals.

2. A treatment effect estimation algorithm that combines the model structure of BCF with the implementation of BART-BMA (BCF-BMA). This method successfully accounts for confounding on observables using the BCF parameterization, while retaining the parsimonious model selection approach of BART-BMA.

3. A simple alternative BART implementation algorithm that uses importance sampling of models (BART-IS). This approach contrasts with existing MCMC and model-search based approaches in that BART-IS makes fast data-independent draws of many sum-of-tree models. The advantages of this approach are that it is straightforward to implement, fast, and trivially parallelizable.

4. Bayesian Causal Forests using Importance Sampling (BCF-IS). This is a combination of the BCF model framework with the BART-IS implementation. BART-IS and BCF-IS exhibit comparable performance to BART-MCMC and BCF across a large number of simulated datasets.

The second chapter also includes some illustrative applications. The methods are extendable to multiple treatments, multivariate outcomes, and panel data methods.

The third chapter of this thesis describes how the methods introduced in the second chapter can be generalized from regression and treatment effect estimation for continuous outcomes, to a range of models with various link functions and outcome variables. As examples of how to apply the general approach, Logit-BART-BMA and Logit-BART-IS are introduced with illustrative applications.

Acknowledgements

I would like to thank my supervisor, Dr. Melvyn Weeks, who coauthored the first chapter of this thesis, for his guidance throughout my PhD studies. I am indebted to Dr. Weeks for his advice and regular feedback on my research. I also acknowledge the assistance of my research advisor, Dr. Debopam Bhattacharya, Prof. Alexey Onatskiy, and the Econometrics Research Group. I am grateful for feedback received from members of the Cambridge Energy Policy Research Group. I would also like to extend my thanks to Dr. Belinda Hernandez and Prof. Andrew Parnell for providing replication code from a research publication relevant to the second chapter of this thesis, and for giving feedback on improvements and bug fixes that were essential to the implementation of the methods introduced in the second and third chapters of this thesis.

I greatly appreciate the assistance the Economics Faculty IT staff in providing software assistance and making virtual machines available for running of code. I also thank the Economics Faculty administrative staff for their assistance. I am very grateful for the financial support extended to me by Christ's College and the Cambridge Economics Faculty Trust Funds. Finally, I thank my parents and my brother for their encouragement and support.

Contents

Chapter 1

Causal Forest Estimation of Heterogeneous Household Response to Time-Of-Use Electricity Pricing Schemes

Abstract

We examine the household-specific effects of the introduction of Time-of-Use (TOU) electricity pricing schemes. Using a causal forest (Wager & Athey 2018, Athey et al. 2019), we consider the association between past consumption and survey variables, and the effect of TOU pricing on household electricity demand. We compare averages of variables across quartiles of estimated demand response. Households that are younger, more educated, and that consume more electricity, are predicted to respond more to a new pricing scheme. In addition, variable importance measures suggest that some aspects of past consumption information may be more useful than survey information in producing these estimates.[1]

1.1 Introduction

If a policymaker believes the impacts of a particular policy are heterogeneous in a given population, then it is helpful to describe the distribution of household-specific effects of the policy. The critical question is: does the policymaker know ex ante which characteristics of individuals are driving the differences in the impact of the policy?

Researchers often describe subpopulations that are of interest a priori, and which can be defined by a known combination of covariates. However, increasingly researchers have many covariates at their disposal and it may not be clear which covariates should be used to categorise heterogeneity, nor what functional form best describes the association between these covariates and treatment effects.

The introduction of an electricity pricing scheme is an example of a policy with heterogeneous effects. Consumers in different socioeconomic groups and with distinct historical intra-day load profiles and behavioural characteristics, may respond differently to the introduction of tariffs that charge different prices for electricity at different times of the day. Customers who can (cannot) adapt their consumption profile to TOU tariffs will accrue a benefit (cost). Those who consume electricity at more expensive peak periods, and who are unable to change their consumption patterns, could end up paying significantly more.

In assessing whether demographic variables are informative in terms of the impact of TOU tariffs on load profiles, the Customer-Led Network Revolution project (Sidebotham 2015) noted

> .. a relatively consistent average demand profile across the different demographic groups, with much higher variability *within* groups than *between* them. This high variability is seen both in

[1] This chapter is co-authored with Dr. Melvyn Weeks.

total consumption and in peak demand.

In addition, the question of which demographic variables are important when considering the impact of energy policies ignores the fact that many of these variables should be considered together, in a multiplicative fashion. One reason for this finding might be that it is the (unknown) combination of income, household size, education, and daily usage patterns that describes a particularly responsive or unresponsive group.

In this paper we consider the household-specific effects on customers following the introduction of a Time-of-Use (TOU) pricing scheme where the price per kWh of electricity usage depends on the time of consumption. The pricing scheme is enabled by smart meters, which records consumption every half-hour. Using machine learning methods, we describe the association between the effect of TOU pricing schemes on household electricity demand and a range of variables that are observable before the introduction of the new pricing schemes.

We apply a recently developed method, known as a causal forest, which aggregates over estimates from causal trees (Athey & Imbens 2016, Wager & Athey 2018, Athey et al. 2019). This method searches across covariates for good predictors of heterogeneous treatment effects. A causal tree provides an interpretable description of heterogeneity, and causal forests can be used to obtain individual-specific estimates of treatment effects. Heterogeneous effects are described by Conditional Average Treatment Effect (CATE) estimates, which are the expected effects of a treatment for individuals in subpopulations defined by covariates. We characterize the most and least responsive households by applying the methods described by Chernozhukov, Demirer, Duflo & Fernandez-Val (2017).

Given that policy makers are often interested in the factors underlying a given prediction, it is desirable to gain some insight to which variables in the large set of covariates are most often selected. A key challenge follows from that fact that partitions generated by tree-based methods are sensitive to subsampling, whereas the use of an ensemble method such as causal forests produces more stable, but less interpretable estimates.

To address this problem we utilise variable importance measures to consider which variables are chosen most often by the causal forest algorithm. However, in the estimation of variable importance it is important to account for the impact of the varying information content across continuous versus discrete random variables. In particular, tree based methods can be biased towards continuous variables, given the presence of more potential splitting points. We address this issue by including permutation-based tests for our variable importance results. This is particularly important for this analysis given that many of our demographic variables are either binary or categorical.

In section 1.2 we first describe the potential outcomes framework and conditional average treatment effects, then describe causal trees and causal forests. We describe the variable importance measures and outline how we will apply the methods of Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) to describe heterogeneity between the most and least demand responsive households. In section 1.3, we introduce the application to electricity smart meter data, and review existing literature. In section 1.4, we present the results. Section 1.5 concludes.

1.2 Methods for Estimation of Heterogeneous Treatment Effects

The estimand is defined using the potential outcomes framework introduced by Neyman (1923) and developed by Rubin (1974). Let X_i be a vector of covariates for individual i. Suppose that there is one treatment group of interest. $Y_i(1)$ ($Y_i(0)$) denotes the potential outcome if individual i is allocated to the treatment (control)

group. The causal effect of a treatment on individual i is therefore $Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that we do not observe the causal effect for any i (Holland 1986).

The estimand that we consider is the Conditional Average Treatment Effect (CATE), also referred to as the Individual Treatment Effect (ITE)

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]. \tag{1.1}$$

Whereas the ATE can be estimated by a difference in means $\bar{y}_t - \bar{y}_c$, where $\bar{y}_t$ ($\bar{y}_c$) is the mean of the outcome variable for the treated (control) group, the CATE can be thought of as a subpopulation average treatment effect.[2] [3] The CATE is identified under unconfoundedness, i.e. $Y_i(1), Y_i(0) \perp T_i|X_i$, and overlap, i.e. $0 < \Pr(T_i = 1|X_i = x) < 1 \; \forall \; x$, where T_i denotes the treatment indicator variable.

A CATE estimate can be obtained from a linear model by including interactions between the treatment indicators and the conditioning variable(s) of interest. The inclusion of interaction terms in a linear model is a common technique for exploring the heterogeneity of treatment effects in areas ranging from biomedical science to the social sciences.[4]

It is possible to search for heterogeneity in treatment effects simply by separately estimating CATEs using many possible conditioning variables and repeatedly estimating the standard linear regression model, and conducting tests of multiple hypotheses. However, a clear problem is false discovery and the need to adjust significance levels for multiple hypothesis testing which can limit the power of a test to find heterogeneity.

A number of alternative machine learning methods allow the researcher to explore more complex forms of heterogeneity. Recent methods for ITE estimation include LASSO (Imai et al. 2013, Weisberg & Pontes 2015, Tian et al. 2014), BART (Hill 2011, Hahn et al. 2017, Logan et al. 2019), other tree-based methods (Powers et al. 2017, Oprescu et al. 2018, Lu et al. 2018, Lechner 2019), the R-learner (Nie & Wager 2017*b*), neural networks (Shalit et al. 2017, Farrell et al. 2018, Atan et al. 2018, Shi et al. 2019), Generalized Adversarial Networks (Yoon et al. 2018), and many more.

In this study we are interested in allowing for many possibly nonlinear interactions between covariates. Forest methods perform well in capturing nonlinear interactions. Furthermore, causal forests perform reasonably well relative to other methods and have known asymptotic properties (Knaus et al. 2018, Alaa & Van Der Schaar 2019, Athey et al. 2019). Therefore we apply the causal forest method for ITE estimation.

Regression and Causal Trees

Causal forests (Wager & Athey 2018, Athey et al. 2019) average the predictions of many causal trees (Athey & Imbens 2016). Causal trees are decision trees for treatment effect estimation, and can be viewed as a variation on standard regression trees, with a different splitting criterion, and different terminal node estimates.

A single regression tree is constructed as follows (Friedman et al. 2009, Breiman et al. 1984). Suppose there are p covariates and N observations. The objective is to partition the covariate space $\mathbb{X}$ into M mutually exclusive regions $R_1, ..., R_M$, where the outcome for an individual with covariate vector x in region R_m is estimated as the mean of the outcomes for training observations in leaf R_m. The following algorithm

[2] In instances where we condition on x being in some subset of the covariate space, i.e. $x \in \mathbb{A} \subset \mathbb{X}$, and $\tau_{\mathbb{A}} = \mathbb{E}[Y_i(1) - Y_i(0)|x \in \mathbb{A}]$, we also refer to this as the CATE (with suitably re-defined covariates).

[3] Another estimand is the average treatment effect conditional upon observed covariates $\bar{\tau} = \frac{1}{N}\sum_{i=1}^{N} \tau(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x_i]$. Imbens & Rubin (2015) refer to this as the conditional average treatment effect, but we shall use the above definition of the CATE.

[4] A description of the application of linear regression methods for the purpose of estimating treatment effects in randomized experiments can be found in Athey & Imbens (2017).

is used to apply binary splits of the data:

Let X_j be a splitting variable and s be a split point. Define $R_1(j,s) = \{X|X_j \leq s\}$ and $R_2(j,s) = \{X|X_j > s\}$.[5] The algorithm selects the pair (j,s) that solves:

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_1(j,s))^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_2(j,s))^2 \right] \tag{1.2}$$

where $\bar{y}_1(j,s)$ and $\bar{y}_2(j,s)$ are the mean outcomes in $R_1(j,s)$ and $R_2(j,s)$ respectively. When the data has been split into two regions, the same process is applied separately to each region. Then the process is repeated on each of the four resulting regions, and so on.

Trees can be fully grown, or grown up to a stopping rule, or a penalty term can be included in the splitting criterion that penalizes the tree size (Friedman et al. 2009). Causal tree (Athey & Imbens 2016) leaf estimates are differences in means between treated and untreated observations, and the splitting criterion is different to (1.2) because the goal is to minimize the expected mean square error of these treatment effect estimates.

Athey & Imbens (2016) also note that estimates produced by standard regression tree algorithms are biased because the same data is used for tree construction and for estimating the terminal node means. Athey & Imbens (2016) therefore suggest separating the training data into two independent subsamples, one for construction of the tree, and one for estimation of the terminal node means. This so-called *honest* estimation ensures unbiased estimates.

Random and Causal Forests

The prediction of a random forest (Friedman et al. 2009) is the average of many (B) unpruned regression trees. Each tree, T_b (b indexing the bootstrap samples), is produced using a bootstrap sample of size N without replacement from the training data. At each split in the tree, the algorithm uses a random subset of the set of all covariates as potential splitting variables. Each tree is fully grown up to a minimum leaf size.

The prediction for an individual with a vector of covariates x is then $\frac{1}{B}\sum_{b=1}^{B} T_b(x)$, where $T_b(x)$ is the estimate produced by tree b. The trees are not independent because two bootstrap samples can have some common observations, and therefore the correlation between trees limits the benefits of averaging. However, this correlation is reduced through the random selection of the input variables.

Similar aggregations over causal trees, known as causal forests, can improve the accuracy of treatment effect estimates. Wager & Athey (2018) outline the properties of causal forests and show that, under certain assumptions, the predictions from causal forests are asymptotically normal and centred on the true treatment effect for each individual. Recent applications of causal forests can be found in articles by Davis & Heller (2017*a*,*b*) and Bertrand et al. (2017).

Athey et al. (2019) introduce a generalization of random forests which can be viewed as an adaptive kernel method. This generalized random forest (GRF) framework can be used for estimation of a variety of models, including treatment effect estimation. The causal forest method introduced by Wager & Athey (2018) is almost equivalent to the GRF implementation of a causal forest without centering. GRF involves an approximate, gradient-based loss criterion, and orthogonalizes the outcome treatment variables from estimates produced by separate forests before fitting the causal forest.

[5] If a splitting variable is categorical with q unordered values, then we can consider all $2^{q-1} - 1$ possible splits of the q values into two groups, or we can use binary variables for each category.

Variable Importance

A general issue which applies to standard regression trees and random forests is the trade-off between interpretability and stability. A single causal tree splits the data into relatively few leaves. The results are easy to interpret given that a simple tree diagram allows the researcher to quickly identify the subgroup to which any household belongs by following a set of decision rules. Breiman (2001) and Strobl (2008) note that single trees can be unstable with small changes in the training data resulting in a very different model (tree). However, although stable forests generate better predictive performance, the interpretability of a single tree is lost when we move to an ensemble method, such as a causal forest.

Across the many trees within a forest, it is not immediately clear what covariates most strongly influence the final estimates, and how different covariates interact. Variable importance measures describe which variables are chosen most often by the causal forest algorithm. However, in the estimation of variable importance it is important to account for the impact of the varying information content across continuous versus discrete random variables. In particular, tree based methods can be biased towards continuous variables, given the presence of more potential splitting points. We address this issue by including permutation-based tests for our variable importance results.

We apply the default variable importance measure for the command **causal_forest** in the **R** package **grf**. This variable importance measure is based upon a count of the proportion of splits on the variable of interest up to a depth of 4, with a depth-specific weighting.[6]

$$imp(x_j) = \frac{\sum_{k=1}^{4} \left[\frac{\sum_{all\ trees} number\ depth\ k\ splits\ on\ x_j}{\sum_{all\ trees} total\ number\ depth\ k\ splits} \right] k^{-2}}{\sum_{k=1}^{4} k^{-2}} \tag{1.3}$$

Permutation Test for Causal Forest Variable Importance

Following the method of Altmann et al. (2010) for random forests,[7] and Bleich et al. (2014) for BART, we compute "p-values" for the variable importances. The adjusted importance measures are referred to as "p-values" but do not truly have the properties of p-values for a test of the null hypothesis of conditional independence. Therefore these measures can be viewed as corrected variable importances (which take lower values for more important variables), but should not be considered as reflective of rigorous hypothesis testing (Nembrini 2019). The calculation of "p-values" involves permuting the dependent variable 1000 times and obtaining variable importances for all variables from 1000 causal forests fitted separately using the 1000 permutations as dependent variables. The variable importances are also obtained from a causal forest using the original, unpermuted dependent variable. Then, using the "local" test described by Bleich et al. (2014), we obtain a "p-value" for each variable by finding the proportion of the 1000 causal forests for which the variable had a greater variable importance measure than that obtained from the causal forest with the unpermuted dependent variable.

If the splits in trees spuriously occur (in the sense that variables might not be as important, or strongly associated with the outcome, as suggested by the number of splits) more often on continuous variables and variables with more categories, then this should also occur when the dependent variable is permuted. In this

[6] Variable importances for categorical variables are the sum of the variable importances of binary variables. The parameters we set for the **causal_forest** command are: 15000 trees, bootstrap samples of half the data, one third of covariates randomly drawn as potential splitting variables, and minimum node size of 5.

[7] Altmann et al. (2010) show that "p-values" based on permutation of the dependent variable can address the issues of bias towards variables with more categories, and masking of the importance of groups of highly correlated variables.

instance, the "p-value" should be unaffected unless the extent of the over-selection of variables for splitting is also dependent on the true importance of the variables. We investigate this issue in further detail in Appendix A, which contains a simple simulation study of this permutation based variable importance test. The simulations suggest that the "p-values" are potentially unaffected by the bias of variable splitting towards variables with more possible splitting points.

Nembrini (2019) investigates the properties of permutation based variable importance tests for random forests, and notes the limitations of permutation of the dependent variable as a method for hypothesis testing. The "p-values" obtained from the permutation-based "test" should be viewed as corrected importance measures, rather than interpreted as actual p-values for a hypothesis test.

Testing and Describing Treatment Effect Heterogeneity

In the results section we provide point ITE estimates with confidence intervals obtained using the methods described by Athey et al. (2019). However, we are also interested in describing heterogeneity through *features* of the treatment effect function $\tau(x)$ (Equation 1.1), which requires a different approach to inference (Chernozhukov, Demirer, Duflo & Fernandez-Val 2017) involving repeatedly obtaining two random subsamples, and training a causal forest on one subsample, and performing a statistical test of interest on the other subsample. A sample split allows for valid inference conditional on the subsample of data used for constructing the causal forest, and repeated sample splitting is used in accounting for the uncertainty induced by the random sampling. This requires the training of many causal forests, in contrast to the requirement of a single causal forest for valid ITE prediction intervals.

We apply the methods of Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) to first test for the presence of heterogeneity, and then characterize the association between covariates and demand response. This approach, summarized below, involves repeated data splitting to avoid overfitting and to achieve validity.

Let Y be the outcome variable, D be the treatment indicator variable, and Z be all other covariates. We split the data in half into a main sample $Data_M$ and auxiliary sample $Data_A$ 1000 times. For each split we train a causal forest on $Data_A$ and also a regression forest on the untreated observations in $Data_A$. Then we obtain treatment effect estimates, $S(Z)$ by applying the trained causal forest to $Data_M$, and we obtain baseline outcome estimates, $B(Z)$ by applying the trained regression forest to $Data_M$.[8] This will result in 1000 sets of parameter estimates that can be used for valid inference on the parameters. See below for a description of the parameters of interest, and see Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) for a description of the inference methods. This approach accounts for estimation uncertainty conditional on the auxiliary sample and splitting uncertainty induced by random partitioning of the data into main and auxiliary samples. (Chernozhukov, Demirer, Duflo & Fernandez-Val 2017).

First, we test for heterogeneity using the Best Linear Predictor (BLP) of the CATE (Chernozhukov, Demirer, Duflo & Fernandez-Val 2017). We obtain the following estimated model by weighted OLS:

$$Y_i = \hat{\alpha_0} + \hat{\alpha_1} B(Z_i) + \hat{\beta}_1 (D_i - p(Z_i)) + \hat{\beta}_2 (D_i - p(Z_i))(S(Z_i) - \overline{S(Z)}) + \hat{\varepsilon}_i \tag{1.4}$$

where the weights are $\{\hat{p}(Z)(1-\hat{p}(Z))\}^{-1}$ and $\overline{S(Z)} = |M|^{-1} \sum_{i \in M} S(Z_i)$. For the randomized controlled trial dataset used in this paper, we set $\hat{p}(Z)$ equal to the sample proportion of treated individuals.

The parameter β_2 reflects the extent to which the estimated treatment effect is a proxy for the true treatment effect function (1.1). Rejection of the null hypothesis $\beta_2 = 0$ implies that there is heterogeneity

[8] $B(Z)$ is included to improve efficiency. Inference would still be valid if we removed $B(Z)$ from equations 1.4 and 1.5.

and $S(Z)$ is a relevant predictor. Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) outline how to perform valid inference on β_2. For each of the 1000 data splits into main and auxiliary samples, we keep the estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and upper and lower bounds of 95% confidence intervals. The medians of the $\hat{\beta}_1$ and $\hat{\beta}_2$ are the final β_1 and β_2 estimates. Similarly, medians of upper and lower bounds define the confidence intervals.[9] The confidence level of the final interval is 90%, and accounts for splitting uncertainty.

We use the estimated treatment effects $S(Z)$ to divide the main sample into groups G_1 to G_4, where G_1 is the 25% of the data that has the lowest (i.e. most negative) treatment effect estimates and G_4 has the highest treatment effect estimates. Sorted Group Average Treatment Effect (GATE) estimates (Chernozhukov, Demirer, Duflo & Fernandez-Val 2017) are obtained from the following estimated model:

$$Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 B(Z_i) + \sum_{k=1}^{K} \hat{\gamma}_k \mathbb{I}(G_k) + \hat{\epsilon}_i \tag{1.5}$$

where $\mathbb{I}(G_k) = 1$ if individual i is in group G_k and 0 otherwise. The weights are the same as in equation 1.4. Inference on γ_k and the difference $\gamma_4 - \gamma_1$ is made using the same approach as for β_2 in (1.4).

Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) also outline how to perform valid inference on the average of any function of the outcome and pre-trial covariates, $g(Y, Z)$, within group G_k, and differences in these averages between groups G_1 and G_4. This is referred to as Classification Analysis (CLAN) and we utilise this method to test for differences in the outcome and pre-trial covariates between the most and least affected 25%.

In summary, the methods of Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) allow us to test for the existence of heterogeneity; test the relevance of our ITE estimates; and to characterise heterogeneity in the treatment effects by describing the most and least affected individuals.

1.3 Heterogeneity of Household Electricity Demand Response

TOU tariffs are becoming more implementable through the roll-out of smart metering technology. The subsequent increase in the availability of large amounts of past electricity consumption data allows for more household specific targeting of electricity pricing and other demand stimuli. Furthermore, in a world where energy suppliers rely increasingly on renewables which are intermittent in nature, measures to reduce peak demand are required as part of the need to balance supply and demand. Understanding heterogeneity in household responses to TOU pricing is of interest to both regulators and retailers.

The British energy regulator, Ofgem (2013), is interested in the impact of new pricing schemes upon vulnerable and low income customers. Faruqui et al. (2010) postulate that two potentially offsetting forces influence how we expect low-income customers to be impacted differently by new electricity pricing schemes. First, lower income customers can have a greater proportion of their demand in off-peak hours, and therefore can benefit from TOU pricing without adjusting their daily demand profile. Second, we might not expect these customers to shift and reduce load as much as other customers because they have lower usage levels in general and less discretionary usage. Faruqui et al. (2010) confirm these hypotheses using US data, and find that low income customers change their electricity usage less than higher income customers.

Counter to some of this evidence, studies by Lower Carbon London (Schofield et al. 2014) and Fron-

[9] The final upper bound is the lower median of the 1000 upper bounds, and the final lower bound is the upper median of the 1000 lower bounds. The final estimates of β_1 and β_2 are mid-points of lower and upper medians of $\hat{\beta}_1$ and $\hat{\beta}_2$ (Chernozhukov, Demirer, Duflo & Fernandez-Val 2017).

tier Economics and Sustainability First (DECC 2012) have noted the generally low associations between demographic variables and demand response, and in particular, the lack of evidence pertaining to differing responses of low-income and vulnerable customers. One possible reason for this is that individuals most affected by energy policies might be identified through the interaction of a number of variables. For example, the Centre for Sustainable Energy produced a report (Preston et al. 2013) which used interactions of variables to define the groups of households predicted to face the largest increase in household bills as a result of changes in energy policy.

In this study we examine the importance of variables constructed from historic load profiles. Relatively few studies have conditioned upon past usage data when estimating treatment effects of electricity pricing schemes. Some recent examples include a study using US data by Harding & Lamarche (2016), who split the sample into low, medium, and high usage customers. The results suggest that high usage customers decrease peak usage to a greater extent, which is expected because these customers have more reducible usage. However, surprisingly low-income customers appear to increase consumption in off-peak time periods. The authors note that this substantial load-shifting by low-income customers demonstrates the difficulty in anticipating the impact of new pricing schemes for some customer segments. A number of recent studies have used past electricity usage data for the estimation of household-specific treatment effects. Bollinger & Hartmann (2015) condition upon the empirical distribution of past electricity usage and consider how a utility can gain from targeting based upon ITE estimates. Balandat (2016) estimates ITEs by comparing predictions of electricity usage under control group allocation to realised usage under treatment allocation during the trial period.

Data

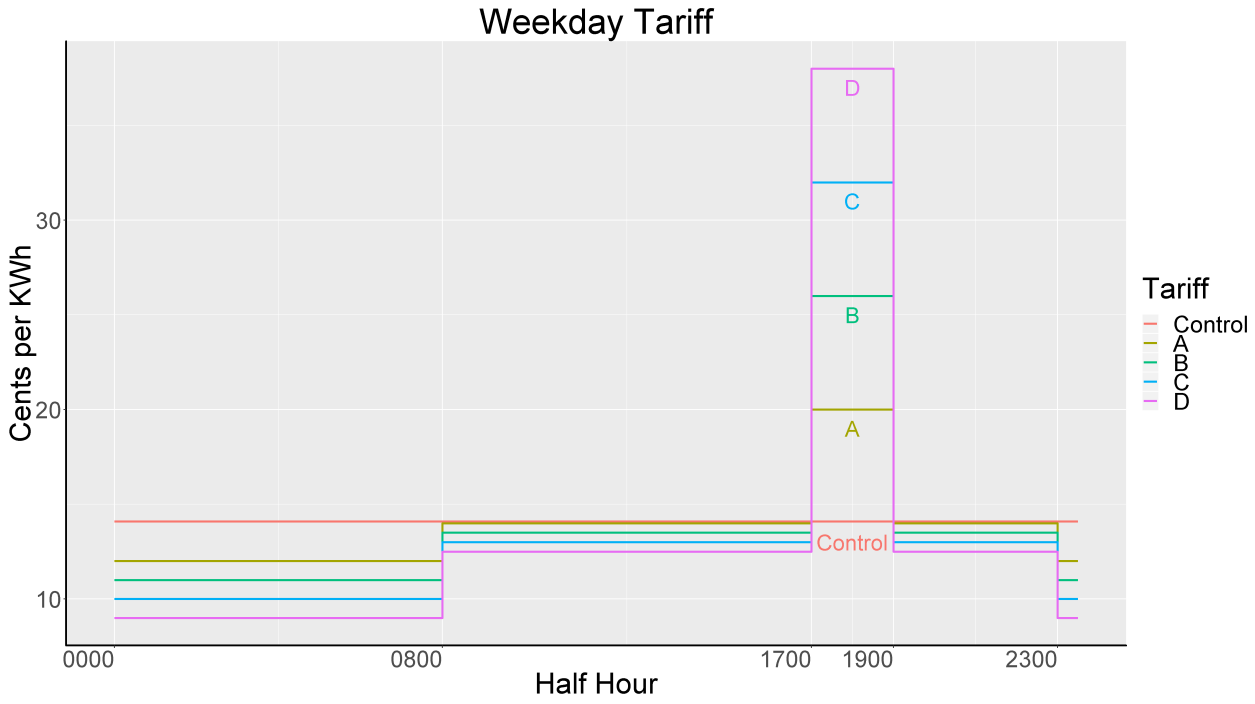
The dataset used in this project is from the Electricity Smart Metering Customer Behavioural Trial conducted by the Irish Commission for Energy Regulation (CER 2011). The CER note that this is "one of the largest and most statistically robust smart metering behavioural trials conducted internationally to date" (CER 2011). The dataset consists of half hourly residential electricity demand observations for 4225 households over 536 days. The benchmark period began on 14th July 2009 and ended on 31st December 2009. Households were then randomly allocated to either a control group or various TOU Pricing Schemes and Demand Side Management stimuli from 1st January 2010 to 31st December 2010.

All households were charged the normal Electric Ireland tariff of 14.1 cents per kWh during the benchmark period. During the trial period the control group remained on the tariff of 14.1 cents per kWh while the test group were allocated to tariffs A, B, C, or D.[10] The tariffs A to D were structured as shown in Table 1.1, and are graphed in Figure 1.1a.

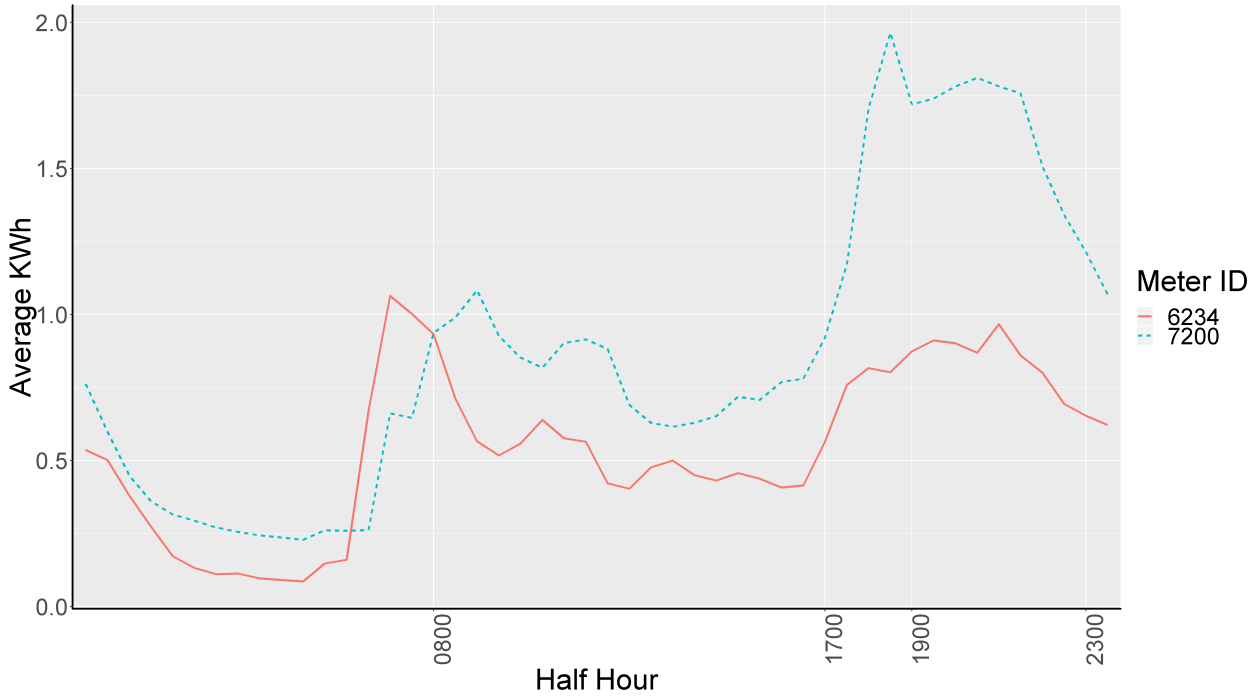
Households in the test group were also allocated to one of the following Demand Side Management (DSM) stimuli: Bi-monthly detailed Bill; Monthly detailed bill; Bi-monthly detailed bill and In-Home Display (IHD); Bi-monthly detailed bill and Overall Load Reduction (OLR) incentive.

The identification of ATEs depends upon unconfoundedness and overlap. The CER took a number of steps to ensure that the samples for treatment groups were representative and did not exhibit notable biases. A stratified random sampling framework was used with phased recruitment. Non-respondents and attriters were surveyed and adjustments were made accordingly. Those who opted in were compared to the national profile. The full dataset contains 4225 households, with 768 households in the control group and 233 households

[10] There was also a Weekend tariff group, which we exclude from this study.



(a) Trial period TOU tariffs



(b) Pre-trial average half-hourly demand for two households

Figure 1.1: Prices and examples of demand profiles

Table 1.1: TOU Tariff details

TOU Tariffs (cents per kWh)	**Night** 23.00-08.00	**Day** 08.00-17.00 every day 19.00-23.00 every day 17.00-19.00 weekends and holidays	**Peak** 17.00-19.00 Mon-Fri Excluding holidays
Tariff A	12.00	14.00	20.00
Tariff B	11.00	13.50	26.00
Tariff C	10.00	13.00	32.00
Tariff D	9.00	12.50	38.00

facing the combination of tariff C and IHD stimulus, which will be the treatment group of interest in this study.

Figure 1.1b gives an example of average half hourly usage on weekdays before the trial period for households with similar survey responses. The two households both have four people in a 3 bedroom semi-detached house, in which the chief earner is an employee and lower middle class with 3rd level education. Both households also typically have one person at home during the day, own their home, have timed oil heating, and have a similar stock of appliances. This figure shows that even households that are similar across multiple characteristics do not necessarily have the same patterns of demand use. Therefore these type of survey variables are of limited use in describing demand heterogeneity.[11]

1.4 Results

The outcome variable is average half-hourly peak time electricity consumption during the trial period (measured in kWh), excluding weekends. We restrict attention to Tariff C in combination with the In-Home Display (IHD). The IHD stimulus is of greater interest than the other information stimuli, and tariff C has a high ratio of peak to off-peak prices and more observations than any other tariff combined with the IHD.[12]

The standard ATE estimates for the tariff C with IHD range from -0.073 to -0.092 kWh for an average peak half hour, depending on the set of controls.[13] Mean half-hourly peak consumption for the control group during the trial period (one full year) was 0.799 kWh, and mean peak consumption for all households during the pre-trial period (half a year) was 0.828 kWh. Therefore these treatment effects are of the order of 10% of peak consumption.

Below we present two estimates of single causal trees as an example of the instability of single tree estimates and small sample size. Causal forest Individual Treatment Effect (ITE) estimates are then plotted with confidence intervals and summarized in density plots. We also apply the methods described by Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) to test the hypothesis $\beta_2 = 0$ in equation (1.4), and confirm that there is heterogeneity of treatment effects and that Causal forest Individual Treatment Effect (ITE) estimates are relevant predictors of the true ITEs. We test the association between Causal forest Individual Treatment Effect (ITE) estimates and a set of pre-trial variables using the approach of Chernozhukov, Demirer, Duflo & Fernandez-Val (2017). Finally, variable importance measures are presented in order to consider which variables are the strongest determinants of the structure of the trees in the forest.

[11] We make use of pre-trial survey data, but we cautiously avoid using post-trial survey information. Prest (2017) applies a causal tree method to this data, but the estimates are potentially biased by conditioning on post-trial survey information. Our methods also differ from those of Prest (2017) in that we use a causal forest.

[12] 343 households were allocated to Tariff C with the IHD. Only 126 households were allocated to tariff D with the IHD.

[13] These results are obtained by linear regression of average peak usage on the treatment indicator.

Causal Trees

Figures 1.2a and 1.2b show estimated causal trees.[14] The set of potential splitting variables is given in Table 1.2.[15]

Table 1.2: Potential splitting variables for Causal Trees and Causal Forest

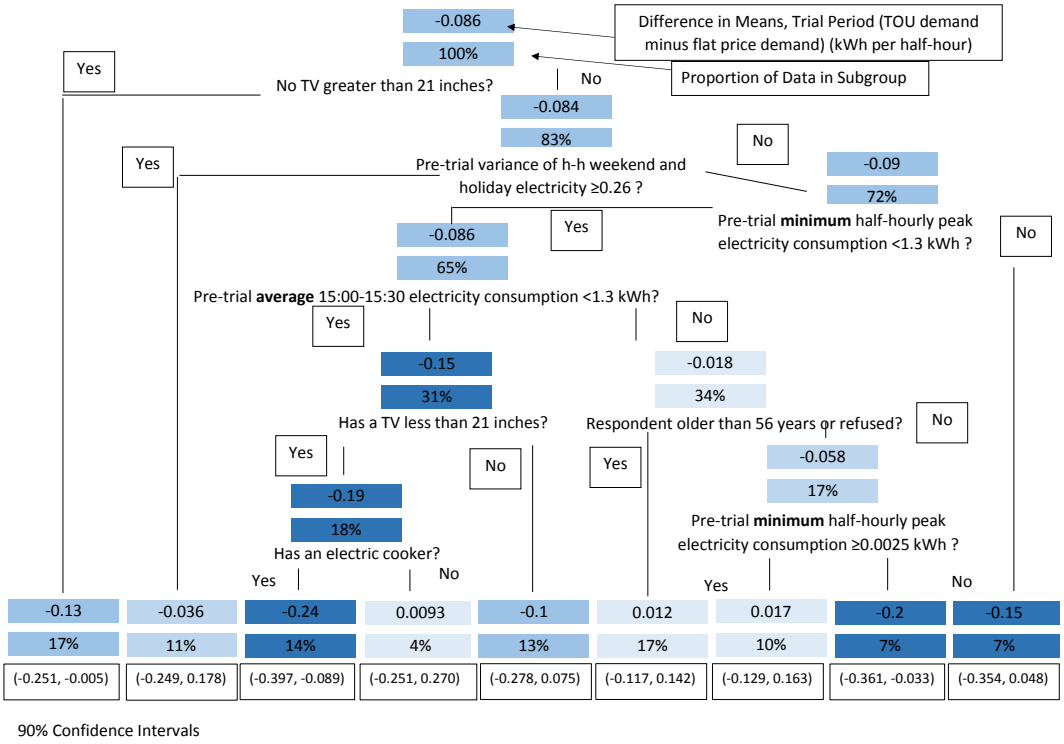
Name of variable	
Survey variables (categorical)	
Age of respondent	Sex of respondent
Class of chief income earner	Regular internet use
Employment status of chief income earner	Other reg. internet users
Number of bedrooms	Education of chief earner
Type of home	Electric central heating
Alone or other occupants	Electric plugin heating
Own or rent the home	Central water heating
Number of electric cookers - number	Immersion water heating
Internet access	Instant water heating
Approximate age of home	Number of washing machines
Lack money for heating	Number of tumble dryers
Number of dishwashers	Number of instant electric showers
No. showers elec. pumped from hot tank	Type of cooker
Number of plug-in convector heaters	Number of freezers
Number of water pumps or electric wells	Number of immersion water heaters
Number of small TVs	Number of big TVs
Number of desktop PCs	Number of laptop PCs
Number of games consoles	Has an energy rating
Proportion of energy saving lightbulbs	Prop. double glazed windows
Lagging jacket	Attic insulation
External walls insulated	
Electricity usage variables (continuous)	
Mean usage	Min. usage
Variance of usage	Max. usage
Mean peak usage	Mean nonpeak usage
Variance of peak usage	Variance of nonpeak usage
Mean night usage	Mean daytime usage
Variance of night usage	Variance of daytime usage
Mean usage - weekdays	Mean peak usage - weekdays
Variance of usage - weekdays	Var. peak usage - weekdays
Mean night usage - weekdays	Mean daytime usage - weekdays
Variance of night usage - weekdays	Var. daytime usage - weekdays
Mean daily maximum usage	Mean usage - weekends
Mean daily minimum usage	Variance of usage - weekends
Mean of half-hour coefficients of variation	Mean usage - each month (July-Dec)
Avg. night usage/ avg. daily usage	Var. of usage - each month (July-Dec)
Avg. lunchtime usage/ Avg. daily usage	Mean usage - each half-hour
Mean night usage - weekends	Mean daytime usage - weekends
Variance of night usage - weekends	Var. daytime usage - weekends

The only difference in estimation of the two trees is the seed for random number generation.[16] The diagrams contain 90% confidence intervals.

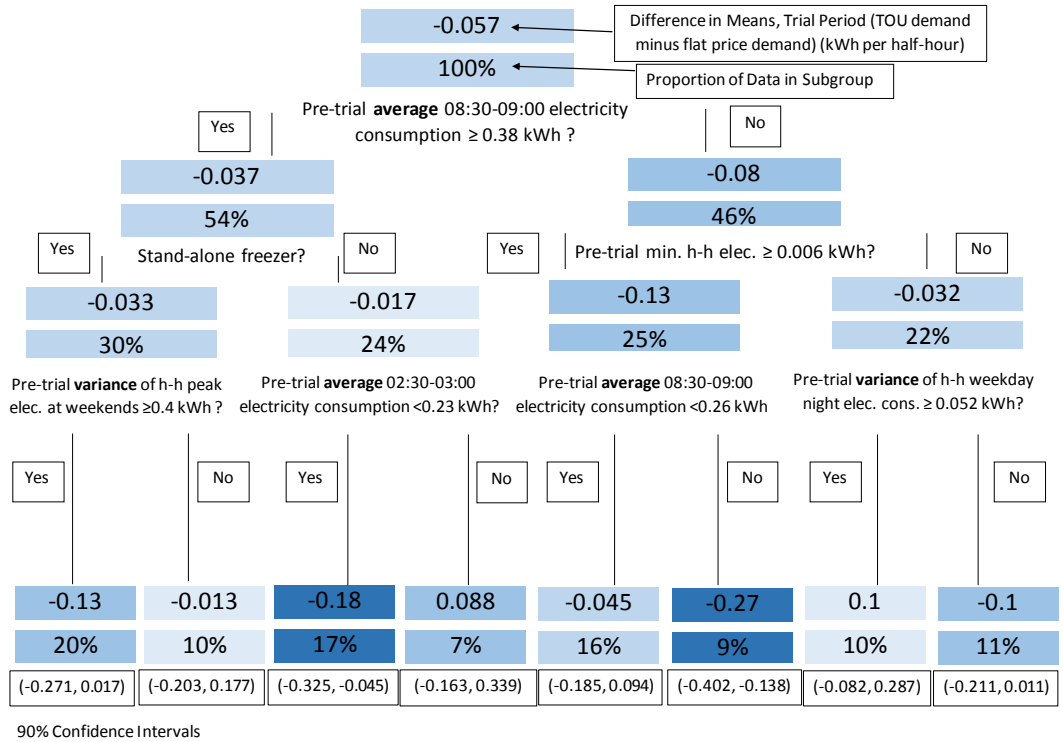
[14] The trees were obtained using the **R** package **causalTree**.

[15] The trees are "honest", i.e. separate data is used for obtaining splitting points and for obtaining terminal node estimates. Half of the data is used for creating the splits in the tree, and half is used for honest estimation. The minimum number of treatment and control observations required for a leaf split is set to ten.

[16] The seed determines the subsampling of the data into splitting and estimation data, and subsampling for cross-validation.



(a) Single Tree Example 1



(b) Single Tree Example 2 - Different seed

It can be immediately observed from these trees that the partition of the data generated by the causal tree algorithm is sensitive to the input data. This can be viewed as partly a sample size issue. Sample size, in combination with sample splitting for honest estimation, also has implications for statistical significance. There were 500 observations used for splitting, and 501 observations for estimation of treatment effects. The causal tree output contains few subgroups with significantly non-zero treatment effects at the 5% level. In contrast, CATE estimates obtained from a low-variance method, such as a linear model interacting treatment with different levels of education and including control variables, can result in multiple groups with significant effects.

The above instability can be addressed by the use of a causal forest. The instability of the output (i.e. sensitivity to the random separation of the data into splitting and estimation subsamples) is less of a problem when aggregation of predictions occurs over a large number of honest causal trees.

Causal Forest

We fitted a causal forest to the dataset containing a set of control households and households allocated to tariff C and the IHD stimulus (1001 households).[17] Each individual honest tree is fitted using a bootstrap sample consisting of half of the data, with half of this sample used for splitting and half used for estimation.[18] The number of individual trees fitted is 15000.[19] For each tree in the forest, a random subsample of one third of the set of covariates are used as potential splitting variables. [20] The minimum number of leaf observations is set to the default of five.

First, we applied the BLP test of Chernozhukov, Demirer, Duflo & Fernandez-Val (2017) to test for heterogeneity, as outlined in the methods section. The results are presented in Table 1.3. The test strongly rejects the null hypothesis that $\beta_2 = 0$, suggesting that there is heterogeneity in demand response and the causal forest ITE estimate is a relevant predictor, i.e. the ITE estimates have a non-zero coefficient when interacted with the treatment indicator variable. A comparison with the results obtained from other machine learning methods in Table 1.3 suggests that the casual forest ITE estimates are more relevant linear predictors of the true ITEs than the estimates produced by the other methods.

Table 1.4 provides a further description of demand response heterogeneity. The average peak demand reduction per half-hour is -0.150kWh for the 25% of households that reduce demand the most, while average demand reduction is not significantly nonzero for the 25% least responsive households.

Tables 1.5 and 1.6 suggest that our estimates provide a reasonable characterisation of heterogeneity.[21] In Table 1.5 we test for differences in averages of pre-trial electricity consumption variables between the 25% of households with the highest and lowest demand response. Unsurprisingly, the most responsive households consume significantly more, and have significantly more variable consumption than the least responsive households. Table 1.5 contains tests for differences in averages of binary survey variables. For example, we can observe that for the first quartile of treatment effects, i.e. the quartile of most responsive households, 40.5% of households have a respondent with third level education.

[17] The causal forest algorithm was implemented using the **R** package **grf**.

[18] Bertrand et al. (2017) also use these sizes of bootstrap samples and training and estimation subsamples. Wager & Athey (2018) divide bootstrap samples in half for honest estimation.

[19] This is somewhat arbitrary, and between the values of 10000 and 25000 used by Bertrand et al. (2017) and Davis & Heller (2017*b*).

[20] The choice of one third of the total number of covariates is commonly used for random forests.

[21] For some variables of interest, particularly binary variables with few non-zero values, confidence interval could not be obtained because for some sample splits there was not sufficient variation within quartiles for a confidence interval to be calculated.

Table 1.3: Best Linear Predictor of Average Half-hourly Peak Demand (kWh)

Causal Forest			
ATE(β_1)	HET(β_2)		
-0.095	1.620		
(-0.128, -0.062)	(0.636, 2.600)		
Elastic Net		*Boosted Tree*	
ATE(β_1)	HET(β_2)	ATE(β_1)	HET(β_2)
-0.010	0.486	-0.098	0.189
(-0.135, -0.067)	(0.214, 0.762)	(-0.131, -0.064)	(-0.081, 0.463)
Neural Network		*Random Forest*	
ATE(β_1)	HET(β_2)	ATE(β_1)	HET(β_2)
-0.093	0.035	-0.097	0.364
(-0.131, -0.056)	(-0.124, 0.195)	(-0.129, -0.065)	(0.026, 0.707)

Medians over 1000 splits. 90% confidence interval in parenthesis

The ML methods were implemented in **R** using the package **caret** and method names **glmnet**, **gbm**, **pcaNNet**, and **rf**.

HET(β_2) = The heterogeneity predictor loading parameter. This is the coefficient of the interaction of the demeaned ITE estimates and the treatment indicator in equation 1.4. $\beta_2 \neq 0$ indicates heterogeneity of treatment effects.

Table 1.4: Group Average Treatment Effects (GATEs) for most and least peak demand responsive households

Variable	25% most responsive	25% least responsive	Difference
Half-hourly peak consumption (kWh)	-0.150	-0.046	0.105
	(-0.202, -0.098)	(-0.097, 0.006)	(0.028, 0.181)

Medians over 1000 splits. 90% confidence interval in parenthesis

For the vast majority of covariates, we observe associations between the covariates and quantiles of individual effects that we would expect a priori. The most responsive households (i.e. Quartile 1) generally use more electricity, are more educated, younger, higher social class, and have more appliances. This particular result is in agreement with the observation made by Di Cosmo et al. (2014), using the same data, that more educated households are generally more responsive.[22]

Tables A.1, A.2, and A.3 in Appendix A.2 demonstrate that demographic groups that are more likely to contain vulnerable customers (CSE 2012), namely lower class and retired households, together with households for which the respondent was over 65 years old, contain a greater proportion of less responsive households. While this may be largely due to the fact that these groups have less reducible peak usage, this difference in demand response for vulnerable and non-vulnerable groups could be relevant to regulation of potential consumer targeting. The patterns of heterogeneity observed in both Tables 1.5 and 1.6 are mostly maintained when the forest is fitted using only electricity consumption data.[23]

Table 1.5: Classification Analysis (CLAN): Pre-trial electricity consumption variable averages for most and least peak demand responsive households

Variable	25% most responsive	25% least responsive	Difference
Avg. pre-trial half-hourly usage (kWh)	0.804 (0.771, 0.835)	0.229 (0.216, 0.242)	0.574 (0.540, 0.609)
Avg. pre-trial peak half-hourly usage (kWh)	1.412 (1.358, 1.467)	0.344 (0.321, 0.367)	1.068 (1.009, 1.127)
Var. of pre-trial half-hourly usage (kWh)	0.779 (0.722, 0.833)	0.109 (0.097, 0.121)	0.669 (0.613, 0.725)
Var. pre-trial peak half-hourly usage (kWh)	1.307 (1.215, 1.402)	0.168 (0.146, 0.190)	1.139 (1.042, 1.236)
Max half-hour elec. con. (kWh)	7.688 (7.414, 7.960)	3.828 (3.601, 4.055)	3.862 (3.508, 4.217)
Min half-hour elec. cons. (kWh)	0.037 (0.028, 0.047)	0.013 (0.010, 0.016)	0.025 (0.014, 0.035)
Mean daily max (kWh)	3.607 (3.489, 3.723)	1.295 (1.212, 1.377)	2.309 (2.168, 2.451)
Mean daily min (kWh)	0.149 (0.133, 0.165)	0.042 (0.037, 0.048)	0.107 (0.090, 0.123)

[22]Our focus on peak demand response is also justified by the observation by Di Cosmo & O'Hora (2017) that households "reduced consumption rather than shifting consumption from peak".

[23]The results for causal forests fitted using only survey variables or only usage variables are not included in this article, but are available from the authors on request.

Table 1.6: Classification Analysis (CLAN): Survey variable averages for most and least peak demand responsive households

Variable	25% most responsive	25% least responsive	Difference
Male	0.516 (0.427, 0.604)	0.488 (0.399, 0.577)	0.020 (-0.144, 0.105)
Internet access	0.873 (0.814, 0.932)	0.424 (0.336, 0.512)	0.457 (0.359, 0.559)
Electric central heating	0.032	0.048 (0.010, 0.086)	-0.016 (-0.033, 0.064)
Water immersion	0.635 (0.550, 0.720)	0.424 (0.336, 0.512)	0.211 (0.089, 0.333)
Water centrally heated	0.159 (0.094, 0.223)	0.112 (0.056, 0.168)	0.047 (-0.041, 0.128)
Went without heat from lack of money	0.048 (0.010, 0.085)	0.040 (0.005, 0.075)	0.000 (-0.053, 0.048)
Lagging jacket on hot water	0.857 (0.795, 0.919)	0.776 (0.702, 0.850)	0.081 (-0.017, 0.177)
Third level education	0.405 (0.318, 0.492)	0.288 (0.208, 0.368)	0.125 (0.005, 0.241)
Employee	0.563 (0.476, 0.651)	0.328 (0.245, 0.411)	0.235 (0.114, 0.355)
Apartment	0	0.048 (0.010, 0.086)	-0.048 (-0.086, -0.01)
Instantaneous water heater	0.008	0.024	-0.016
Plug-in electric heater	0.032	0.024	0.008 (-0.047, 0.033)

Note: For some binary variables with few non-zero values, confidence interval could not be obtained because for some sample splits there was not sufficient variation within quartiles for a confidence interval to be calculated.

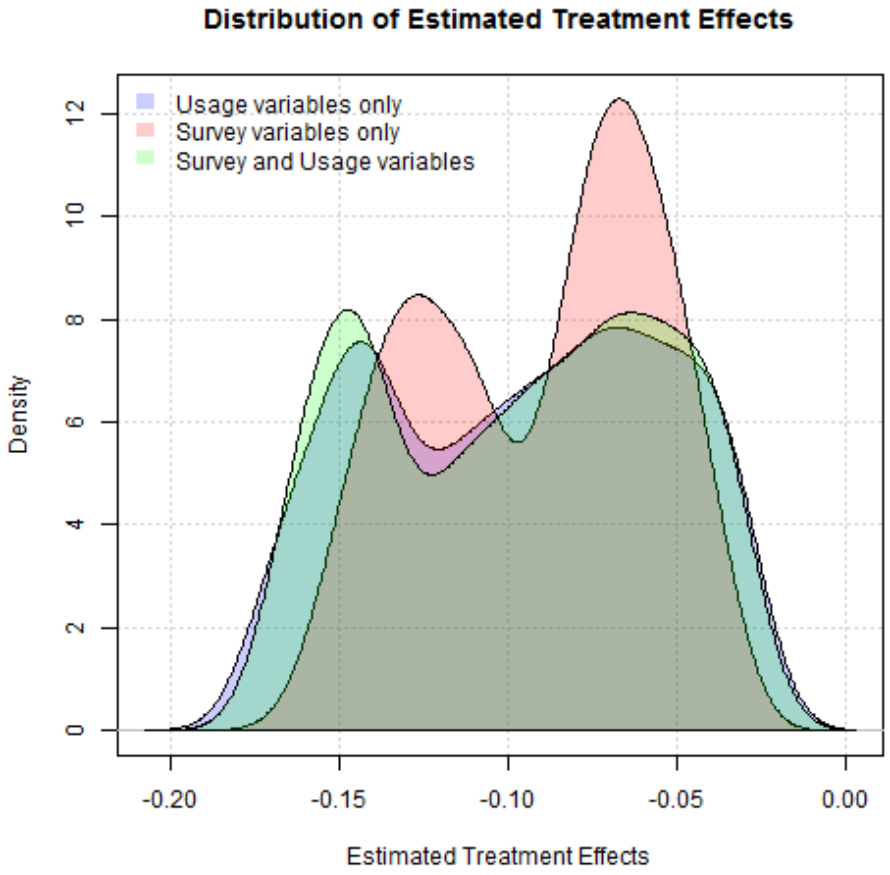


Figure 1.3: Density plots of causal forest household estimates fitted using different sets of variables

To demonstrate this, Figure 1.3 presents a density plot comparing the distributions of the ITE estimates obtained by fitting causal forests with different sets of potential conditioning variables. One forest was fitted using both survey and usage variables, one forest was fitted using only usage variables, and one forest was fitted using only survey variables. This suggests that electricity consumption data contains information related to survey data information that can characterise heterogeneous groups of demand response. This issue may be relevant to firms or policymakers who wish to understand which information to collect in order to predict demand response.

The results suggest that the usage variables exert a greater influence on the causal forest estimates. Furthermore, the density plot suggests potential bimodality in the distribution of individual effects. However, although it is most plausible that past usage variables are more informative than survey variables, we also note the possibility that these results are driven by the bias of variable selection towards continuous variables, which have more potential splitting points. This issue can be addressed by discretizing each continuous usage variable, for example, into indicator variables defined by quantiles.

Figure 1.4 shows ITEs with confidence intervals ordered by size of estimated effect.[24] None of the individual estimates are significantly positive. This accords with economic intuition.

[24]Confidence intervals are produced by the **causal_forest** command of the R package **grf** (Wager & Athey 2018). Each level of a categorical survey variable is represented by a separate binary potential splitting variable because the package currently does not support finding optimal splits of multiple categories.

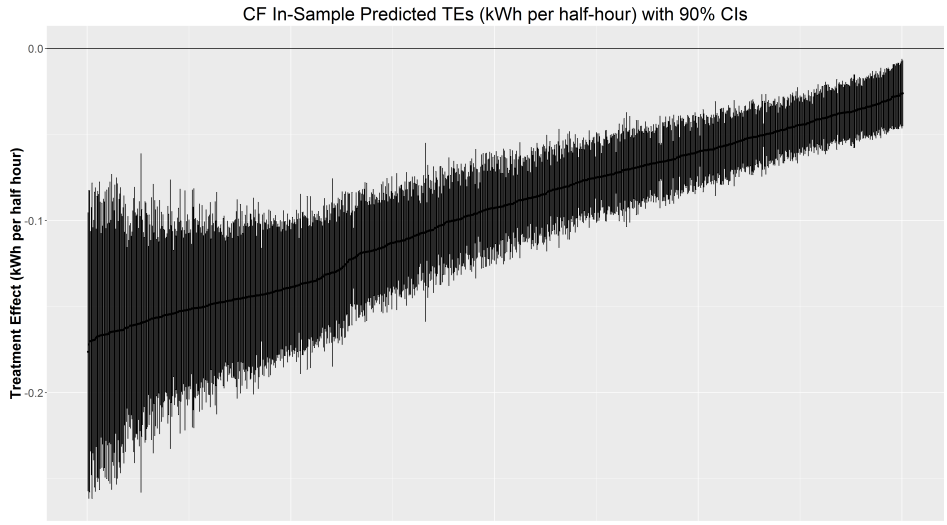


Figure 1.4: 90% Confidence Intervals for ITEs ordered by size of ITE estimate

Variable Importance

In this section we present the results for variable importance utilising the methods outlined in Section 1.2. The variable importance measure is a depth-weighted average of the number of splits on the variable of interest.[25] For the second method we also carry out a permutation-based test, as outlined in Section 1.2.

Columns (1) and (2) of Table 1.7 give the variable names and values for the variable importance measure. The variables are ordered by importance, with larger values indicating greater importance. The importances are scaled such that the most important variable has variable importance equal to 100.

The results indicate that the trees most often split on electricity usage, and specifically variables that indicate the level and variance of weekday electricity consumption. The most important survey variables are number of laptop PCs, number of freezers, and employment status. These variables are likely to be correlated with income and level of electricity usage.

As noted in Section 1.2, given the bias of variable importance measures in favour of variables with more splitting points (Strobl 2008), we implement an alternative *permutation*-based measure of variable importance which is able to address this issue (Altmann et al. 2010). Column (3) shows the "p-value" permutation-based measures for the **grf** variable importances. These measures are referred to as "p-values" because the permutation-based approach is influenced by existing approaches to conditional independence testing. However, these measures do not have the properties of valid p-values of a test of conditional independence between the treatment effect and covariate of interest. See Nembrini (2019) for further discussion of this issue. The variables with lower "p-values" are interpreted as more important in the sense that they are selected more often by the causal forest algorithm when the outcome is the true outcome and not noise (or a permuted outcome). Therefore this measure should be less biased towards values with more potential splitting points, which can be spuriously selected more often by the algorithm, even when none of the covariates can predict the outcome.

The "p-values" confirm the pattern of results observed in column (2) in so far as the past consumption information variables that obtain the highest variable importances also obtain the lowest "p-values", indicating that the high proportion of causal forest splits based on these variables were not only the result

[25] This is the default measure in the **R** package **grf**.

of spurious selection of these variables resulting from the greater number of potential splitting points for continuous variables than for categorical variables. On the other hand, a few of the most important survey variables were also confirmed to be important according to the "'p-value" measure. For example *number of laptop PCs* and *number of freezers* have "p-values" of 0.1 and 0.06 respectively. Furthermore, some of the past electricity consumption information variables were relatively unimportant according to the "p-value" measure, such as *mean daily min. usage* and *mean lunchtime / mean day usage*, with values of 0.72 and 0.92 respectively. Nonetheless, the overall pattern is that the most important variables across both measures are past electricity consumption information variables while most of the least important variables across both measures are categorical survey variables.

Variable name	Variable importance	"p-value"	Variable name	Variable importance	"p-value"
electric plugin heating	0	0.09	mean 14:00-14:30 usage	12.23	0.79
water instantly heated	0	0	mean usage - weekdays	12.9	0.1
unheated, lack of money	0	0.38	var. night usage - weekends	12.95	0.99
number of washing machines	0.02	0.47	mean daytime usage - weekends	13	0.14
electric central heating	0.22	0.9	mean night / mean day usage	13.07	1
prop. double glazed windows	0.25	1	mean 21:30-22:00 usage	13.07	0.73
number of electric cookers	0.46	1	mean 22:30-23:00 usage	13.28	0.9
number of immersion heaters	0.54	1	var. night usage - weekdays	13.35	1
number of dishwashers	0.58	1	mean 13:00-13:30 usage	13.72	0.84
type of cooker	0.84	1	mean 06:30-07:00 usage	14.14	0.98
number of tumble dryers	0.85	1	mean daytime usage - weekdays	14.87	0.16
water centrally heated	1.06	1	mean 14:30-15:00 usage	15.23	0.71
regular internet user	1.07	1	mean usage - weekends	15.3	0.07
sex of respondent	1.23	1	var. daytime usage - weekdays	15.36	0.4
own or rent home	1.24	1	mean 19:00-19:30 usage	15.36	0.61
no. of elec. convector heaters	1.33	1	mean 21:00-21:30 usage	15.49	0.49
water pumped from elec. well	1.58	0.99	mean 07:30-08:00 usage	15.96	1
attic insulated	1.79	1	mean h-h coef. of variation	16.6	1
number of instant elec. showers	2.03	1	var. nonpeak usage - weekdays	16.64	0.26
external walls insulated	2.06	1	mean 23:00-23:30 usage	16.76	0.85
other internet users	2.07	0.49	*number of freezers*	17	0.06
water immersion	2.22	0.98	variance nonpeak usage	17.12	0.12
number of small TVs	2.35	1	mean 00:00-00:30 usage	17.39	0.84
number of hot tank elec. showers	2.36	0.97	*number of laptop PCs*	17.48	0.1
age of home	2.92	1	variance daytime usage	18.06	0.16
number of games consoles	2.94	0.89	mean 10:00-10:30 usage	19.76	0.51
education	3.44	1	mean 20:00-20:30 usage	19.93	0.12
lagging jacket	3.45	0.44	variance of usage	20.1	0.05
has an energy rating	3.46	0.44	mean daily min. usage	20.72	0.72
age of respondent	3.91	1	mean 23:30-00:00 usage	21.12	0.68
prop. elec. saving lightbulbs	4.74	1	mean 18:00-18:30 usage	21.29	0.25
number of bedrooms	4.87	0.98	var. usage - weekends	21.41	0.13
lives alone	5.17	0.69	mean 18:30-19:00 usage	21.82	0.19
mean 06:00-06:30 usage	5.61	1	mean 19:30-20:00 usage	22.01	0.17
mean 02:30-03:00 usage	5.66	1	var. usage - weekdays	22.11	0.05
type of home	5.78	0.97	var. daytime usage - weekends	22.85	0.1
mean 04:00-04:30 usage	6.03	1	mean 16:00-16:30 usage	22.86	0.46
mean 12:00-12:30 usage	6.1	1	mean 09:00-09:30 usage	23.05	0.58
internet access	6.35	0.12	mean November peak usage	24.88	0.15
mean 03:00-03:30 usage	6.44	1	min. half-hourly usage	25.8	0.72
mean 03:30-04:00 usage	6.68	1	mean 16:30-17:00 usage	26.02	0.51
mean 05:00-05:30 usage	6.73	1	var. November peak usage	26.19	0.39
mean night usage	6.76	0.99	max. half-hourly usage	26.96	0.67
mean 04:30-05:00 usage	7.12	1	mean lunchtime / mean day usage	27.47	0.92
number of big TVs	7.42	0.68	mean 15:30-16:00 usage	29.14	0.18
mean 11:00-11:30 usage	7.5	1	mean daily max. usage	29.39	0.06
mean night usage - weekends	7.53	0.99	mean 15:00-15:30 usage	30.3	0.13
mean 05:30-06:00 usage	7.87	1	mean 08:00-08:30 usage	30.85	0.48
mean 11:30-12:00 usage	8.22	1	mean 20:30-21:00 usage	31.45	0.04
mean 00:30-01:00 usage	8.48	1	var. December peak usage	38.92	0.2
social class	8.95	0.64	mean 09:30-10:00 usage	39.05	0.13
mean 01:30-02:00 usage	9.02	0.99	mean 08:30-09:00 usage	39.44	0.21
number of desktop PCs	9.11	0.12	mean peak usage - weekdays	42.61	0
mean 12:30-13:00 usage	9.36	1	mean peak usage	44.2	0
mean night usage - weekdays	9.38	0.96	variance peak usage	44.65	0.01
mean 13:30-14:00 usage	9.83	0.99	var. peak usage - weekdays	47.83	0.03
mean 01:00-01:30 usage	10.31	0.98	mean July peak usage	49.55	0.08
mean nonpeak usage - weekdays	10.31	0.3	mean September peak usage	49.96	0.03
variance night usage	10.47	1	mean 17:00-17:30 usage	57.62	0.02
mean 02:00-02:30 usage	10.78	0.94	mean December peak usage	62.11	0
mean 10:30-11:00 usage	10.86	0.99	var. September peak usage	66.87	0.03
employment	11.22	0.4	var. July peak usage	68.34	0.1
mean nonpeak usage	11.66	0.17	mean August peak usage	68.73	0
mean of usage	11.86	0.1	mean 17:30-18:00 usage	69.41	0
mean 07:00-07:30 usage	11.9	1	var. August peak usage	74.89	0.01
mean 22:00-22:30 usage	11.96	0.85	mean October peak usage	76.14	0
mean daytime usage	12.2	0.19	var. October peak usage	100	0

Survey variables are in italics. The "p-values" are permutation based importance measures (motivated by permutation-based testing methods) and are not valid p-values for hypothesis tests.

Table 1.7: Variable importance results

1.5 Conclusion

In this article we have examined heterogeneity of demand response following the introduction of time-of-use electricity pricing. Variable importance measures, adjusted for differences in information content across past usage and demographic variables, suggest that the causal forest algorithm favours the use of certain functions of past electricity consumption rather than survey information to describe heterogeneity. Tables 1.6 to A.3 reveal notable patterns of heterogeneity across *unimportant* survey variables. For example, the causal forest results suggest that younger, more educated households that consume more electricity exhibit greater demand response to new pricing schemes. In this respect, although survey variables can be less informative than detailed electricity consumption information in terms of selection in the causal forest algorithm, they can also be correlated with *important* past consumption information.

Chapter 2

State-of-the-BART: Simple Bayesian Tree Algorithms for Prediction and Causal Inference

Abstract

Bayesian Additive Regression Trees (BART) (Chipman et al. 2010) and Bayesian Causal Forests (BCF) (Hahn et al. 2020) are state-of-the-art machine learning algorithms for prediction and treatment effect estimation. These methods involve averaging predictions from sum-of-tree models, typically drawn using Markov Chain Monte Carlo (MCMC) methods.

This paper introduces conceptually and computationally simple alternatives to MCMC implementations of BART. A new importance sampling based implementation of BART (BART-IS) builds on the ideas of Hernández et al. (2018) and Quadrianto & Ghahramani (2014). BART-IS samples models from a data independent model prior. This paper also contains an extension to average and individual treatment effect estimation, BCF-IS.

In addition, this paper describes Bayesian Causal Forests using Bayesian Model Averaging (BCF-BMA), an implementation of BCF (Hahn et al. 2020) that extends an improved implementation of BART-BMA (Hernández et al. 2018) to treatment effect estimation. [1]

Three applications are included in this paper: 1. The treatment effect estimation methods introduced in this chapter and existing methods are compared using a Time-of-Use electricity pricing trial dataset. 2. BART-BMA and BART-IS are applied to inflation forecasting. 3. BART-BMA and BART-IS are used to identify determinants of economic growth.

2.1 Introduction

Prediction and treatment effect estimation are key tasks for policy makers (Kleinberg et al. 2015). Economists are increasingly applying machine learning methods for treatment effect estimation (Wager & Athey 2018, Athey & Imbens 2015, Athey 2018).

BART and BCF are Bayesian machine learning methods for prediction and treatment effect estimation (Chipman et al. 2010, Hahn et al. 2020). In this paper, a set of new implementation algorithms are introduced for these methods. BART and BCF can be interpreted as model averages of Bayesian linear regressions with the sets of covariates equal to binary variables indicating if observations fall in terminal nodes of decision trees. The covariates are defined by decision tree structures, and a prior on the tree structures defines a prior

[1] **R** packages implemented in C++ for the methods described in this paper are available at https://github.com/EoghanONeill . Many thanks are due to Belinda Hernandez and Andrew Parnell for providing the original BART-BMA code and providing useful feedback on improvements to the algorithm.

on the space of models. This interpretation of BART provides a link to the existing econometric literature on Bayesian Model Averaging of linear models (Steel 2017, Fernandez et al. 2001*a*,*b*, Brock & Durlauf 2001).

The key contributions of this paper are:

1. An new implementation of BART-BMA (Hernández et al. 2018) with improvements to the model search algorithm, and calculations of model probabilities and prediction intervals. [2]

2. Bayesian Causal Forests using Bayesian Model Averaging (BCF-BMA). This method accounts for confounding on observables using the BCF parameterization of BART (Hahn et al. 2020), while retaining the parsimonious model selection approach of BART-BMA.

3. Simple importance sampling based implementations of BART and BCF (referred to in this paper as BART-IS and BCF-IS), following the approach for single classification trees described by Quadrianto & Ghahramani (2014). This approach provides a link between BART and the implementation of BMA of linear models used by Sala-i Martin et al. (2004), and also shares some similarities with extremely randomized trees (Geurts et al. 2006).

The algorithms presented in this paper provide two contrasting approaches to the implementation of BART and BCF. The BMA implementations involve a deterministic greedy model search that finds suitable splitting points and grows trees to add to sum-of-tree models. The IS implementations involve data independent random draws of models. In some applications, both approaches yield similar results to existing MCMC based implementations. MCMC-based methods can be limited by factors such as poor mixing of chains and lack of parallelizability. [3] Therefore this paper explores the viability of alternatives to MCMC implementations of BART. The algorithms introduced in this paper provide alternatives options to the implementation of BART that may be suited to particular datasets or computational constraints. A further motivation for this paper is the generalizability of the new implementations beyond models for continuous outcomes (see third chapter).

These methods are conceptually simple, in that conjugate priors give a tractable closed form for the predictive distribution (e.g. of the Average Treatment Effect). The appeal of BART-IS and BCF-IS is that they are straightforward to implement and very parallelizable. The output of BCF-BMA contains relatively few sum-of-trees models. Under the default settings, each model include five trees describing the treatment effect function and each tree contains at most five splits. Therefore the output is more interpretable than that of standard MCMC implementations, which usually draw thousands of models, each of which contains a sum of a hundred or more trees.

The range of different implementation methods for BART is analogous to the range of possible implementations for BMA of linear models (Hoeting et al. 1999). BART-BMA and BCF-BMA follow the Occam's

[2] BART-BMA applies a greedy model search algorithm to find trees to append to sum-of-tree models, and keeps a small number of models with highest posterior probability. The search for trees is based on residuals calculated from models found in previous rounds of the model search algorithm. The new version of the algorithm calculates residuals after re-estimating the whole sum-of-tree model in each round, whereas the old version fits a single tree to the residuals and adds this to a previously estimated model in a manner similar to boosting. Other improvements include bug fixes and a different method for calculation of credible intervals. Furthermore the new implementation is entirely deterministic. The original implementation by Hernández et al. (2018) and standard MCMC implementations of BART rely on random sampling.

[3] Hill et al. (2020) note that "Posterior computation has improved since the initial implementation of BART, but room for further improvement remains. Most BART implementations can handle hundreds of covariates and tens of thousands of observations, although mixing of the MCMC algorithm tends to degrade as either the sample size or dimension gets larger. Scaling to larger data sets (both in terms of the number of observations and the number of predictors) would naturally be quite useful. In all likelihood this will be more than an engineering exercise, and more efficient algorithms for posterior inference will be necessary."

window approach (Madigan & Raftery 1994, Volinsky et al. 1997). Standard BART-MCMC is similar to Stochastic Search Variable Selection (George & McCulloch 1995) and Markov Chain Monte Carlo Model Composition.[4] BART-IS and BCF-IS are similar to importance sampling of linear models (Clyde et al. 1996, Stewart 1987, Sala-i Martin et al. 2004).[5] [6]

Early examples of Bayesian Model Averaging of tree-based models include examples of single tree models for classification (Buntine 1992, Kwok & Carter 1990). An importance sampling based approach for single classification trees is described by Quadrianto & Ghahramani (2014). Table 2.1 places the methods introduced in this paper in the existing Bayesian Tree literature.

	Single Tree Regression/Classification	Sum-of-Trees Regression	Sum-of-Trees Treatment Effects
MCMC	Chipman et al. (1998) Bayesian regression tree	Chipman et al. (2010) BART	Hahn et al. (2017) BCF
BMA	Buntine (1991)	Hernandez et al. (2018) BART-BMA	**BCF-BMA**
Importance sampler	Quadrianto and Ghahramani (2015) safe-Bayesian-RF	**BART-IS**	**BCF-IS**

The methods introduced in this paper are in **blue** text.

Table 2.1: Summary of Bayesian tree algorithms.

This paper includes a comparison of methods across simulated datasets. The BMA and IS implementations give comparable results to MCMC-based implementations of BART and BCF. We also illustrate the applicability of the algorithms to:

1. Treatment effect estimation, using a Time-of-Use electricity pricing trial dataset (CER 2011)
2. Forecasting, using an inflation time series dataset (Garcia et al. 2017)
3. Variable selection, using a growth determinant dataset (Sala-i Martin et al. 2004).

The remainder of this paper is structured as follows: section 2.2 provides a review of BART and BART-BMA, section 2.3 describes some improvements to BART-BMA and outlines how BART-BMA is applicable to treatment effect estimation, section 2.4 introduces BART-IS, section 2.5 provides a comparison of BART-MCMC, BART-BMA, and BART-IS using simulated data, 2.6 introduces BCF-BMA and BCF-IS and com-

[4] An implementation of BART fully analogous to Markov Chain Monte Carlo Model Composition (Madigan et al. 1995, Raftery et al. 1997) is possible, and has been implemented for single tree models (Chipman et al. 1998). This approach would differ from standard BART-MCMC in that it involves marginalization of the variance of the error term. This approach is not implemented in this paper, although future work may compare the performance of this approach to the methods introduced in this paper. Boatman et al. (2020) apply this approach in the context of combining primary source and supplementary source data for causal effect estimation.

[5] An interesting topic for future research is Bayesian Adaptive Sampling (BAS) of BART Models (Clyde et al. 2011). BAS involves sampling without replacement and possibly adjusting sampling probabilities by predicting the marginal likelihood of unsampled models. While BAS has been applied to sampling of linear models, further research is required for application of this approach to tree-based models. This hypothetical alternative approach to BART (BART-BAS) is not to be confused with standard BAS which uses a binary tree structure to represent the model space of standard *linear* regression models. It is also distinct from the existing literature that applies BART-MCMC to guide adaptive sampling of linear models (Yu et al. 2010, 2012, Yu & Li 2020).

[6] In the context of sampling/estimation of parameters in a single model, it has been observed (e.g. Chopin et al. (2017)) that sophisticated methods such as MCMC do not notably outperform importance sampling on some datasets. However, the viability of simple importance sampling of models in the context of BMA has not been thoroughly studied beyond the work of Clyde et al. (1996), Stewart (1987), Sala-i Martin et al. (2004), Quadrianto & Ghahramani (2014) and Clyde et al. (2011).

pares the performance of these methods and existing methods on a range of simulated datasets, section 2.7 includes three example applications of the methods introduced in this paper, and 2.8 concludes the paper.

2.2 Review of BART and BART-BMA

This section describes the BART model (Chipman et al. 2010), reviews BART implementations, summarizes applications of BART, and describes BART-BMA (Hernández et al. 2018).

2.2.1 Overview of BART

Description of Model and Priors

Suppose there are n observations, and the $n \times p$ matrix of explanatory variables, X, has i^{th} row $x_i = [x_{i1}, ..., x_{ip}]$. Following the notation of Chipman et al. (2010), let T binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and let $M = \{\mu_1, ..., \mu_b\}$ denote a set of parameter values associated with each of the b terminal nodes of T. The decision rules are binary splits of the predictor space of the form $\{x \in A\}$ vs $\{x \notin A\}$ where A is a subset of the range of x. These are typically of the form $\{x_{is} \leq c\}$ vs $\{x_{is} > c\}$ for continuous x_s ($s \in \{1, ..., p\}$). Each observation's x_i value is associated with a single terminal node of T by the sequence of decision rules from top to bottom, and is then assigned the μ value associated with this terminal node. For a given T and M, we use $g(x_i; T, M)$ to denote the function which assigns a $\mu \in M$ to x_i. This gives the single tree model $Y \sim g(x_i; T, M) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ (Chipman et al. 1998).

For the standard BART model, the outcome is determined by a sum of trees,

$$Y_i = \sum_{j=1}^{m} g(x_i; T_j, M_j) + \varepsilon_i$$

where $g(x_i; T_j, M_j)$ is the output of a decision tree. T_j refers to decision tree $j = 1, ..., m$, where m is the total number of trees in the model. M_j are the terminal node parameters of T_j, and $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

Prior independence is assumed across trees T_j and across terminal node means $M_j = (\mu_{1j} ... \mu_{b_j j})$ (where $1, ..., b_j$ indexes the terminal nodes of tree j). The form of the prior used by Chipman et al. (2010) is:

$$p(M_1, ..., M_m, T_1, ..., T_m, \sigma) \propto \left[\prod_j \left[\prod_k p(\mu_{kj} | T_j) \right] p(T_j) \right] p(\sigma)$$

In standard BART, $\mu_{kj} | T_j \overset{i.i.d}{\sim} N(0, \sigma_0^2)$ where $\sigma_0 = \frac{0.5}{e\sqrt{m}}$ and e is a user-specified hyper-parameter.

Chipman et al. (2010) set a regularization prior on the tree size and shape $p(T_j)$ to discourage any one tree from having undue influence over the sum of trees. The probability that a given node within a tree T_j is split into two child nodes is $\alpha(1 + d_h)^{-\beta}$, where d_h is the depth of (internal) node h and α and β are parameters which determine the size and shape of T_j respectively. There are also priors on the splitting variables and splitting points in each tree. Chipman et al. (2010) use a uniform prior on available splitting variables, and a uniform prior on the discrete set of available splitting variables. Chipman et al. (2010) assume that the model precision σ^{-2} has a conjugate prior distribution $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$ with degrees of freedom v and scale λ.

BART predictions are averages from sum-of-tree models of the form described above. Therefore model uncertainty is taken into account and there are two levels of regularization. Firstly, greater prior probability is placed on models with shallower trees with fewer splitting points. Secondly, over-fitting is further avoided through the prior on the terminal node parameters μ_{kj}, as in standard Bayesian linear regression.[7]

Existing BART Implementations

Samples can be taken from the posterior distribution $p((T_1, M_1), ..., (T_m, M_m), \sigma|y)$ by a Bayesian backfitting MCMC algorithm. This algorithm is a Gibbs or Metropolis Hastings sampler, involving m successive draws from $(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y$ for $j = 1, ..., m$ [where $T_{(j)}, M_{(j)}$ are the trees and parameters for all trees except the j^{th} tree] followed by a draw of σ from the full conditional $\sigma|T_1, ..., T_m, M_1, ..., M_m, y$.

A set of draws induces the sum of trees function $f^*(.) = \sum_{j=1}^{m} g(. \, ; T_j^*, M_j^*)$. After burn-in, the sequence of f^* draws, $f_1^*, ..., f_Q^*$ may be regarded as an approximate, dependent sample of size Q from $p(f|y)$. To estimate the unknown function $f(x)$,[8] a natural choice is $\frac{1}{Q}\sum_{q=1}^{Q} f_q^*(x)$, which approximates $E(f(x)|y)$. Prediction intervals can be obtained from quantiles of the draws $f_q^*(x)$.

A number of papers describe faster BART implementation algorithms and improved sampling methods, including parallelized BART (Pratola et al. 2014), particle Gibbs algorithms (Lakshminarayanan et al. 2015), more efficient Metropolis-Hastings proposals (Pratola et al. 2016), Consensus Monte Carlo (Scott et al. 2016), a likelihood-inflated sampling algorithm (Entezari et al. 2018), and Accelerated BART (X-BART, which uses a stochastic hill climbing algorithm as a greedy stochastic approximation to MCMC) (He et al. 2018). An alternative to the MCMC BART implementation is Approximate Bayesian Computation Bayesian Forests (Liu et al. 2018), which has been shown to be consistent for variable selection under certain conditions.

BART-BMA (Hernández et al. 2018), in contrast to other BART implementations, does not involve MCMC methods. A greedy model search algorithm adds trees to sum-of-tree models by first restricting the set of potential splitting points using a changepoint detection algorithm, and only keeping sum-of-tree models with posterior model probabilities within a distance, known as Occam'a window (Madigan & Raftery 1994), of the highest probability model currently in the set of selected models. [9] See sections 2.2.2 and 2.3 for more details.

BART Theory

Recent papers have discussed the asymptotic properties of BART. Posterior concentration rates are derived by Rockova & van der Pas (2017), Linero & Yang (2017) and Rocková & Saha (2018). Castillo & Rockova (2019) obtain uncertainty quantification results. Asymptotic properties of variable selection are derived by Liu et al. (2018). Asymptotic results for estimating ITEs using Bayesian methods more generally are described by Alaa & van der Schaar (2018).

Review of Extensions and Applications of BART

BART has been extended to a wide range of applications (Hill et al. 2020, Yao et al. 2018). Starling et al. (2018) describe a BART method for functional data analysis that parameterizes each tree's terminal nodes

[7] Careful calibration of these priors can play an important role

[8] $Y_i = \sum_{j=1}^{m} f(x_i) + \varepsilon_i \approx \sum_{j=1}^{m} g(x_i; T_j, M_j) + \varepsilon_i$

[9] In the original implementation of BART-BMA, a Gibbs sampler was used for constructing prediction intervals. In the improved implementation, we obtain intervals from a closed form for the model averaged posterior predictive distribution. Therefore BART-BMA provides an implementation of BART that does not require any random number generation

with smooth functions of a target covariate. Another smooth variant of BART is BART with "soft" decision trees (Linero & Yang 2017).

Some variations on the BART priors have been suggested for variable selection, including a Dirichlet hyperprior on the probability that a variable is used for a split (Linero 2018) and spike and tree priors (Rockova & van der Pas 2017, Liu et al. 2018). An overlapping group Dirichlet hyperprior has been applied to splitting probabilities for a dataset in which the variables have an overlapping group structure (Du & Linero 2019). A prior for interaction detection has been proposed by Du & Linero (2018).

BART can be applied to data without i.i.d normally distributed error terms. Heteroscedastic BART models the error as a product of trees (Pratola et al. 2017), and fully nonparametric BART (George et al. 2018) models the error using a Dirichlet process mixture.

BART has been adapted for different outcome variables, including Bayesian quantile additive regression trees (Kindo, Wang, Hanson & Peña 2016), Multiclass Bayesian Additive Classification Trees (Kindo et al. 2013), BART methods for multinomial outcomes (Agarwal et al. 2014, Kindo, Wang & Peña 2016), loglinear BART (Murray 2017), random intercept BART (Tan et al. 2016), BART for survival analysis (Bonato et al. 2010, Sparapani et al. 2016), BART for competing risks models (Sparapani et al. 2019), and BART modelling of recurrent events (Sparapani et al. 2018). A general framework for extending BART to different tasks is described by Tan & Roy (2019).

BART can also be applied to data with multiple outcomes. Chakraborty (2016) applies BART to Seemingly Unrelated Regression, and Linero et al. (2019) describe shared Bayesian Forests. BART has been used for the imputation of missing data (Xu et al. 2016, Tan et al. 2018) and the modelling of missing data in longitudinal studies (Zhou et al. 2019).

BART has been applied to treatment effect estimation (Hill 2011, Green & Kern 2012, Taddy et al. 2015, Henderson et al. 2017). Data analysis competitions (Dorie et al. 2019, Hahn et al. 2019, Carvalho et al. 2019) have shown that BART is among the most accurate treatment effect estimation methods. Hahn et al. (2020) introduce Bayesian Causal Forests (BCF), a BART based method for treatment effect estimation that allows the prior regularization of the treatment effect estimate to be specified separately to the prior regularization of the rest of the model for the outcome.

Hahn et al. (2020) also note that standard BART treatment effect estimates can be improved by including the propensity score as a potential splitting variable. Santos & Lopes (2018) study the performance of this approach on sparse data using the Dirichlet hyperprior described by Linero (2018). BCF has been extended to Instrumental Variable estimation of treatment effects by Bargagli-Stoffi et al. (2019). Deshpande et al. (2020) extend BCF to a linear varying coefficient framework (VC-BART), and demonstrate theoretical near-optimality and derive posterior concentration rates in settings with independent and correlated errors.

2.2.2 Overview of BART-BMA

BART-BMA (Hernández et al. 2018) applies the same priors as standard BART (section 2.2.1), except the variance of the terminal node parameters is proportional to the variance of the error term, $\mu_{ij}|T, \sigma \sim N(0, \frac{\sigma^2}{a})$, as suggested by Chipman et al. (1998).[10] Integration of the likelihood with respect to the μ parameters and σ results in a closed form expression proportional to the marginal likelihood.

[10]Moran et al. (2018) argue against the use conjugate priors in Bayesian linear regression. However, this issue will not be discussed in further detail in this paper. Nonetheless, it is worth noting that the methods introduced in this paper can be improved further by careful calibration of the a parameter, e.g. by cross-validation.

The marginal likelihood can be derived as follows. Let $Y = (Y_1, ..., Y_n)$ be the outcome vector. For a given sum of trees model $\mathcal{T}$, the likelihood of Y is:

$$Y|\mathcal{T}, M_1, ..., M_m, \sigma^{-2} \sim N(\sum_{j=1}^{m} J_j M_j, \sigma^2 I)$$

where J_j (which depends on the original matrix of covariates X) is an $n \times b_j$ binary matrix whose (i,k) element denotes the inclusion of observation $i = 1, ..., n$ in terminal node $k = 1, ..., b_j$ of tree j.

Let $W = [J_1 ... J_m]$ be an $n \times b$ matrix , where $b = \sum_{j=1}^{m} b_j$, and let $\underline{\boldsymbol{\mu}} = (M_1^T ... M_m^T)^T$ be a vector of size b of terminal nodes assigned to trees $T_1, ..., T_m$. We can then write $W\underline{\boldsymbol{\mu}} = \sum_{j=1}^{m} J_j M_j$, [11] and therefore

$$Y|\underline{\boldsymbol{\mu}}, \sigma^{-2} \sim N(W\underline{\boldsymbol{\mu}}, \sigma^2 I)$$

which, with $\underline{\boldsymbol{\mu}} \sim N(0, \frac{\sigma^2}{a} I_b)$, where I_b is a $b \times b$ identity matrix, implies that the marginal likelihood is given by a multivariate student distribution with ν degrees of freedom[12]

$$p(Y) = \frac{\Gamma(\frac{\nu+n}{2})(\lambda v)^{\frac{\nu+n}{2}}}{\Gamma(\frac{v}{2}) v^{\frac{n}{2}} \pi^{\frac{n}{2}} \lambda^{\frac{n}{2}} (\frac{1}{a})^{\frac{b}{2}} det\,(aI_b + W^T W)^{\frac{1}{2}}} \left[\lambda v + Y^T Y - Y^T W (aI_b + W^T W)^{-1} W^T Y\right]^{-\frac{\nu+n}{2}}$$

Anything that does not depend on W or b will cancel out when calculating the model weights, therefore it is only necessary to calculate:

$$\propto \frac{1}{(\frac{1}{a})^{\frac{b}{2}} det\,(aI_b + W^T W)^{\frac{1}{2}}} \left[\lambda v + Y^T Y - Y^T W (aI_b + W^T W)^{-1} W^T Y\right]^{-\frac{\nu+n}{2}}$$

and the log marginal likelihood is proportional to $\frac{b}{2}\log(a) - \frac{1}{2}\log(det(M)) - \frac{\nu+n}{2}\log(\lambda v + Y^T Y - Y^T W M^{-1} W^T Y)$ where $M = aI_b + W^T W$.

A deterministic model search algorithm first reduces the set of potential splitting variables by a change-point detection algorithm, and then recursively adds splits to trees that are potentially to be appended to models in the set of currently selected sum of tree models. After a set of single tree models are selected, changepoints in the residuals are used as potential splitting variables for constructing the next set of trees to potentially append to the selected models.[13] Then a new set of residuals is constructed for the new set of sum-of-two-tree models, changepoints are detected, and trees are appended to create a set of sum-of-three-tree models, and so on.

The set of models to be averaged over are those with posterior probability within some distance of the highest probability model found by the model search algorithm. i.e. For all proposed models, $\mathcal{T}_\ell$, indexed by ℓ, the algorithm obtains

$$p(Y|\mathcal{T}_\ell, X)p(\mathcal{T}_\ell) \propto p(\mathcal{T}_\ell|Y, X) = \frac{p(Y|\mathcal{T}_\ell, X)p(\mathcal{T}_\ell)}{p(\mathbf{y})}$$

[11] $W\underline{\boldsymbol{\mu}} = \sum_{j=1}^{m} J_j M_j$ is analogous to $X\boldsymbol{\beta}$ in standard linear regression notation.

[12] Each sum-of-tree model is a ridge regression with each covariate being a dummy variable for a terminal node. $Y \sim MVST_v(0, \lambda(I_n + \frac{1}{a} W W^T)$

[13] In the original paper, Hernández et al. (2018) construct residuals by subtracting from the outcomes the sum of single tree model predictions (for each tree in the sum-of-tree model). In this paper we present the results of an improved algorithm that calculates the residuals by subtracting from the outcome the posterior mean of the whole sum-of-tree model. i.e. correlations across trees and the whole set of parameters for all trees influence the predictions.

And keeps the models such that

$$\arg\max_{\ell'}(\log(p(\mathcal{T}_{\ell'}|Y,X))) - \log(p(\mathcal{T}_{\ell}|Y,X)) \leq \log(o)$$

where o is Occam's window,[14] and the minimum is over the set of all proposed models.

The original BART-BMA algorithm derived prediction intervals by Gibbs sampling from full conditionals for the model parameters for each selected model. However, the posterior predictive distributions for the selected models are multivariate t-distributions, as the models are Bayesian linear regressions with covariates equal to indicator variables for terminal node parameters. Therefore posterior distributions and credible intervals can be obtained without random number generation (see section 2.3 for further details).

2.3 Improved BART-BMA Algorithm

2.3.1 Summary of Improvements

The BART-BMA algorithm searches for trees to add to sum-of-tree models. The set of potential splitting points to be used in searching for a tree is restricted by applying a grid search algorithm or Pruned Exact Linear Time changepoint detection algorithm to the residuals (Killick et al. 2012, Hernández et al. 2018). [15] [16] The improved implementation differs in the calculation of residuals.

First, the residual from a sum-of-tree model currently in Occam's window is obtained, then the grid search approach considers a fixed number of equally spaced splitting points for each covariate, and orders the potential splitting points by squared error of the predictions of the residual resulting from a binary split. A percentage of splitting points, set by the user, are kept for constructing trees. The original BART-BMA algorithm approximated the residuals of each sum-of-tree model by subtracting single tree predictions each time a tree was appended to the model. The new implementation introduced in this chapter uses residuals from the full sum-of-tree models instead of an approximation.

A notable difference between the original implementation and the new implementation is that the original algorithm estimated the parameters of a new tree by fitting a single tree to the residuals of an existing model. The original implementation therefore does not take account of correlations across trees, nor penalize the contribution of previously added trees to model complexity. [17] The new approach appends a potential tree to the model and re-estimates the entire model. In this sense, the new implementation adjusts the BART-BMA model search in a manner analogous to how methods such as LPBoost adjust AdaBoost by re-estimating coefficients at each step, e.g. by backfitting or linear programming (Freund & Schapire 1995, Freund et al. 1996, Demiriz et al. 2002). This approach is particularly useful for the extension to BCF-BMA, because the parameters of interest are the terminal node parameters of treatment effect trees, and orthogonalization from control trees plays an important role.

Other improvements include bug fixes and more precise calculations of the marginal likelihood and prior.[18]

[14] o can be set arbitrarily or by cross-validation. Computational constraints may also affect the choice of o.

[15] In the first round of the algorithm, when single tree models are created, the changepoint detection algorithm is applied to the outcomes.

[16] For details on how the PELT algorithm is used in BART-BMA, see Hernández et al. (2018),

[17] This is not an issue for the finally chosen models, given that a Gibbs sampler is used for the final estimates in the original implementation, however, it does have implications for the residuals used in the model search algorithm.

[18] The **R** package **bartBMA**, available on CRAN, is based on this improved implementation. Options are included for alternative tree priors described by Quadrianto & Ghahramani (2014) and Rockova & van der Pas (2017). Many parameters options are included that can be used to adjust the model search algorithm. For example, the set of potential splitting points can be updated

The new implementation avoids the use of a Gibbs sampler for calculation of prediction intervals by using the standard closed form of the posterior predictive distribution. While this is not necessary for BART-BMA, it avoids potential issues regarding convergence of the sampler and is particularly useful for BART-IS and BCF-IS, for which a much larger number of models are averaged. For a given sum-of-tree model the posterior distribution for the vector of terminal node parameters is

$$\underline{\boldsymbol{\mu}}|Y,\mathcal{T} \sim MVSt_{\nu+n}\left(M^{-1}W^TY\ ,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]M^{-1}\right)$$

where $M = aI_b + W^TW$.[19] The posterior distribution of $W\underline{\boldsymbol{\mu}} = f(x)$ (for in-sample estimates) is:

$$W\underline{\boldsymbol{\mu}}|Y,\mathcal{T} \sim MVSt_{\nu+n}\left(WM^{-1}W^TY\ ,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]WM^{-1}W^T\right)$$

The posterior predictive (out-of-sample) distribution for a sum-of-tree model is:

$$\tilde{Y}|Y,W,\tilde{W},\mathcal{T}_\mu,\mathcal{T}_\tau \sim MVSt_{\nu+n}\left(\tilde{W}M^{-1}W^TY,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY](I_{\tilde{n}} + \tilde{W}M^{-1}\tilde{W}^T)\right)$$

where the tilde notation indicates numbers or random variables relating to out-of-sample data.

Therefore, unconditional on the model, the posterior predictive distribution for BART-BMA is a posterior model probability weighted mixture of multivariate t-distributions. For pointwise prediction intervals, we only need to obtain the marginal posterior (predictive) distribution, which is (for each model) a univariate t-distribution with location and scale. Then the marginal mixture distribution has a closed form PDF, and a CDF that can be evaluated by numerical integration methods. Prediction intervals can therefore be constructed by obtaining the quantiles of the (marginal) mixture distribution's CDF by a root finding algorithm (e.g. bisection).[20] This approach is also used to obtain prediction intervals for BART-IS and BCF-IS, which involve averaging of a much larger set of predictive distributions.

A number of areas for further research are outlined in appendix B.1. These include methods for setting the regularization parameter a, further improvements to computational methods, testing of alternative priors, and an implementation involving OLS estimation and model weights based on squared errors as in Bayesian Averaging of Classical Estimators Sala-i Martin et al. (2004).

The spike and tree prior (Rockova & van der Pas 2017) can also be applied to the space of sum-of-tree models instead of the standard BART prior. Details for this prior are included in appendix B.5.

2.3.2 BART-BMA for Treatment Effect Estimation

This subsection outlines how BART-BMA can be applied to treatment effect estimation in an approach similar to that described by Hill (2011), but using the conjugate priors of BART-BMA to obtain a closed form posterior distribution for Individual Treatment Effects (ITEs) and the Conditional Average Treatment Effect (CATE).

Following the approach of Hill (2011), let the BART-BMA ITE estimate be defined as $\hat{\tau}(x) = \hat{f}_1(x) -$

after each split is added to a tree, or the same set of points can be used in constructing an entire tree.

[19] An alternative would be to draw $\{\underline{\boldsymbol{\mu}}^{(q)}, \sigma^{2(q)}\}_{q=1}^{Q}$ from $\underline{\boldsymbol{\mu}}^{(q)} \sim MVN(M^{-1}W^TY, \sigma^{2(q)}M^{-1})$ and $\sigma^{2(q)} \sim \Gamma^{-1}\left(\frac{\nu+n}{2}, \frac{\nu\lambda}{2} + \frac{1}{2}[Y^TY - Y^TWM^{-1}W^TY]\right)$

[20] It is also possible to directly sample from the mixture of multivariate t-distributions and obtain pointwise quantiles, or to separately sample from a mixture of univariate t-distributions for each individual in the out-of-sample data.

$\hat{f}_0(x)$, where $\hat{f}(x)$ is obtained from fitting a BART-BMA regression of the outcome on the covariate and treatment (i.e. include the treatment indicator as a covariate). $\hat{f}_1(x)$ ($\hat{f}_0(x)$) is the estimate obtained for covariate vector x when the treatment status is set to 1 (0).

Let $W_1\underline{\boldsymbol{\mu}} = f_1(x)$, and $W_0\underline{\boldsymbol{\mu}} = f_0(x)$, where W_1 is the W matrix obtained if all Z values are reset to 1 (and similarly W_0 is obtained by setting $Z = 0$). Note that some splits can be on Z, and this determines how W changes with Z.

Consider the posterior predictive distribution of the ITE for a given sum-of-trees model.

$$ITE = f_1(x) - f_0(x) = W_1\underline{\boldsymbol{\mu}} - W_0\underline{\boldsymbol{\mu}} = (W_1 - W_0)\underline{\boldsymbol{\mu}}$$

Let $W_{diff} = W_1 - W_0$ Then the in-sample posterior distribution of the vector of ITEs for all individuals (in the sample) is:

$$W_{diff}\underline{\boldsymbol{\mu}}|Y,\mathcal{T} \sim MVSt_{\nu+n}\left(W_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]W_{diff}M^{-1}W_{diff}^T\right)$$

[21] The in-sample posterior distribution of the CATE, $\frac{1}{n}\sum_{i=1}^{n}\tau(x)$, is:

$$\frac{1}{n}\mathbf{1}^TW_{diff}\underline{\boldsymbol{\mu}}|Y,\mathcal{T} \sim$$
$$MVSt_{\nu+n}\left(\frac{1}{n}\mathbf{1}^TW_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\frac{1}{n}\mathbf{1}^TW_{diff}M^{-1}W_{diff}^T\frac{1}{n}\mathbf{1}\right)$$

$\mathbf{1}$ is a vector of ones of length n. [22]

The distribution for the Conditional Average Treatment Effect on the Treated (CATT) can be obtained by replacing $\frac{1}{n}\mathbf{1}^T$ with $\frac{1}{n_{treated}}\mathbf{z}^T$ and the Conditional Average Treatment Effect on the Not Treated (CATNT) distribution can be obtained using $\frac{1}{n_{control}}(\mathbf{1}-\mathbf{z})^T$.

2.4 BART-IS

This section presents BART-IS, which extends the importance sampling approach described by Quadrianto & Ghahramani (2014) from single classification trees to sums of regression trees by utilising the conjugate priors of BART-BMA.

Importance sampling of Bayesian linear regression models involves constructing weights by dividing the prior model probability by the model sampling probability. Therefore the model prior and importance sampler probabilities do not need to be calculated when the models are sampled from the prior. This approach is used by Quadrianto & Ghahramani (2014) in safe-Bayesian Random Forests for classification, and by Sala-i Martin et al. (2004) in their implementation of BMA of linear regressions. For completeness, we provide the option of using different samplers and priors in the **safeBart** package.[23]

[21] For out-of-sample ITEs, let $\tilde{W}_{diff} = \tilde{W}_1 - \tilde{W}_0$. Then $\tilde{W}_{diff}\underline{\boldsymbol{\mu}}|Y,\mathcal{T} \sim MVSt_{\nu+n}\left(\tilde{W}_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\tilde{W}_{diff}M^{-1}\tilde{W}_{diff}^T\right)$. Note that the error term does not enter $f_1(x) - f_0(x)$ and therefore there is no $I_{\tilde{n}}$ term in the variance of the out-of-sample posterior distribution.

[22] The out-of-sample posterior distribution of the CATE is: $\frac{1}{\tilde{n}}\tilde{\mathbf{1}}^T\tilde{W}_{diff}\underline{\boldsymbol{\mu}}|Y,\mathcal{T} \sim$
$MVSt_{\nu+n}\left(\frac{1}{\tilde{n}}\tilde{\mathbf{1}}^T\tilde{W}_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\frac{1}{\tilde{n}}\tilde{\mathbf{1}}^T\tilde{W}_{diff}M^{-1}\tilde{W}_{diff}^T\frac{1}{\tilde{n}}\tilde{\mathbf{1}}\right)$ where $\tilde{\mathbf{1}}$ is a vector of ones of length $\tilde{n}$.

[23] The package is publicly available at https://github.com/EoghanONeill/safeBart .

Bayesian Model Averaging tends towards one model as the number of observations tends to infinity. However, when the model space does not contain the true model, more accurate predictions can be obtained from Bayesian Model Combination, which tends towards a combination of models. Quadrianto & Ghahramani (2014) apply a standard model combination approach by raising the model likelihoods to a power. This makes the approach "safe" in the sense that it does not tend towards one possibly wrong model. The option of raising the likelihood to a power is provided in the **safeBart** package. However, for a fair comparison of BART implementations, the likelihood is not raised to a power in the results presented in this paper.

Preprocessing involves a probability integral transformation of each covariate, with the distribution equal to the empirical cumulative distribution function. The BART-IS algorithm randomly samples all trees in each sum-of-tree model from the independent tree prior, and calculates the marginal likelihood and predictions for each sum-of-tree model. The likelihoods can be raised to a power for a safe-Bayesian approach. The final predictions are a marginal-likelihood weighted average.

The BART-IS algorithm is generalizable in the sense that the prior tree model distribution can be replaced by any prior on partitions of the covariate space. The partitions do not need to be representable in binary tree structures. Provided it is possible to (quickly) draw partitions and construct indicator variables for sets in the partitions, this approach is applicable. Then, for a drawn model, any conjugate Bayesian linear regression priors can be applied given a set of indicator variables as covariates.

BART-IS is applicable to ITE estimation using the distributions outlined in 2.3.2.[24] In principle, BART-IS can also be applied to data with multiple outcomes by applying standard conjugate priors for Bayesian multivariate linear regression. This approach is outlined in appendix B.3. Table 2.2 extends a table from He et al. (2019) to provide a comparison between the methods discussed in this chapter and other tree-based methods.

2.4.1 Description of the BART-IS Algorithm

1. Sample sets of trees from a prior. The prior can be the standard BART prior (Chipman et al. 2010), the prior described by Quadrianto & Ghahramani (2014), or the spike-and-tree prior (Rockova & van der Pas 2017).

2. Obtain the model predictions. If computational speed is desired, particularly for a large number of samples, or for models with many trees, a fast ridge regression algorithm can be applied for model predictions.

3. Obtain model weights. This can optionally involve raising the marginal likelihood to a power, as described by Quadrianto & Ghahramani (2014).[25] If importance sampling is not from the prior, then the likelihood is multiplied by the ratio of the prior model probability to the importance sampler model probability. [26]

4. Obtain the predictive distribution, which is a mixture of multivariate t-distributions. See section 2.3.

[24] This approach to ITE estimation is available in the **safeBart** package available at https://github.com/EoghanONeill/safeBart .

[25] This is because it is possible that none of the set of models is the true model, but BMA tends towards placing all the weight on one model. In practice a Bayesian Model Combination approach, such as the power likelihood approach, might be more accurate.

[26] Alternatively, the construction of weights from residuals instead of the marginal likelihood may also increase computational speed.

It is possible to quickly sample from the prior described by Quadrianto & Ghahramani (2014) and the standard BART prior (Chipman et al. 2010). Appendix B.5.2 contains an outline of how to sample from a spike and tree prior.

	CART	ERT	RF	XGB	XBART	BART-MCMC	BART-BMA	BART-IS
Deterministic	Yes	No	No	No	No	No	Yes	No
Data independent model draws	No	Yes	No	No	No	No	No	Yes
Parallelizable	No	Yes	Yes	limited	limited	limited	limited	Yes
Sequential fitting	No	No	No	Yes	Yes	Yes	Yes	No
Recursion	Yes	Sampling	Yes	Yes	Yes	No	Yes	Sampling
Leaf parameters	optimized with splits	optimized with splits	optimized with splits	optimized with splits	integrate out at split, sample	integrate out at split, sample	integrate out, sampling unnecessary	integrate out, sampling unnecessary
Criteria	Likelihood	Likelihood	Likelihood	Likelihood	Marginal Likelihood	Marginal Likelihood	Marginal Likelihood	Marginal Likelihood

Table 2.2: Comparison of tree-based machine learning algorithms.
CART = Classification and Regression Trees (Breiman et al. 1984), ERT = Extremely Randomized Trees (Geurts et al. 2006), RF = Random Forests (Breiman 2001), XGB = Gradient Boosted Trees (Breiman 1997, Friedman 2001), XBART = Accelerated Bayesian Additive Regression Trees (He et al. 2019, 2018).

2.5 Results for BART-BMA and BART-IS

This section contains the results from the application of the improved BART-BMA algorithm and BART-IS to the data generating process used by Chipman et al. (2010) and Hernández et al. (2018). Section 2.5.1 presents the results for high-dimensional data and section 2.5.2 presents the results for low-dimensional data.

2.5.1 High-Dimensional Data

Figure 2.1 presents the results obtained by applying the following methods to to the commonly used simulations introduced by Friedman et al. (1991): BART-BMA with the standard BART model prior, BART-BMA with the spike and tree prior,[27] BART-IS, BART with 1000 and 10,000 MCMC draws, Dirichlet BART with 1000 and 10,000 MCMC draws, and Random Forests. [28]

The outcome depends on 5 uniformly distributed predictor variables $x_1, x_2, ..., x_5$:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$$

[27]The BART-BMA results presented here are for BART-BMA with the gridpoint method for changepoint detection. Another option is the Pruned Exact Linear Time algorithm (Killick et al. 2012). The results presented here are for BART-BMA with no within-tree updating of potential split points. Another option is to update potential splitting points after a split is added to each tree.

[28]BART-IS was implemented with 1,000,000 draws of sum-of-tree models each containing 30 trees. Each of the 10,000 BART and DART sum-of-tree models contained 200 trees. Many more draws were made for BART-IS because trivially parallelizable data-independent draws can be made much faster than MCMC draws. Furthermore data independent draws of models can potentially be made offline before any data is obtained.

where $\varepsilon \sim \mathcal{N}(0,1)$. Variables $x_6, ..., x_p$ are uniformly distributed. The number of observations is 500. I considered 5 different values of p, the number of covariates: $p = (100, 1000, 5000, 10000, 15000)$. The RMSE were obtained using fivefold cross-validation. The default parameter values were used for RF, BART, and Dirichlet BART (DART).[29]

The results for the new variations of BART-BMA compare favourably to the results obtained for the original implementation (Hernández et al. 2018), which gives RMSE between 2.9 and 3. [30] BART-BMA with the grid-search changepoint detection algorithm and standard priors has RMSE which does not appear to deteriorate as the number of variables increases. BART-BMA variations that update the set of potential splitting points within the construction of individual trees exhibit deteriorating performance as the number of variables is increased.[31]

BART and Dirichlet BART under default parameter settings do not perform very well when the number of variables is increased to 5000. However, the default implementation of BART and DART includes only 1000 draws from the posterior with 100 burn-in draws. Figure 2.1 demonstrates that BART and DART exhibit much better performance in high dimensional data when the number of MCMC samples is increased to 10,000. [32] The superior performance of DART is perhaps unsurprising given that DART involves a sparsity-inducing hyperprior on the probabilities of splitting variables. Therefore the fairer comparison is arguably between BART-MCMC and BART-BMA, and the results confirm that BART-BMA delivers results comparable to BART, as intended. However, BART-BMA performs very well in terms of variable selection, and an alternative explanation, particularly given the better performance of BART-MCMC in low-dimensional data, is that the small number of trees used by default in BART-BMA and BART-IS within sum-of-tree models is insufficient to model the complex functional form in this particular example.[33] It is possible that BART-BMA and BART-IS would produce better results for higher numbers of trees. [34]

BART can be used for variable selection (Linero 2018, Bleich et al. 2014). The Brier scores for the BART, DART, and BART-BMA posterior variable inclusion probabilities (PIP) are given in table 2.3. [35] The Brier score is defined as $\frac{1}{P}\sum_{p=1}^{P}(I_p - PIP_p)^2$ where p indexes the covariates, $I_p = 1$ for truly important variables $x_1, ..., x_5$ and $I_p = 0$ otherwise. The results suggest that BART-BMA outperforms BART and DART in terms of variable selection. The spike-and-tree prior outperforms the standard BART prior.

Prediction intervals obtained directly from the closed form for the point-wise predictive distributions were obtained for the Friedman data simulations, and the results for 95% prediction intervals are presented for BART-BMA, BART, and DART in tables 2.4 and 2.5. BART-BMA gives more precise prediction intervals than BART and DART. BART-IS intervals have comparable coverage to BART-MCMC and DART-MCMC, although the intervals for DART and BART are notably narrower for low dimensional simulations. [36]

[29] Random Forest was implemented using the **R** package **ranger**. BART was implemented using the **wbart** function in the **R** package **BART**. DART was implemented using the **wbart** function in the **R** package **BART** with the following parameter setting sparsity = TRUE.

[30] See original paper by Hernández et al. (2018). The RMSEs for the old BART-BMA implementation in figure 2.1 are approximate readings from the corresponding table in the original paper.

[31] The results for BART-BMA with updating of splitting points within the construction of trees are not included in Figure 2.1.

[32] A comparison of computational times is included in appendix B.2.

[33] Chipman et al. (2010) noted the trade-off between the predictive accuracy of models containing a few hundred trees, and the impressive variable importance results from sum-of-tree-models containing 5, 10 or 20 trees.

[34] Preliminary results (not included in this paper) indicate that BART-IS produces more accurate results when the number of trees is set to a few hundred, even if a smaller number of models is sampled. However, for this to be computationally inexpensive this would require implementation of fast ridge regression or Bayesian linear regression, possibly with approximations, for each sum-of-tree model.

[35] The results in table 2.3 are for BART and DART with 10,000 draws.

[36] Also, the BART and DART results might improve with more MCMC draws as this would allow for convergence of the

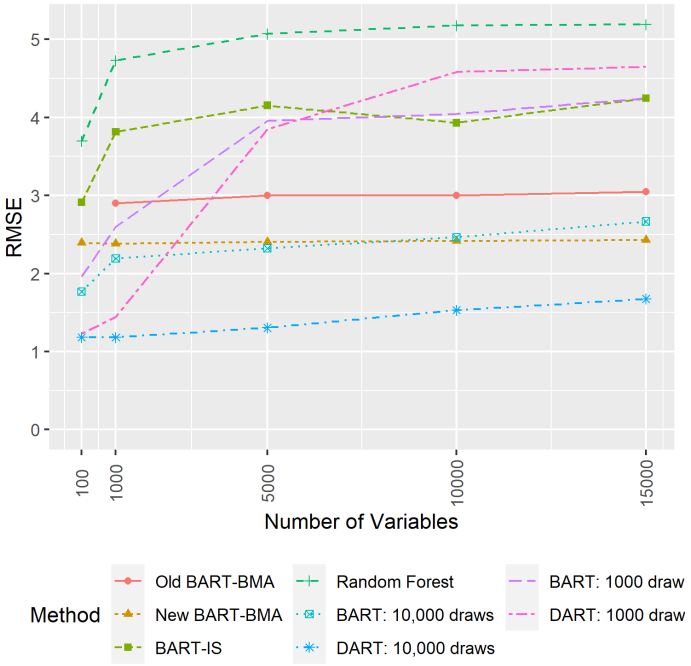


Figure 2.1: RMSEs for High-dimensional Friedman Data Simulations

Number of Variables	Old BART-BMA	New BART-BMA Standard	New BART-BMA Spike-and-tree	BART	DART
100	NA	2.000×10^{-3}	3.560×10^{-32}	7.005×10^{-1}	3.161×10^{-3}
1000	3.26×10^{-3}	4.000×10^{-4}	1.350×10^{-31}	3.974×10^{-2}	7.210×10^{-5}
5000	6.55×10^{-4}	2.000×10^{-4}	2.700×10^{-32}	3.236×10^{-3}	1.513×10^{-4}
10000	3.28×10^{-4}	1.000×10^{-4}	1.350×10^{-32}	1.150×10^{-3}	1.448×10^{-4}
15000	2.18×10^{-4}	8.000×10^{-5}	9.000×10^{-33}	6.648×10^{-4}	1.120×10^{-4}

Table 2.3: Brier Scores for Friedman data simulations

2.5.2 Low-Dimensional Data

Figure 2.2 presents the results for the Friedman simulations described in section 2.5.1, with the number of covariates, p equal to 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The RMSE was averaged across five simulations for each value of p.

BART and DART outperform other methods in terms of RMSE. It can be observed that the predictions of RF, BART, and BART-IS become less accurate as the number of covariates increases. However, it is likely that the accuracy of these methods when applied to high dimensional data would improve with a greater number of draws of models. It is unsurprising that the RMSE of BART-IS predictions degrades as the dimensionality of the data increases. Importance sampling is known to suffer from the curse of dimensionality.[37]

Markov Chain, and more accurate estimation of quantiles of the posterior distribution. The chosen number of MCMC draws for BART and DART is 10,000. For each draw of a sum-of-tree model, 10 draws of the additive error term, ε were made from a normal distribution. It is likely that more accurate intervals could be obtained with a greater number of draws of the error.

[37]The model space is very high-dimensional and further research is required in order to establish the effective (i.e. equivalent to exact model posterior) sample size corresponding to draws from the importance sampler.

Number of Variables	BART	DART	Old BART-BMA	New BART-BMA Standard	New BART-BMA Spike and tree	BART-IS
100	**95.0**	94.4	NA	95.4	94.6	97.4
1000	97.4	96.8	94.4	95.8	**94.6**	97.4
5000	97.0	97.0	93.8	**95.4**	**94.6**	**95.4**
10000	97.6	98.4	94.0	**94.8**	94.6	97.6
15000	98.8	98.2	94.0	**94.8**	94.6	94.0

Table 2.4: Average 95% prediction interval coverage for Friedman data simulations

Number of Variables	BART	DART	Old BART-BMA	New BART-BMA Standard	New BART-BMA Spike and tree	BART-IS
100	6.74	4.77	NA	9.62	10.11	12.65
1000	9.24	5.07	11.69	9.61	10.24	15.81
5000	10.70	6.00	11.67	9.64	10.24	16.10
10000	12.11	8.15	11.66	9.61	10.24	16.61
15000	12.89	10.13	11.68	9.61	10.24	16.39

Table 2.5: Average 95% prediction interval width for Friedman data simulations

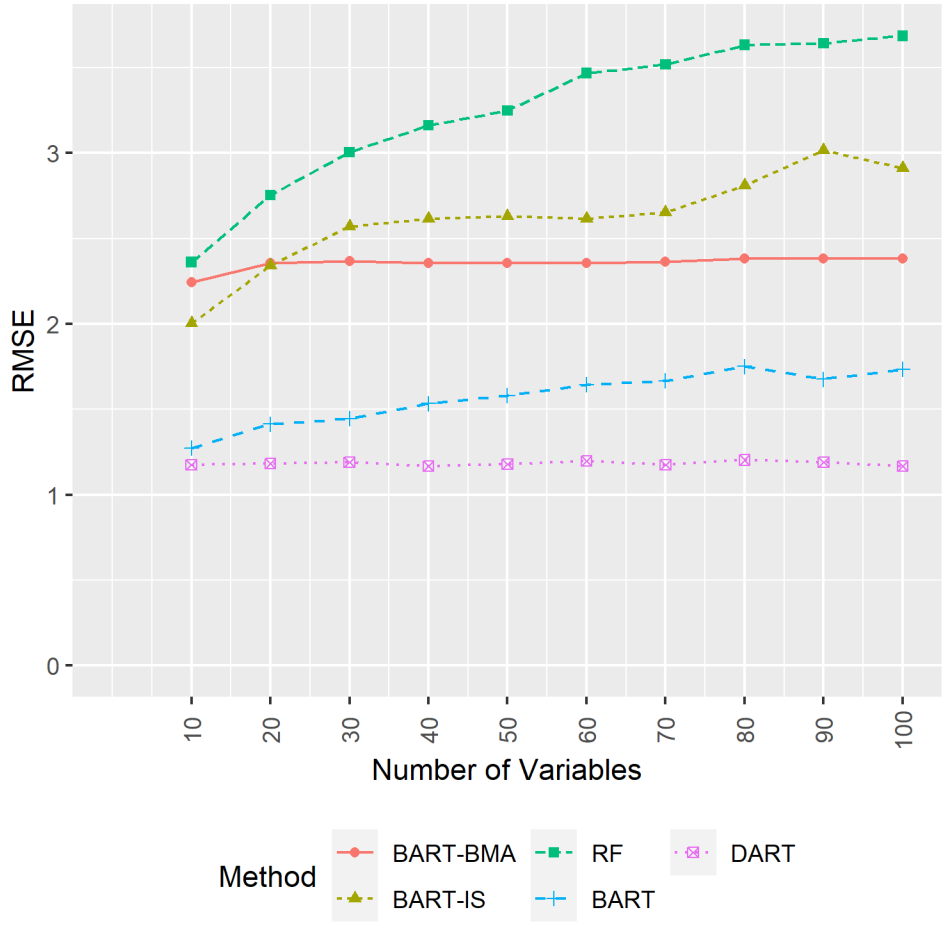


Figure 2.2: RMSEs for Low-dimensional Friedman Data Simulations

2.6 BCF-BMA and BCF-IS

This section introduces a combination of the BCF parameterization of BART for treatment effect estimation (Hahn et al. 2020) and the BART-BMA model search implementation of BART (Hernández et al. 2018). [38] Section 2.6.1 reviews BCF, section 2.6.2 describes the BCF-BMA model, and section 2.6.3 outlines the BCF-BMA algorithm.[39] Bayesian Causal Forests using Importance Sampling (BCF-IS) is briefly described in section 2.6.4. Finally results are presented for ITE estimation on simulated data, giving a comparison between BCF-BMA, BART-BMA, BART-IS, BCF-IS and existing state-of-the-art methods BCF, BART, and causal forests (Wager & Athey 2018, Athey et al. 2019).

2.6.1 BCF

BCF controls for confounding by including the estimated propensity score as a splitting variable, and allows the treatment effect function to be regularized separately to the rest of the model.

BCF Summary

Hill (2011) proposed the use of BART to estimate treatment effects by including the treatment variable Z in the set of splitting variables, and estimating the model $Y_i = f(x_i, Z_i) + \epsilon_i,\ \epsilon \sim N(0, \sigma^2)$. The treatment effect can be expressed as $\tau(x_i) = f(x_i, 1) - f(x_i, 0)$. If an individual has a vector of covariates x, then the difference in predictions for $(X = x,\ Z = 1)$, and $(X = x,\ Z = 0)$ is the estimated treatment effect. [40]

However, the implications of the prior on $f(x, z)$ for the induced prior on τ are difficult to understand, and the induced prior on τ will vary with the number of covariates. Furthermore, the estimates can be biased in the presence of confounding. Hahn et al. (2020) propose an alternative approach, and elaborate on an issue referred to as "Regularization Induced Confounding" (Hahn et al. 2018). Regularization priors tend to adversely bias treatment effect estimates by over-shrinking control variable regression coefficients. In the presence of confounding, the finite sample bias of the treatment effect estimator will be influenced by the prior regularization, and it is desirable to directly control regularization of the treatment effect function. This emphasis on separately regularizing the prognostic effect and treatment effect functions is related to other methods, including double machine learning (Belloni et al. 2014, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins et al. 2017, Yang et al. 2015).

Confounding can be mitigated by including the estimated propensity score as a potential splitting variable (Hahn et al. 2020). Hahn et al. (2020) propose a re-parameterization that allows for an independent prior to be placed on τ and also include the estimated propensity score, $\hat{\pi}_i$, as a potential splitting variable.

$$f(x_i, z_i) = \mu(x_i, \hat{\pi}_i) + \tau(x_i) z_i$$

where $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ are both sums of trees.

Different BART parameters (e.g. the number of trees, depth penalty, splitting probability, scale of terminal node outputs) are used for the sums of trees denoted by $\mu(x_i)$ and $\tau(x_i)$, and $\tau(x_i)$ priors are set such that it is more strongly regularized than $\mu(x_i)$.

[38] The **R** package for BCF-BMA is publicly available at https://github.com/EoghanONeill/bcfbma .

[39] See appendix B.6 for more details on the BCF-BMA algorithm.

[40] Another common approach is to separately fit a model on observations for which $z_i = 1$, and on observations for which $z_i = 0$, and let $\hat{\tau}_i$ be the difference in the predictions of these two models.

Hahn et al. (2020) demonstrate that BCF can perform well in simulations in terms of MSE of individual treatment effect estimates relative to: BART (Hill 2011), including the propensity score estimates in standard BART, fitting BART separately to treated and control groups, and causal forests (Wager & Athey 2018, Athey et al. 2019).

BCF Priors

Chipman et al. (2010) assume that the model precision σ^{-2} has a conjugate prior distribution $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$ with degrees of freedom v and scale λ. The same prior is used for the model precision in BCF.

The probability of a single tree structure is $p(T_j) = \prod_{h=1}^{b_j-1} \alpha(1+d_h)^{-\beta} \prod_{k=1}^{b_j}(1-\alpha(1+d_k)^{-\beta})$, where h indexes the internal nodes of the tree T_j, and k indexes the terminal nodes. Different splitting probabilities are applied to $\mu(x)$ and $\tau(x)$ trees. In particular, $\alpha = 0.95$ and $\beta = 2$ for $\mu(x)$ trees, and $\alpha = 0.25$ and $\beta = 3$ for $\tau(x)$ trees. This regularizes the treatment effect function to a greater extent than the rest of the model.[41]

2.6.2 Outline of BCF-BMA

BCF (Hahn et al. 2020) is an average of models of the form $f(x_i, z_i) = \mu(x_i, \hat{\pi}_i) + \tau(x_i)z_i$, where $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ are separate sum of tree models.[42] Let $T_{\mu j}$ and $T_{\tau j}$ denote trees in $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ respectively, and let $M_{\mu j}$ and $M_{\tau j}$ denote the terminal node parameters for $T_{\mu j}$ and $T_{\tau j}$ respectively. The BCF prior can be written as:

$$p(M_{\mu 1}, ..., M_{\mu m_\mu}, T_{\mu 1}, ..., T_{\mu m_\mu}, M_{\tau 1}, ..., M_{\tau m_\tau}, T_{\tau 1}, ..., T_{\tau m_\tau}, \sigma)$$

$$\propto \left[\prod_j \prod_i p(\mu_{ij}|T_{\mu j})p(T_{\mu j})\right] \left[\prod_j \prod_i p(\tau_{ij}|T_{\tau j})p(T_{\tau j})\right] p(\sigma)$$

For BCF-BMA, I suggest placing the prior $\mu_{ij}|T_\mu, \sigma \sim N(0, \frac{\sigma^2}{a_\mu})$ and $\tau_{ij}|T_\tau, \sigma \sim N(0, \frac{\sigma^2}{a_\tau})$. These priors are somewhat different to those proposed by Hahn et al. (2020) who place different priors on the scales of μ_{ij} and τ_{ij}. Different scales are directly specified through the choice of a_μ and a_τ. The BCF-BMA prior, like the BART-BMA prior, provides a closed form for the marginal likelihood and a multivariate t-distribution for posterior predictions.

BCF-BMA Marginal Likelihood

Let $Z = (Z_1, ..., Z_n)$ be the treatment indicator variable. Let $J_{\mu j}$ and $J_{\tau j}$ be matrices denoting inclusion of observations in terminal nodes of tree j in $\mu(x)$ and $\tau(x)$ respectively. The BCF-BMA likelihood can be written as:

$$Y|\mathcal{T}_\mu, M_\mu \mathcal{T}_\tau, M_\tau, \sigma^{-2} \sim N\left((\sum_{j=1}^{m_\mu} J_{\mu j} M_{\mu j}) + Diag(Z)(\sum_{j=1}^{m_\tau} J_{\tau j} M_{\tau j}), \sigma^2 I\right)$$

Now, let $W_\mu = [J_{\mu 1} ... J_{\mu m_\mu}]$ be an $n \times b_\mu$ matrix, where $b_\mu = \sum_{j=1}^{m_\mu} b_{\mu j}$, and let $\underline{\boldsymbol{\mu}} = [M_{\mu 1}^T ... M_{\mu m_\mu}^T]^T$ be an $b_\mu \times 1$ vector. Similarly let $W_\tau = [J_{\tau 1} ... J_{\tau m_\tau}]$ be an $n \times b_\tau$ matrix, where $b_\tau = \sum_{j=1}^{m_\tau} b_{\tau j}$, and let

[41] Hahn et al. (2020) also suggest simply including the estimated propensity score as a potential splitting variable in standard BART, and then using the approach introduced by Hill (2011). Therefore, later in this section, there is a similar comparison between BCF-BMA and standard BART-BMA with the estimated propensity score included as a potential splitting variable.

[42] The BCF-BMA package allows for the inclusion of zero, one, or more than one set of propensity score estimates as potential splitting variables in the $\mu(x)$ function.

$\underline{\boldsymbol{\tau}} = [M_{\tau 1}^T ... M_{\tau m_\tau}^T]^T$ be an $b_\tau \times 1$ vector. Then we can write

$$Y|\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\tau}}, \sigma^{-2} \sim N\left(W_\mu \underline{\boldsymbol{\mu}} + Diag(Z)W_\tau \underline{\boldsymbol{\tau}}, \sigma^2 I\right)$$

Now let $W_{BCF} = [W_\mu \ (Diag(Z)W_\tau)]$ be an $n \times (b_\mu + b_\tau)$ matrix, and let $\underline{\boldsymbol{\theta}} = [\underline{\boldsymbol{\mu}}^T \underline{\boldsymbol{\tau}}^T]^T$ be a $(b_\mu + b_\tau) \times 1$ matrix. Then $Y|\underline{\boldsymbol{\theta}}, \sigma^2 \sim N(W_{BCF}\underline{\boldsymbol{\theta}}\ ,\ \sigma^2 I)$, and the BCF-BMA marginal likelihood is:

$$p(Y|X, \mathcal{T}_\mu, \mathcal{T}_\tau) = \int\int p(Y|\underline{\boldsymbol{\theta}}, \sigma^{-2}) p(O) p(\sigma^{-2}) dO d\sigma^{-2}$$

The first b_μ elements of $\underline{\boldsymbol{\theta}}$ have independent prior distributions $\mu \sim N(0, \frac{\sigma^2}{a_\mu})$, and the last b_τ elements of $\underline{\boldsymbol{\theta}}$ also have independent normal priors, with different variance, $\tau \sim N(0, \frac{\sigma^2}{a_\tau})$. This implies that $\underline{\boldsymbol{\theta}}|\sigma^{-2} \sim N(0, \sigma^2 A^{-1})$ where $A = \begin{pmatrix} a_\mu I_{b_\mu} & 0 \\ 0 & a_\tau I_{b_\tau} \end{pmatrix}$ is a diagonal matrix with the first b_μ diagonal elements equal to a_μ, and the next b_τ diagonal elements equal to a_τ.

Therefore $Y|\sigma^{-2} \sim \text{W}_{\text{BCF}}\ \varepsilon_2 + \varepsilon_1$, where $\varepsilon_1 \sim N(0, \sigma^2 I)$ and $\varepsilon_2 \sim N(0, \sigma^2 A^{-1})$. This implies that $Y|\sigma^{-2} \sim N(0, \sigma^2(I_n + WA^{-1}W^T))$ and therefore marginalization over σ gives

$$Y \sim MVSt_\nu\left(0, \lambda(I_n + WA^{-1}W^T)\right)$$

$$p(Y) = \frac{1}{(det(A))^{-\frac{1}{2}}\, det(I_n + WA^{-1}W^T)^{\frac{1}{2}}} \left[\lambda\nu + Y^TY - Y^TW(A + W^TW)W^TY\right]^{-\frac{\nu+n}{2}}$$

And the log of this expression is the log marginal likelihood:

$$\frac{b_\mu}{2}\log(a_\mu) + \frac{b_\tau}{2}\log(a_\tau) - \frac{1}{2}\log(det(M)) - \frac{\nu+n}{2}\log\left(\lambda\nu + Y^TY - Y^TWM^{-1}W^TY\right)$$

where $M = A + W^TW$.

BCF-BMA Posterior ITE Distribution

Let $V = [\mathbf{0}_{n\times b_\mu}\ W_\tau]$, where $\mathbf{0}_{n\times b_\mu}$ is a matrix of zeros of dimensions equal to those of W_μ. The posterior distribution of $\tau(x)$ is:

$$V\underline{\boldsymbol{\theta}} \quad | \quad Y, \mathcal{T}_\mu, \mathcal{T}_\tau \quad \sim \quad MVSt_{\nu+n}\left(VM^{-1}W^TY,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]VM^{-1}V^T\right)$$

where $M = A + W^TW$. For out of sample predictions, replace V with $\tilde{V} = [\mathbf{0}_{\tilde{n}\times b_\mu}\ \tilde{W}_\tau]$.

BCF-BMA CATE Posterior Distribution

The posterior distribution of $\tau(x)$ given in the previous subsection is the posterior distribution of what is often referred to as the Individual Treatment Effect (ITE). However, $\tau(x)$ can also be referred to as the Conditional Average Treatment Effect (CATE) Function. In this paper, the term CATE refers to the expectation of the average of the ITEs, i.e. $\frac{1}{n}\sum_{i=1}^n \tau(x)$.

The posterior distribution of $\frac{1}{n}\sum_{i=1}^{n}\tau(x)$ for a given model in BCF-BMA is:

$$\frac{1}{n}\mathbf{1}^T V\underline{\boldsymbol{\theta}} \mid Y, \mathcal{T}_\mu, \mathcal{T}_\tau \sim$$
$$MVSt_{\nu+n}\left(\frac{1}{n}\mathbf{1}^T VM^{-1}W^T Y\ ,\ \frac{1}{\nu+n}[\nu\lambda + Y^T Y - Y^T WM^{-1}W^T Y]\frac{1}{n}\mathbf{1}^T VM^{-1}V^T\frac{1}{n}\mathbf{1}\right)$$

where $M = A + W^T W$ and $\mathbf{1}$ is a vector of 1s of length n. Note that this is a univariate t-distribution with location and scale. For out of sample predictions, replace V with $\tilde{V} = [\mathbf{0}_{\tilde{n}\times b_\mu}\ \tilde{W}_\tau]$ and replace $\frac{1}{n}\mathbf{1}$ with $\frac{1}{\tilde{n}}\tilde{\mathbf{1}}$, where $\tilde{\mathbf{1}}$ is a vector of 1s of length $\tilde{n}$.

2.6.3 Description of the BCF-BMA Algorithm

The BCF-BMA model search algorithm is similar to the improved BART-BMA algorithm, except in each round either $\mu(x)$ trees or $\tau(x)$ trees can be appended to existing models.[43] The model selection criterion is the posterior model probability. In constructing $\tau(x)$ trees to be potentially appended to the model, potential splitting points are selected from a changepoint detection algorithm applied to treated observations only. [44] Pseudocode for the BCF-BMA algorithm is given in Appendix B.6.

2.6.4 BCF-IS

The BCF-IS algorithm is the algorithm outlined in section 2.4.1, with some adjustments. The $\mu(x)$ trees and $\tau(x)$ trees are drawn from separate priors. The marginal likelihood is the same as that described in section 2.6.2. The BCF-IS algorithm is intended for estimation of treatment effects, not the outcome.

The default model prior for BCF-IS is the standard BART prior, and different priors can be applied to $\mu(x)$ trees and $\tau(x)$ trees as described for BCF-BMA. Similarly, for the prior described by Quadrianto & Ghahramani (2014), different splitting probabilities can be specified for $\mu(x)$ trees and $\tau(x)$ trees. For the Spike and Tree prior (Rockova & van der Pas 2017), different prior parameters can be specified for the Poisson distribution for the number of terminal nodes, and different hyperparameters can be specified for the beta hyperprior distribution on the variable inclusion probabilities.

2.6.5 BCF-BMA and BCF-IS Results for Simulated Datasets

Simulation from `bcf` R Package

This section contains a comparison of BCF-BMA and standard BCF (Hahn et al. 2020) using a simulation example from the **bcf** package in **R**.

The simulated dataset contains n observations of p standard normally distributed covariates $x_1, ..., x_p$. The outcome is set equal to

$$Y = \mu(x) + \tau(x)T + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma)$ and $\sigma = max(\mu(x_i) + \tau(x_i)\pi(x_i)) - min(\mu(x_i) + \tau(x_i)\pi(x_i))$.

$$\mu(x) = -\mathbb{I}\{x_1 > x_2\} + \mathbb{I}\{x_1 < x_2\}$$

[43] An option is also provided in the BCF-BMA package for adding a mu tree, then a tau tree, and then a mu tree, and so on in an alternating sequence.

[44] Another option, provided in the **R** package **bcfbma** available at https://github.com/EoghanONeill/bcfbma, is to apply the changepoint detection algorithm to Horowitz-Thompson transformed residuals

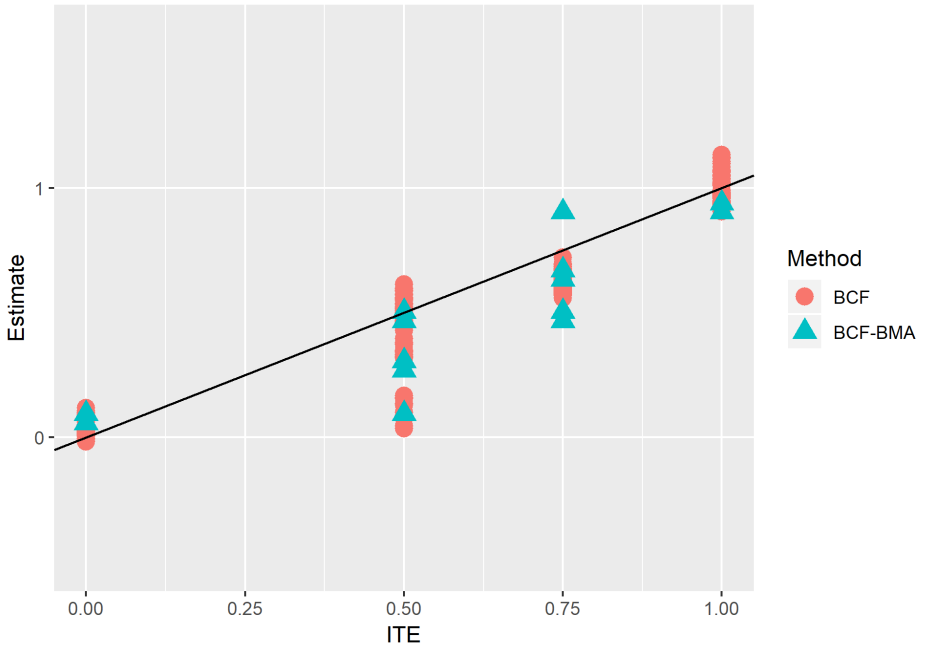


Figure 2.3: Example results for BCF and BCF-BMA. True ITE on x-axis, estimated ITE on y-axis.

where $\mathbb{I}$ is an indicator function. Let the probability of treatment be $\pi(x) = \Phi(\mu(x))$. i.e. there is confounding. The treatment variable is Z. Let the treatment effect function be

$$\tau(x) = 0.5\mathbb{I}\{x_3 > -0.75\} + 0.25\mathbb{I}\{x_3 > 0\} + 0.25\mathbb{I}\{x_3 > 0.75\}$$

Suppose the propensity score estimates are exact, i.e. the true propensities are known $\hat{\pi}(x) = \pi(x)$.

The results for one simulation of the data generating process outlined above with $n = 250$ and $p = 3$ are included in Figure 2.3. It can be observed that BCF and BCF-BMA yield similar predictions. BCF-BMA has the added advantage that the output contains a small number of models, and each model (under the default settings) contains only 5 $\mu(x)$ trees and 5 $\tau(x)$ trees, each of which contains a small number of splits.[45] Therefore it is possible to directly observe the important splitting variables and splitting points from the tree structures in the output of the algorithm.

Simulations used by Hahn et al. (2020)

Hahn et al. (2020) simulate the following eight data generating processes, corresponding to the various combinations of three two-level settings: homogeneous versus heterogeneous treatment effects, a linear versus nonlinear conditional expectation function, and two different sample sizes ($n = 250$ and $n = 500$).

Five variables comprise x; the first three are continuous, drawn as standard normal random variables, the fourth is a dichotomous variable and the fifth is unordered categorical, taking three levels (denoted 1,2,3). The treatment effect is either $\tau(x) = 3$ (i.e. homogenous) or $\tau(x) = 1 + 2x_2x_5$ (i.e. heterogeneous). The prognostic function is either $\mu(x) = 1 + g(x_4) + x_1x_3$ (linear) or $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$ (nonlinear)

[45] The maximum number of splits under the default settings for BCF-BMA is 5 per tree.

where $g(1) = 2$, $g(2) = -1$, and $g(3) = -4$, and the propensity function is

$$\pi(x_i) = 0.8\Phi(3\mu(x_i)/s - 0.5x_1) + 0.05 + u_i/10$$

where s is the standard deviation of μ taken over the observed sample and $u_i \sim Uniform(0, 1)$. The variance of the additive Gaussian error term is set equal to 1.

Comparisons with other methods are made in tables 2.6 to 2.9 in terms of RMSE, coverage, and interval length for Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE) estimates. Tables 2.6 to 2.9 contain results for $n = 250$. The set of methods includes: BCF,[46] BCF-BMA without updates of potential splitting variables within trees (BCF-BMA 1), BCF-BMA with updates of potential splitting variables within trees (BCF-BMA 2), BART-BMA [47], BART,[48] BCF-IS,[49], BART-IS, [50] and standard causal forests (Athey et al. 2019, Wager & Athey 2018). [51]

For all methods except causal forest, the propensity score is estimated using the function **pbart** in the **R** package **BART**.

	ATE			ITE		
	RMSE	coverage	length	RMSE	coverage	length
BCF	0.24	0.86	0.86	0.45	0.97	1.99
BCF-BMA 1	0.40	0.69	0.99	0.58	0.78	1.22
BCF-BMA 2	0.22	0.92	0.92	0.34	0.90	1.06
BART-BMA	0.30	0.84	1.04	0.51	0.86	1.64
BART	0.23	0.87	0.83	0.37	0.97	1.74
BCF-IS	0.27	0.90	1.00	0.34	0.98	1.74
BART-IS	0.29	0.88	1.03	0.35	0.97	1.62
CF	0.41	0.67	1.08	0.53	0.78	1.35

Table 2.6: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = 1 + g(x_4) + x_1x_3$, $n = 250$, 200 replications.

	ATE			ITE		
	RMSE	coverage	length	RMSE	coverage	length
BCF	0.19	0.95	0.85	0.50	0.96	1.91
BCF-BMA 1	0.24	0.98	1.33	0.78	0.86	1.86
BCF-BMA 2	0.25	0.98	1.31	0.73	0.87	1.79
BART-BMA	0.45	0.72	1.38	1.10	0.81	2.17
BART	0.19	0.94	0.88	0.43	0.98	2.20
BCF-IS	0.24	0.98	1.35	0.39	0.99	2.34
BART-IS	0.19	0.99	1.38	0.39	1.00	2.90
CF	0.60	0.49	1.26	0.67	0.66	1.57

Table 2.7: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 250$, 200 replications.

[46] Implemented with the **R** package **bcf** with default parameter values.
[47] See section 2.3.2
[48] Implemented with the **R** package **BART** with default parameter values.
[49] Implemented with 100,000 draws of models from the importance sampler. Each model contains 50 $\mu(x)$ trees and 25 $\tau(x)$ trees.
[50] Implemented with 100,000 draws of models from the importance sampler, each model includes 30 trees, with the propensity score and treatment included as potential splitting variables. See section 2.3.2 for details on the posterior distributions for individual models.
[51] Causal forests are estimated using the **R** package **grf** and 4000 trees.

	ATE			ITE		
	RMSE	coverage	length	RMSE	coverage	length
BCF	0.27	0.81	0.92	0.92	0.90	2.79
BCF-BMA 1	0.44	0.62	1.03	1.29	0.45	1.49
BCF-BMA 2	0.29	0.77	0.90	1.09	0.55	1.36
BART-BMA	0.46	0.61	1.05	1.35	0.42	1.40
BART	0.29	0.80	0.90	1.01	0.82	2.21
BCF-IS	0.34	0.81	1.08	1.15	0.75	2.31
BART-IS	0.29	0.88	1.09	1.12	0.82	2.70
CF	0.45	0.72	1.21	1.25	0.57	1.75

Table 2.8: Hahn et al. (2018) simulations, $\tau(x) = 1+2x_2x_5$, $\mu(x) = 1+g(x_4)+x_1x_3$, $n = 250$, 200 replications.

	ATE			ITE		
	RMSE	coverage	length	RMSE	coverage	length
BCF	0.22	0.92	0.93	1.03	0.88	2.81
BCF-BMA 1	0.29	0.93	1.36	1.35	0.59	2.09
BCF-BMA 2	0.29	0.92	1.30	1.29	0.63	2.00
BART-BMA	0.48	0.61	1.01	1.59	0.41	1.30
BART	0.22	0.91	0.94	0.91	0.89	2.69
BCF-IS	0.27	0.97	1.41	1.26	0.79	2.73
BART-IS	0.22	1.00	1.42	1.17	0.88	3.37
CF	0.66	0.54	1.38	1.32	0.57	2.00

Table 2.9: Hahn et al. (2018) simulations, $\tau(x) = 1 + 2x_2x_5$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 250$, 200 replications.

The results for these simulations suggest that standard BART generally outperforms the other methods in terms of RMSE, followed by BCF, although BART-IS and BCF-IS are competitive with BART and BCF for some DGPs, and generally outperform standard causal forests.[52] In some cases the coverage of credible intervals is better for the new algorithms described in this paper than for BART or BCF, although it should be noted that the 100% or nearly 100% coverage observed, for example in Table 2.9 is not desirable, and the prediction intervals for the new methods are notably wider than those of BART and BCF.

The RMSE of ITE estimates for simulations with heterogeneous treatment effects is worse for BCF-BMA than for BART and BCF. This is expected because the default setting for BCF-BMA are 5 $\mu(x)$ trees, and 5 $\tau(x)$ trees, each of which has a maximum of 5 splits. Therefore the estimates are less heterogeneous than those produced by BART and BCF with many trees. However, the relatively small set of simpler models averaged by BCF-BMA is more interpretable and still performs reasonably well, particularly for ATE estimation.

Data Challenge Datasets

The annual Atlantic Causal Inference Conference (ACIC) has run a data analysis competition for treatment effect estimation methods. BART and BCF have performed well in this competition (Dorie et al. 2019, Hahn et al. 2019).

Table 2.10 presents a comparison between BCF, BCF-IS, BART-IS, BART, and CF applied to the

[52]The performance of BART-IS and BCF-IS improves with the number of samples drawn. There is therefore a trade-off between computational time and accuracy, although this is less of an issue when the draws are parallelized across many threads. The extent to which the results would improve with a greater number of draws is a potential topic for future research.

publicly available data from the 2019 ACIC Data Challenge.[53] The results are restricted to the 1200 datasets in the low-dimensional category with less than 1000 observations and a continuous dependent variable.[54] In all cases the estimates and intervals are produced for $\frac{1}{N}\sum_{i=1}^{N}\tau(x_i)$, and the RMSE and coverage are calculated using the true population ATE.

	ATE		
	RMSE	coverage	length
BCF	0.18	0.88	0.67
BCF-IS	0.17	0.91	0.69
BART-IS	0.19	0.95	0.93
BART	0.23	0.93	0.99
CF	0.22	0.93	1.01

Table 2.10: Results for ACIC Data Challenge low-dimensional datasets with less than 1000 observations and a continuous dependent variable.

BCF-IS attains the lowest RMSE, but the results for BCF and BART-IS are similar. BART-IS achieves the most accurate coverage of prediction intervals.

2.7 Applications

This section includes three applications of the methods introduced in this chapter. First, the usefulness of the methods in treatment effect estimation is demonstrated on an electricity Time-of-Use pricing trial dataset. The second application is a demonstration of how the methods introduced in this chapter can be generically used in direct forecasting of inflation. The third example is an application of variable importance measures for identifying determinants of economic growth.[55]

2.7.1 Time-of-Use Electricity Pricing Trial

This subsection revisits the application introduced in chapter 1 of this thesis. The data is from the Electricity Smart Metering Customer Behavioural Trial conducted by the Irish Commission for Energy Regulation (CER 2011). The dataset consists of half hourly residential electricity demand observations for 4225 households over 536 days. The benchmark period began on 14th July 2009 and ended on 31st December 2009. Households were then randomly allocated to either a control group or various TOU Pricing Schemes and Demand Side Management stimuli from 1st January 2010 to 31st December 2010. See the first chapter of this thesis for further details.

This subsection presents results for the application of ITE estimation methods to a subset of the data containing control households and households allocated to tariff C and the IHD stimulus (1001 households in total). All households were charged a tariff of 14.1 cents per kWh (c/kWh) during the benchmark (pre-treatment) period. The control group paid 14.1 c/kWh for all half-hours during the trial period. The treatment group paid 10 c/kWh from 11pm to 8am, 32 c/kWh at the peak hours of 5pm-7pm on weekdays,

[53] Results are not presented for BCF-BMA or BART-BMA, because the current implementations can require a large quantity of RAM, and this can lead to errors/crashes.

[54] The current implementations of BART-IS and BCF-IS are slow when applied to datasets with many observations. The methods presented in this chapter are designed for data with a continuous dependent variable. See chapter 3 of this thesis for the results for ACIC 2019 datasets with binary outcomes.

[55] This topic has received much attention in the econometric literature on BMA of linear models (Steel 2017, Sala-i Martin et al. 2004, Fernandez et al. 2001*a*, Doppelhofer & Weeks 2009, Eicher et al. 2011).

and 13 c/kWh at all other half-hours including weekends. The outcome variable of interest is average half-hourly peak demand over the whole trial period (in kWh per half-hour). Covariates include pre-treatment consumption information and responses to a survey. See Chapter 1 for a full list of variables.

The methods compared are: Causal forest,[56] BART-MCMC, [57], BART-BMA, BCF-BMA, BART-IS, BCF-IS, and the following three linear models:

A model only including a treatment dummy variable

$$peak_i = \beta_0 + \beta_1 TOU_i + \varepsilon_i \tag{2.1}$$

where $peak_i$ is average trial period half-hourly consumption and TOU_i is a dummy variable equal to one if the household is in the TOU group and zero otherwise.

A model including a pre-trial consumption control variables.[58]

$$peak_i = \beta_0 + \beta_1 TOU_i + \beta_2 pre\_trial\_avg\_peak_i + \beta_3 pre\_trial\_var\_peak_i + \beta_4 pre\_trial\_avg_i + \beta_5 pre\_trial\_avg\_off\_peak_i + \varepsilon_i \tag{2.2}$$

where $pre\_trial\_avg\_peak$ is average half-hourly peak consumption during the pre-treatment period, $pre\_trial\_var\_peak_i$ is the sample variance of half-hourly peak consumption during the pre-treatment period, $pre\_trial\_avg_i$ is average pre-trial consumption across all half-hours, and $pre\_trial\_avg\_off\_peak_i$ is average pre-trial consumption across off-peak daytime hours.

A model including an interaction between treatment and pre-trial average peak consumption

$$peak_i = \beta_0 + \beta_1 TOU_i + \beta_2 pre\_trial\_peak_i + \beta_3 TOU_i * pre\_trial\_peak_i + \varepsilon_i \tag{2.3}$$

A key difficulty in assessing the performance of Individual Treatment Effect estimation methods on real-world datasets is the fact that the ground truth is never known. The true treatment effect for an individual, $Y_i(1) - Y_i(0)$ can never be observed. This is known as the "fundamental problem of causal inference". If the true treatment effect, $\tau(x)$ were observable, then a suitable measure of accuracy of ITE estimation methods would be the MSE, i.e. $\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}(x) - \tau(x))^2$, which is also known as the Precision of Estimating Heterogeneous Effects (PEHE) (Hill 2011).

A number of approaches have been suggested for estimation of the accuracy of ITE estimation methods (Schuler et al. 2018, Saito & Yasui 2019, Alaa & Van Der Schaar 2019). Schuler et al. (2018) review the literature and find that the $\widehat{\tau - risk}_R$ measure proposed by Nie & Wager (2017*a*) most consistently selects the highest performing model. Therefore the $\widehat{\tau - risk}_R$ measure will be used to compare the performance of treatment effect estimation methods on the TOU pricing trial dataset. The $\widehat{\tau - risk}_R$ measure is defined as:

$$\widehat{\tau - risk}_R = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}^{|\mathcal{V}|}((y_i - \breve{m}(x_i)) - (T_i - \breve{p}(x_i))\hat{\tau}(x_i))^2$$

where $\mathcal{V}$ denotes the validation dataset, $\breve{m}(x_i)$ is an estimate of $\mathbb{E}[Y|X]$ obtained by regressing Y on X without using the treatment T,[59] and $\breve{p}(x_i)$ is the estimated propensity score, which in this example is simply set equal to the proportion of households allocated to the treatment group because treatment is randomized.

[56] The causal forest algorithm was implemented using the **R** package **grf**.

[57] Standard BART is implemented using the **R** package **BART**

[58] Note that treatment is randomized and therefore orthogonality between TOU_i and $pre_trial_peak_i$ ensures that this does not bias the treatment effect estimates.

[59] In this example, $\breve{m}(x_i)$ is estimated by gradient-boosted trees using the **xgbTree** option in the **R** package **caret**.

Table 2.11 presents the results from the application of the $\widehat{\tau - risk}_R$ measure to the CER electricity trial data with tenfold cross validation. For each validation fold, the ITE estimation algorithms are trained on the other nine folds and $\widehat{\tau - risk}_R$ measure is calculated using the validation fold. The final result is the average across all ten validation folds. It can be observed from table 2.11 that BCF-IS minimizes τ-risk and other measures do not perform notably better than a linear model without controls. However, these results should be interpreted with caution given the limitations of methods for assessing the accuracy of ITE estimation methods.

Method	τ-risk
CF	0.0021
BART	0.0016
BART-IS	0.0011
BCF-IS	**0.0007**
BART-BMA	0.0021
BCF-BMA	0.0034
LM	0.0013
LM with controls	0.0019
LM with interaction	0.0025

Table 2.11: Tau-risk measure of accuracy of ITE estimates applied to CER electricity trial data with tenfold cross-validation.

Table 2.12 presents sample correlations of ITE estimates. BART-IS did not detect any heterogeneity in treatment effects, and therefore correlations are unavailable for the BART-IS estimates.[60] LM refers to the linear model with interactions, (equation 2.3). The results form the causal forest, BART-BMA, and linear model are highly correlated. Somewhat surprisingly, BCF-IS and BCF-BMA are not highly correlated with the linear model.

	CF	BART	BCF	BART-IS	BCF-IS	BART-BMA	BCF-BMA	LM
CF	1							
BART	0.81	1						
BCF	0.77	0.59	1					
BART-IS	NA	NA	NA	1				
BCF-IS	0.33	0.31	0.19	NA	1			
BART-BMA	0.86	0.66	0.69	NA	0.23	1		
BCF-BMA	0.51	0.42	0.39	NA	0.23	0.37	1	
LM	0.93	0.72	0.75	NA	0.3	0.88	0.36	1

Table 2.12: Correlations of ITE estimates for CER data

Figure 2.4 plots ITE estimates on the y-axis and pre-treatment average peak electricity consumption on the x-axis. The estimated treatment effect function from the linear model with an interaction (equation 2.3) is given by the black line. Note that the linear model does not give the true treatment effect function, although it is expected that the treatment effect increases in magnitude nearly linearly with the level of consumption as the amount of reducible consumption is a key determinant of a household's ability to make energy savings. Tree-based methods may be limited in their ability to capture this smooth association between past consumption and demand response. Nonetheless, figure 2.4 shows that all methods produce estimates that are associated to an extent with past peak consumption. The standard causal forest produces

[60]The fact that homogeneity of treatment effects in this dataset was strongly rejected in chapter 1 of this thesis suggests that BART-IS has some limitations, at least when applied to some datasets.

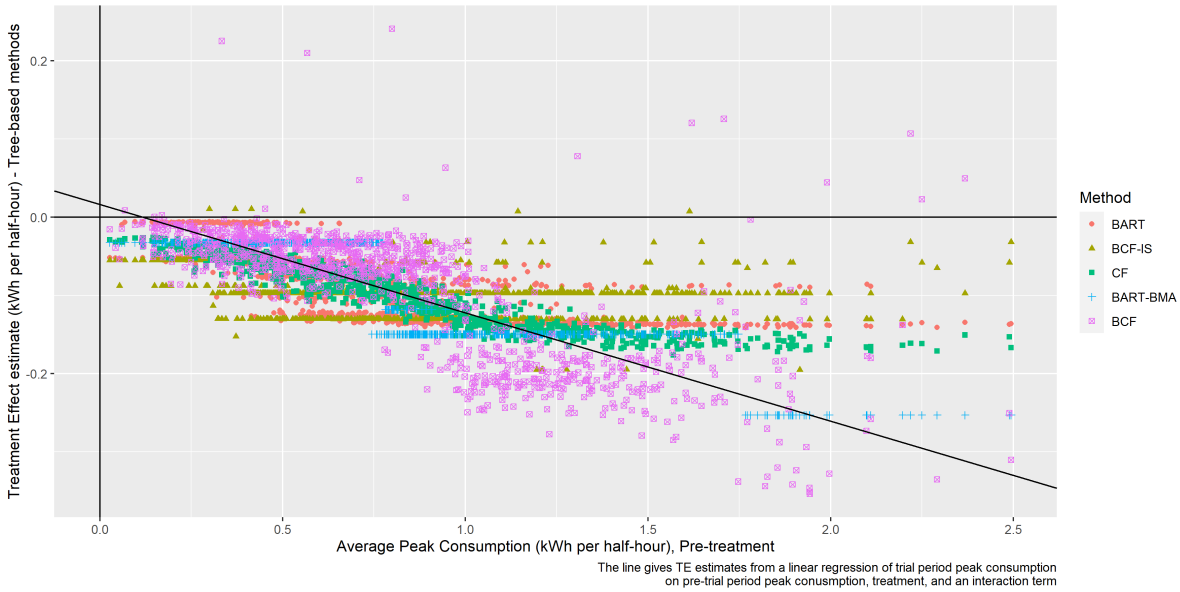


Figure 2.4: ITE estimates (kWh) for CER data (y-axis) against pre-trial average half-hourly peak consumption (kWh) (x-axis).

a particularly impressive near-linear association between the treatment effect and past consumption which breaks down for households with very high past consumption. It is realistic for there to be a limit to the demand response of households with very high levels of pre-treatment consumption because these households are likely to have high-income and be relatively price-inelastic. This provides support for the choice of causal forest in the first chapter of this thesis.

However, it is possible that variables other than past peak consumption are more often selected by algorithms other than the standard causal forest. The causal forest does not appear to capture much heterogeneity beyond that which can be captured by a linear model with an interaction. The high level of heterogeneity, which increases with pre-trial peak consumption for BART-based methods is also arguably to be expected from this data. The lower $\tau - risk$ score for BART and BCF-IS may reflect the ability of these methods to find other drivers of heterogeneity of demand response.

The standard BCF produces some very unrealistic estimates, with some households estimated to increase their peak consumption, and others estimated to decrease peak consumption to an implausibly large extent.[61] This issue is investigated in further detail in Appendix B.7. This suggests that standard BART and the alternative implementations introduced in this chapter are preferable to standard BCF in the application to this dataset.

2.7.2 Inflation Forecasting

This subsection compares BART implementations, Random Forests (RF) and LASSO in a generic application to direct forecasting of inflation data. The dataset, taken from Garcia et al. (2017), consists of monthly inflation data from Brazil from 31 January 2003 to 31 December 2015. There are 58 covariates, which include price indices, electricity consumption, industrial production, unemployment, income, exchange rates, interest

[61] A demand response of -0.2 kWh per half-hour is on the order of 20% of peak consumption.

rates, and government fiscal statistics, and the money supply. See Garcia et al. (2017) for further details on the dataset.

The results reported here are for BART-MCMC, BART-BMA, BART-IS, LASSO, and Random Forests (RF) applied generically in direct forecasting to 1, 3, 6, and 12 step ahead forecasts. This is straightforward to implement using the **R** package **forecastML** (R Core Team 2020, Redell 2020). All graphs, tables and code for this example are adapted from an introductory example in the **forecastML** documentation.[62]

In addition, results are included for a diffusion index model (Stock & Watson 2002) with an autoregressive component and factor lags. The model has the form:

$$y_{t+h}^{h} = \alpha_h + \sum_{j=1}^{3} \boldsymbol{\beta}_{hj}' F_{t-j+1} + \sum_{j=1}^{3} \gamma_{hj} y_{t-j+1} + \epsilon_{t+h}^{h}$$

where F_{t-j+1} is the vector of the first three principal components of the data matrix at time period $t-j+1$, [63] [64] $\boldsymbol{\beta}_{hj}'$ is a vector of three coefficients for the factors in time period $t-j+1$, γ_{hj} is a coefficient of the outcome lag y_{t-j+1}, and ϵ_{t+h}^{h} is the idiosyncratic disturbance. The h-step ahead forecasts are equal to

$$y_{T+h|T} = \hat{\alpha}_h + \sum_{j=1}^{3} \hat{\boldsymbol{\beta}}_{hj}' \hat{F}_{T-j+1} + \sum_{j=1}^{3} \hat{\gamma}_{hj} y_{T-j+1}$$

where $\hat{F}_{T-j+1}$ is the vector of estimated principal components, and $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\gamma}$ are parameter estimates.

The models for all forecast horizons (1, 3, 6, and 12 step ahead forecasts) make use of contemporaneous values and one and two period lagged values of covariates and the dependent variable.[65] Overall, there are 156 months of observations. The final 12 months (January to December 2015) are held out as test data. The inflation time series is graphed in figure 2.5.

Within the training data, the algorithms are first assessed on three validation windows of length 12, 12, and 9 months. The windows are April 2004 - March 2005, April 2009 - March 2010, and April 2014 to December 2014. The validation windows are only used for LASSO parameter tuning because a key appeal of BART and RF is that the algorithms tend to perform well without parameter tuning, and therefore the goal is to test the performance of these algorithms without tuning. For the same reason and fair comparison across models, the decision was made not to perform a search for the optimal choice of lags to include in the diffusion index model. The accuracy of the algorithms on the validation data is summarized in table 2.13. The diffusion index model (DI-AR-Lag) performs slightly worse than the other models in the validation data, and the other models have similar accuracy.

Finally, all algorithms are retrained using all the data up to December 2014, and accuracy is assessed on the holdout data. Figure 2.6 plots the hold-out predictions and actual observations. Table 2.14 presents

[62] The original example is available at https://cran.r-project.org/web/packages/forecastML/vignettes/package_overview.html (Redell 2020). I replaced the dataset with the inflation data from Garcia et al. (2017) and added BART-MCMC, BART-BMA, and BART-IS as methods for direct forecasting.

[63] The data matrix used for estimating principal components does not include lags of the covariates (except in so far as the original dataset includes lags). The lags of the principal components are constructed using the lags of the covariates. This ensures that the same number of lags are used by all methods included in this comparison. An alternative would be to include more lags in the initial matrix used for construction of principal components, although this would require the lags of principal components to be constructed from higher order lags not used by te other methods. Stock & Watson (2002) observe that forecasts based on larger "stacked" data generally perform worse than forecasts based on "unstacked" data.

[64] The choice of three principal components is entirely arbitrary. A search across different model specifications is not implemented in this paper to ensurer a fairer comparison across methods. The results of Stock & Watson (2002) suggest that "most of the forecast gains seem to come from using a single factor".

[65] In the case of tree-based methods the contemporaneous and lagged features and outcome are are included as potential splitting variables.

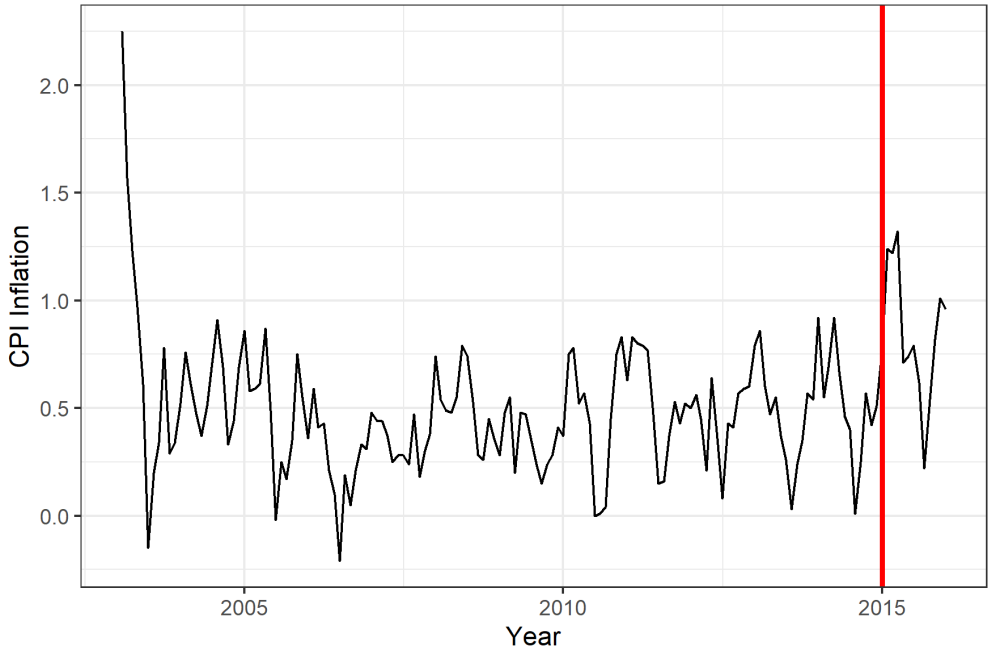
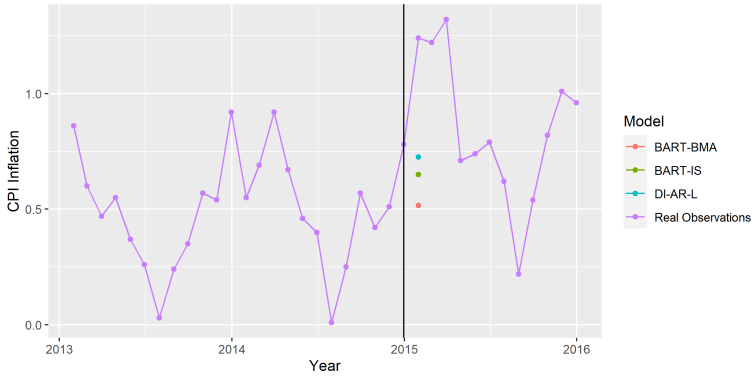


Figure 2.5: CPI inflation time series. The data to the right of the red vertical line is held out as test data.

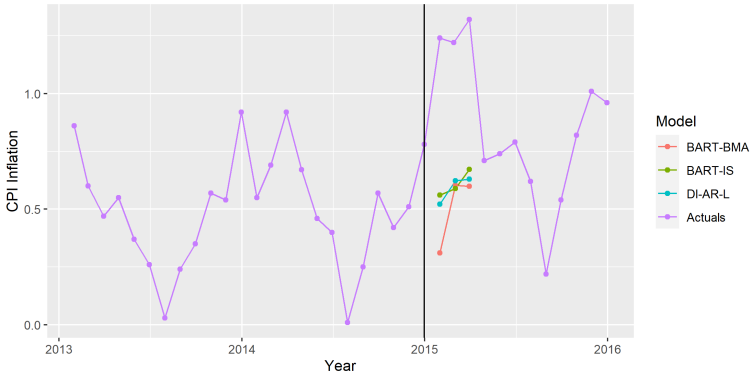
Model	MAE	MAPE	MDAPE	SMAPE	RMSE
BART-BMA	0.19	65.64	35.20	51.41	0.22
BART-IS	0.17	45.34	29.33	43.09	0.21
BART-MCMC	0.16	38.93	26.94	40.28	0.21
LASSO	0.17	43.11	28.41	40.98	0.21
RF	0.16	42.05	26.57	39.72	0.21
DI-AR-Lag	0.20	57.17	31.79	49.59	0.24

Table 2.13: Measures of accuracy of inflation forecasts, averaged across all validation windows and forecast horizons (1,3,6,12 step forecasts). MAE = Mean Absolute Error, MAPE = Mean Absolute Percentage Error, MDAPE= Median Absolute Percentage Error, SMAPE = Symmetric Mean Absolute Percentage Error, RMSE = Root Mean Squared Error

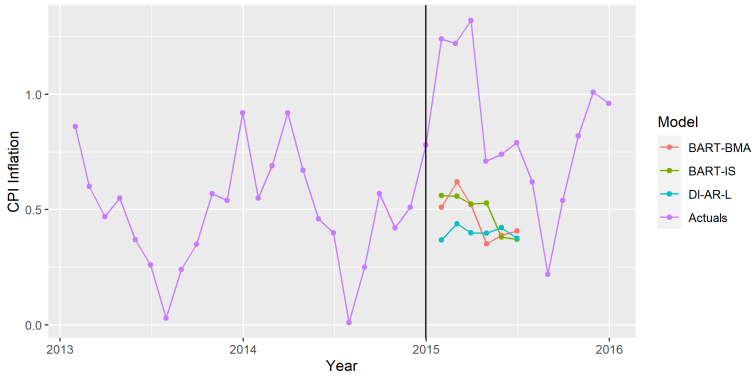
measures of accuracy for the predictions on the holdout data. LASSO performs best in terms of RMSE, although there is no clear winner across all measures of forecast accuracy. A comparison of the ranking of methods between Table 2.13 and Table 2.14 suggests that the relative performance of methods may be sensitive to the size of the available training dataset. While BART-IS does not outperform LASSO or a diffusion index model, it is encouraging to note that it outperforms BART-MCMC across all measures of forecast accuracy in this example. These results suggest that that BART with a naive direct forecasting approach is not particularly well suited to inflation forecasting. However, recently introduced methods such as Bayesian Additive Vector Autoregression Trees (BAVART) (Huber & Rossini 2020) might yield better results.



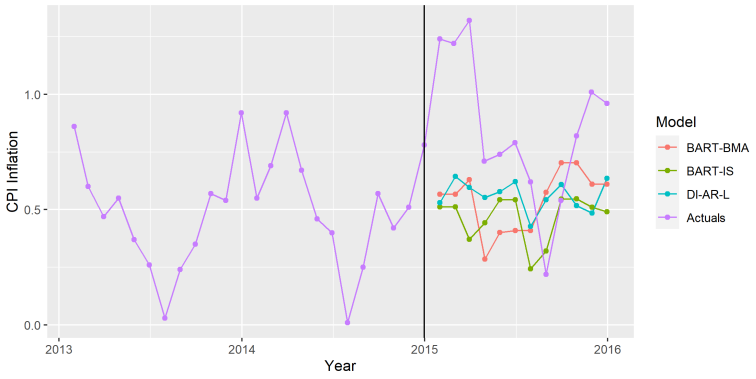
(a) Inflation forecasts, 1 month ahead



(b) Inflation forecasts, 3 months ahead



(c) Inflation forecasts, 6 months ahead



(d) Inflation forecasts, 12 months ahead

Figure 2.6: Hold-out data predictions and actual observations for one-month, three-month, six-month, and 12-month ahead forecasts of CPI inflations.

Model	MAPE	MDAPE	SMAPE	RMSE
BART-BMA	58.58	79.91	0.69	2.57
BART-IS	50.43	67.64	0.61	2.29
BART-MCMC	54.47	68.73	0.63	2.36
LASSO	51.01	65.43	0.59	2.21
RF	51.83	66.20	0.60	2.25
DI-AR-L	48.33	65.80	0.61	2.27

Table 2.14: Measures of accuracy of inflation forecasts in hold-out data, averaged across all forecast horizons (1,3,6,12 step forecasts). MAPE = Mean Absolute Percentage Error, MDAPE= Median Absolute Percentage Error, SMAPE = Symmetric Mean Absolute Percentage Error, RMSE = Root Mean Squared Error

2.7.3 Growth Determinants

There is an extensive literature on the application of Bayesian Model Averaging to macroeconomic datasets for the discovery of determinants of economic growth. Early empirical studies on growth determinants include those by Levine & Renelt (1992), Barro (1996*b*,*a*) and Sala-i Martin (1997). Examples of studies applying Bayesian Model Averaging of linear models include those by Fernandez et al. (2001*b*,*a*), Sala-i Martin et al. (2004), Doppelhofer & Weeks (2009), and many subsequent papers. See Steel (2017) for a comprehensive review of the literature.

There are a small number of growth determinant studies that move beyond standard BMA of linear models. Dobra et al. (2010) apply Gaussian Graphical Models to account for dependency between variables, Durlauf et al. (2012), Lenkoski et al. (2014) and Karl & Lenkoski (2012) account for endogeneity of potential growth determinants, and Doppelhofer et al. (2016) account for measurement error. Moral-Benito (2016) and Leon-Gonzalez & Montolio (2015) accounts for endogeneity in BMA of panel models of economic growth.

Few, if any, existing papers in the growth determinant literature allow for complex nonlinearities and interactions. We consider the usefulness of BART in selecting determinants of economic growth. The following illustrative example does not include any attempt to take account of endogeneity.[66] The dataset is from a paper by Sala-i Martin et al. (2004) on Bayesian Averaging of Classical Estimators (BACE), which involves an approximation to BMA of linear models. All countries with missing observations for any covariates are removed from the dataset. The data contains 67 covariates for 88 countries. The dependent variable is the average growth rate of GDP from 1960 to 1996. The variable names with descriptions are given in table 2.15.

[66] An interesting topic for future research would be how to obtain variable selection measures from BART models that account for endogeneity.

Description of Variable	Variable Name	Description of Variable	Variable Name
Absolute Latitude	ABSLATIT	Frac. of Land Near Navigable Water	LT100CR
Air Distance to Big Cities	AIRDIST	Malaria Prevalence in 1960s	MALFAL66
Ethnolinguistic Fractionalization	AVELF	Fraction GDP in Mining	MINING
British Colony Dummy	BRIT	Fraction Muslim	MUSLIM00
Fraction Buddhist	BUDDHA	Timing of Independence	NEWSTATE
Fraction Catholic	CATH00	Oil Producing Country Dummy	OIL
Civil Liberties	CIV72	Openness measure 1965-74	OPENDEC1
Colony Dummy	COLONY	Fraction Orthodox	ORTH00
Fraction Confucian	CONFUC	Fraction Speaking Foreign Language	OTHFRAC
Population Density 1960	DENS60	Primary Schooling in 1960	P60
Population Density Coastal in 1960s	DENS65C	Average Inflation 1960-90	PI6090
Interior Density	DENS65I	Square of Inflation 1960-90	SQPI6090
Population Growth Rate 1960-90	DPOP6090	Political Rights	PRIGHTS
East Asian Dummy	EAST	Fraction Population Less than 15	POP1560
Capitalism	ECORG	Population in 1960	POP60
English Speaking Population	ENGFRAC	Fraction Population Over 65	POP6560
European Dummy	EUROPE	Primary Exports 1970	PRIEXP70
Fertility in 1960s	FERTLDC1	Fraction Protestants	PROT00
Defence Spending Share	GDE1	Real Exchange Rate Distortions	RERD
GDP in 1960 (log)	GDPCH60L	Revolutions and Coups	REVCOUP
Public Educ. Spending Share GDP, 1960s	GEEREC1	African Dummy	SAFRICA
Public Investment Share	GGCFD3	Outward Orientation	SCOUT
Nominal Govt. GDP Share 1960s	GOVNOM1	Size of Economy	SIZE60
Government Share of GDP in 1960s	GOVSH61	Socialist Dummy	SOCIALIST
Gov. Consumption Share 1960s	GVR61	Spanish Colony	SPAIN
Higher Education 1960	H60	Terms of Trade Growth in 1960s	TOT1DEC1
Religion Measure	HERF00	Terms of Trade Ranking	TOTIND
Fraction Hindus	HINDU00	Fraction of Tropical Area	TROPICAR
Investment Price	IPRICE1	Fraction Population In Tropics	TROPPOP
Latin American Dummy	LAAM	Fraction Spent in War 1960-90	WARTIME
Land Area	LANDAREA	War Participation 1960-90	WARTORN
Landlocked Country Dummy	LANDLOCK	Years Open 1950-94	YRSOPEN
Hydrocarbon Deposits in 1993	LHCPC	Tropical Climate Zone	ZTROPICS
Life Expectancy in 1960	LIFE060		

Table 2.15: Names of variables in growth determinant regression

The standard measure of the importance of growth determinants in the existing literature on BMA of growth regressions is the Posterior Inclusion Probability (PIP). The PIP is the model-probability weighted average of an dummy variable equal to one if the variable of interest is included in the model. Let the dummy variable γ_i equal 1 if variable x_i is included in a model. Let the model space be denoted by $\mathcal{M}$, and let j index the set of models. Then the PIP can be written as

$$PIP_i = p(i|\boldsymbol{y}) = \sum_{\forall j \in \mathcal{M}} \mathbf{1}(\gamma_i = 1|\boldsymbol{y}, M_j) p(M_j|\boldsymbol{y})$$

Posterior Inclusion Probabilities can be calculated for BART-MCMC,[67] BART-IS, and BART-BMA by setting $\gamma_i = 1$ if any splitting rules in any trees in the sum-of-tree model are based on variable x_i. Alternatively,

[67] In this example, BART-MCMC is implemented with 5 trees per model for comparability of PIPs and variable importance. Chipman et al. (2010) recommend a small number of trees (5, 10, or 20) for obtaining variable importance measures because this results in more parsimonious models that make use of fewer splitting variables.

variable importance can be assessed using a model weighted average of the fraction of splitting rules based on the variable of interest. This alternative measure is simply referred to as "variable importance" in this paper.

Table 2.16 gives the posterior inclusion probability results for BART based methods and the original BACE results obtained by Sala-i Martin et al. (2004). A number of key variables receive the highest PIP and variable importance across all three BART implementations. Therefore, this is an example of an economic application for which the methods introduced in this paper are viable alternatives to BART-MCMC for identifying important variables. Some key variables have a relatively high PIP across all methods, such as the East Asia dummy variable and fraction Confucian. However, there are some notable differences in PIPs across methods. For example, BART-IS and BART-MCMC do not place a high PIP on GDP in 1960 or enrolment in primary education in 1960, while BACE and BART-BMA place high PIP on these variables. One possible explanation for this result is that some pairs of variables are substitutes (i.e. both explain the same underlying effect) while others are complements that have a higher probability of both being included or excluded in the model. Therefore a thorough analysis of correlations and jointness (Doppelhofer & Weeks 2009) may explain some of these patterns. However, it is also possible that BART-based methods can find interactions and non-linearities that are not captured by the linear model approach. Table 2.17 presents the variable importance results for BART-BMA, BART-IS, and BART-MCMC. The pattern for variable importances is similar to the pattern observed for PIPs.

Variables	BART BMA	BART IS	BART MCMC	BACE	Variables	BART BMA	BART IS	BART MCMC	BACE
ABSLATIT	**1**	0.231	0.045	0.033	LT100CR	0	0.060	0.079	0.019
AIRDIST	0	0.063	0.007	0.039	MALFAL66	**1**	0.227	**0.780**	0.252
AVELF	0	0.103	0.102	0.105	MINING	0	0.057	0.007	0.124
BRIT	0	0.033	0.045	0.027	MUSLIM00	0	0.075	0.066	0.114
BUDDHA	**1**	**0.556**	**0.396**	0.108	NEWSTATE	0	0.063	0.020	0.019
CATH00	0	0.149	0.010	0.033	OIL	0	0.066	0.008	0.019
CIV72	0	0.023	0.010	0.029	OPENDEC1	0	0.054	0.016	0.076
COLONY	0	0.064	0.011	0.029	ORTH00	0	0.042	0.003	0.015
CONFUC	**1**	**0.339**	**0.318**	0.206	OTHFRAC	0	0.148	0.078	0.080
DENS60	0	0.123	0.071	0.086	P60	**1**	0.177	0.151	**0.796**
DENS65C	0	0.161	0.083	0.428	PI6090	0	0.049	0.103	0.020
DENS65I	0	0.038	0.016	0.015	SQPI6090	0	0.039	0.002	0.018
DPOP6090	0	0.055	0.021	0.019	PRIGHTS	**1**	0.062	0.003	0.066
EAST	**1**	**0.757**	**1.000**	**0.823**	POP1560	0	0.049	0.027	0.041
ECORG	0	0.078	0.110	0.015	POP60	0	0.046	0.019	0.021
ENGFRAC	0	0.044	0.011	0.020	POP6560	0	0.100	0.064	0.022
EUROPE	0	0.069	0.132	0.030	PRIEXP70	0	0.061	0.043	0.053
FERTLDC1	0	0.071	0.029	0.031	PROT00	0	0.046	0.020	0.046
GDE1	0	0.083	0.055	0.021	RERD	**1**	0.082	0.078	0.082
GDPCH60L	0.584	0.034	0.020	**0.685**	REVCOUP	0	0.042	0.010	0.029
GEEREC1	**1**	0.039	0.304	0.021	SAFRICA	0	0.103	0.137	0.154
GGCFD3	0	0.031	0.010	0.048	SCOUT	0	0.061	0.025	0.030
GOVNOM1	**1**	0.026	0.018	0.063	SIZE60	0	0.065	0.023	0.020
GOVSH61	0	0.045	0.031	0.036	SOCIALIST	0	0.155	0.005	0.020
GVR61	0	0.055	0.078	0.104	SPAIN	0	0.064	0.036	0.123
H60	**1**	**0.242**	0.023	0.061	TOT1DEC1	0	0.046	0.048	0.021
HERF00	0	0.072	0.084	0.020	TOTIND	0	0.099	0.021	0.016
HINDU00	0	0.034	0.010	0.045	TROPICAR	0	0.232	**0.454**	**0.563**
IPRICE1	**1**	0.168	0.230	**0.774**	TROPPOP	**1**	0.143	0.020	0.058
LAAM	0	0.056	0.044	0.149	WARTIME	0	0.06	0.045	0.016
LANDAREA	0	0.065	0.021	0.016	WARTORN	0	0.069	0.013	0.015
LANDLOCK	0	0.039	0.017	0.021	YRSOPEN	**1**	0.227	0.090	0.119
LHCPC	0	0.124	0.018	0.025	ZTROPICS	0	0.151	0.066	0.016
LIFE060	0	**0.553**	0.136	0.209					

Table 2.16: PIPs for growth determinant regressions. BART based methods and original BACE results from Sala-i Martin et al. (2004).

Variables	BART BMA	BART IS	BART MCMC	Variables	BART BMA	BART IS	BART MCMC
ABSLATIT	**0.095**	0.024	0.010	LT100CR	0	0.009	0.014
AIRDIST	0	0.008	0.001	MALFAL66	**0.095**	0.030	**0.136**
AVELF	0	0.014	0.017	MINING	0	0.005	0.001
BRIT	0	0.005	0.006	MUSLIM00	0	0.007	0.011
BUDDHA	0.076	**0.074**	**0.068**	NEWSTATE	0	0.009	0.003
CATH00	0	0.022	0.002	OIL	0	0.008	0.001
CIV72	0	0.003	0.001	OPENDEC1	0	0.006	0.002
COLONY	0	0.008	0.002	ORTH00	0	0.004	0.001
CONFUC	**0.114**	**0.039**	**0.055**	OTHFRAC	0	0.019	0.012
DENS60	0	0.021	0.011	P60	**0.095**	0.025	0.023
DENS65C	0	0.03	0.013	PI6090	0	0.006	0.017
DENS65I	0	0.005	0.003	SQPI6090	0	0.005	0
DPOP6090	0	0.008	0.003	PRIGHTS	0.057	0.009	0
EAST	**0.095**	**0.107**	**0.174**	POP1560	0	0.008	0.004
ECORG	0	0.010	0.002	POP60	0	0.004	0.003
ENGFRAC	0	0.005	0.002	POP6560	0	0.014	0.010
EUROPE	0	0.010	0.022	PRIEXP70	0	0.009	0.007
FERTLDC1	0	0.010	0.005	PROT00	0	0.005	0.003
GDE1	0	0.009	0.009	RERD	0.076	0.010	0.013
GDPCH60L	0.011	0.004	0.003	REVCOUP	0	0.006	0.001
GEEREC1	0.038	0.004	0.049	SAFRICA	0	0.012	0.024
GGCFD3	0	0.004	0.002	SCOUT	0	0.009	0.004
GOVNOM1	0.019	0.003	0.003	SIZE60	0	0.011	0.004
GOVSH61	0	0.006	0.005	SOCIALIST	0	0.028	0.001
GVR61	0	0.008	0.013	SPAIN	0	0.008	0.006
H60	0.019	0.020	0.004	TOT1DEC1	0	0.007	0.007
HERF00	0	0.012	0.013	TOTIND	0	0.010	0.003
HINDU00	0	0.004	0.002	TROPICAR	0	**0.031**	**0.083**
IPRICE1	0.057	0.022	0.036	TROPPOP	0.057	0.020	0.003
LAAM	0	0.006	0.007	WARTIME	0	0.006	0.007
LANDAREA	0	0.009	0.004	WARTORN	0	0.010	0.002
LANDLOCK	0	0.006	0.003	YRSOPEN	**0.095**	0.018	0.014
LHCPC	0	0.019	0.002	ZTROPICS	0	0.017	0.010
LIFE060	0	**0.078**	0.022				

Table 2.17: Variable Importances for growth determinant BART results.

2.8 Conclusion

2.8.1 Limitations of Importance Sampling of Models

Importance sampling is well known to have limitations in high-dimensional settings (Agapiou et al. 2017). In particular, without an appropriate choice of sampler, the IS approximation can have high or infinite variance. When there are many covariates, Bayesian Model Averaging of linear models, or of tree-based models involves sampling from a high-dimensional model space. Therefore, simple importance sampling-based approaches to model averaging suffer from the curse of dimensionality. In the context of linear models, methods for addressing this issue include orthogonalization of the data matrix combined with sampling from approximate model inclusion probabilities, sampling without replacement, and adaptive sampling (Clyde et al. 1996, 2011,

Yu et al. 2010).[68]

Quadrianto & Ghahramani (2014) noted that simple importance sampling from the prior is known not to work so well, and explain that this choice of sampling scheme is due to a trade-off between predictive accuracy and computational time.[69] Lakshminarayanan et al. (2013) similarly justify the use of a prior as a proposal tree sampler within a MCMC algorithm.

Importance sampling schemes for BART are likely to suffer from the curse of dimensionality, and may fail to sample models with high posterior probability. This leads to the following topics for future research. 1. The combination of BART-IS with screening methods or adaptive sampling schemes to improve the variable-selection properties of the algorithm, or 2. Accepting the fact that BART-IS should fail to give an accurate representation of a posterior probability weighted average (or at least a highly variable approximation), is there a potential explanation for the observation that BART-IS can exhibit comparable performance to BART-MCMC on some datasets?

A few studies have combined BART with screening methods. For example, BART-BMA Hernández et al. (2018) relies on a changepoint detection algorithm to reduce the number of potential splitting points to be used in constructing trees. Another approach, RS-BART, combines random subspace methods with BART by applying BART a number of times to subsamples of the set of covariates, using a data-informed sampler for the covariates similar to Sure Independence Screening (Wang et al. 2019, Fan & Lv 2008). It is possible to implement RS-BART with BART-IS instead of BART-MCMC, or to make use of the same data-informed covariate sampler within the standard BART-IS algorithm.[70]

An area for future research is the combination of BART with adaptive sampling methods. Some initial test results suggest that BART-IS in combination with a straightforward update of sampling probabilities based on posterior inclusion probabilities from already sampled models can lead to improved accuracy in moderately high dimensional datasets.[71] This approach moves towards a data-informed stochastic search for sum-of-tree models, analogous to the existing literature for linear models.

There are some similarities between adaptive forms of BART-IS and Thompson Variable Selection (TVS) (Liu & Rockova 2020). TVS makes use of a multi-armed bandits approach that involves iteratively sampling from a distribution of "rewards" that are used to create variable subsamples, applying BART-MCMC to the subsample of the covariates, and using counts of covariate splits to update the reward distribution. This approach exhibits impressive variable selection properties. Such stochastic variable selection approaches can be used to find the Median Probability Model (MPM) rather than a true Bayesian model average, but may yield impressive predictive performance nonetheless. It is in principle possible to replace BART-MCMC with BART-IS in the TVS algorithm.

Liu et al. (2018) introduce Approximate Bayesian Computation Bayesian Forests (ABC-BF). The "naive" implementation of ABC-BF is similar to BART-IS in that it involves independent sampling of tree models from a prior. However, the main difference is that, instead of applying marginal likelihood weights (as in BART-IS), ABC-BF simulates data from the drawn models and accepts or rejects the drawn models based on the distance between the simulated and observed data. ABC-BF involves a spike-and-tree prior that first samples a subset of covariates, and then the model draw is conditional on these covariates. Furthermore,

[68] However, similar limitations can also apply to MCMC based approaches to model averaging, and this provides some motivation for development of stochastic search algorithms (Heaton & Scott 2010, Clyde & Ghosh 2012).

[69] Further improvements to the BART-IS code are required before a thorough comparison of computational speed against BART-MCMC.

[70] Preliminary results suggest that this gives some improvement in predictive accuracy in datasets with low to moderately high covariate dimension.

[71] This is essentially Bayesian Adaptive Sampling (Clyde et al. 2011), albeit sampling models with replacement.

the naive approach is improved on by taking model draws conditional on a subsample of the data (i.e. one BART-MCMC draw), separate to the data for the acceptance rule (conditional on the drawn covariates). This approach is shown by Liu et al. (2018) to have desirable variable selection properties. The resulting robust and stable Posterior Inclusion Probabilities can be used to select the Median Probability Model.[72]

Friedman et al. (2003, 2008) describe how a wide range of methods, including boosting, bagging, random forests, and BMA fall into the framework of importance sampling of the parameters of weak learners.[73] The location and scale of the parameters are both important to the success of the algorithm. On the one hand it is desirable to average over learners with parameters that give minimal predictive risk. On the other hand, if there is insufficient variation in the parameters, each sample provides little additional information. Partial importance sampling locates the parameter sampling distribution near the optimal values (e.g. a single regression tree deterministically fitted to all the data). An ensemble of all strong or all weak base learners will perform poorly. Ideally, base learners should be moderately strong and not very highly correlated. For example, random forest increases the scale and decorrelates the learners by subsampling the training data and randomly sampling the potential splitting variables.

The data-independent sum-of-tree model samples in BART-IS potentially have excessively high scale. Furthermore, unlike MCMC-based algorithms, BART-IS does not sample from the posterior and therefore the samples might not be close to the posterior mode (highest probability model). However, BMA will place all the mass on one model as the number of observations tends to infinity, and this might not be desirable if the "true model" is not representable as a sum of trees. This can be partly addressed in BART-IS by raising the marginal likelihood to a power (Quadrianto & Ghahramani 2014, Grünwald 2012).[74]

Given the above limitations, it is perhaps worthwhile attempting to explain why BART-IS can produce reasonably accurate predictions. One possibility is that, since BART-IS is very similar to other purely random forest methods, it might share some of the desirable properties of these methods (Arlot & Genuer 2014). Methods such as Perfect Ensemble Random Trees and Extremely Randomized Trees (Cutler & Zhao 2001, Geurts et al. 2006) apply equal weights to a set of weak learners, and can exhibit impressive performance.

2.8.2 Summary and Discussion of Future Research

Many MCMC implementations of BART have been demonstrated to be effective in a variety of applications. This paper explores potential alternatives to MCMC based BART. The BCF-BMA algorithm extends an improved version of BART-BMA (Hernández et al. 2018), to treatment effect estimation. This paper also describes BART-IS and BCF-IS, which, in notable contrast with BART-BMA and BCF-BMA, do not involve a deterministic data driven model search, but instead involve simple importance sampling from a data independent model prior.[75]

The BART-IS and BCF-IS sampling schemes are unlikely to be as effective as MCMC methods, despite the marginalization of terminal node parameters. However, the simple importance sampling framework allows for straightforward implementation of BART and testing of different priors and other variations on the model.[76]

[72] Liu et al. (2018) also describe an ABC Forest Fit algorithm that involves, for each random sample of variables and data, sampling predictions from an average over MCMC model draws to be used in the accept/reject step. In principle, the MCMC algorithm could be replaced by BART-IS (applied to the sub-sampled data) in ABC Forest Fit.

[73] Parameters in this context include, for example, splitting variables and splitting points in trees.

[74] Even in finite samples, it may be desirable to reduce the weight applied to correlated models with high marginal likelihoods.

[75] While the BMA implementations in this paper are entirely deterministic, the IS implementations are potentially strongly influenced by data-independent random sampling of models.

[76] See chapter 3 of this thesis for examples. e.g. extensions to binary outcomes.

Interesting potential topics for further research include faster implementations of BART-IS, multivariate BART-IS, [77] semi-parametric BART-IS, [78] and Bayesian stacking of sum-of-tree models. Another area for future research is Bayesian Adaptive Sampling (BAS) of BART Models (Clyde et al. 2011). BAS involves sampling without replacement and possibly adjusting sampling probabilities by predicting the marginal likelihood of unsampled models. While BAS has been applied to sampling of linear models, further research is required for application of this approach to tree-based models.[79]. Furthermore, the potential implementation of a safe-Bayesian approach, as suggested by Quadrianto & Ghahramani (2014) has not been fully explored in this paper. In most examples, BART-IS places a very high posterior probability on a few models. A safe-Bayesian approach can ensure that probability mass is not placed on one model as the number of observations tends to infinity (Grünwald 2012).

[77] See appendix B.3.

[78] See Zeldow et al. (2019) for a description of semi-parametric BART (BART plus a linear model) and an MCMC implementation. See appendix B.4.

[79] This hypothetical alternative approach to BART is distinct from the existing literature that applies BART-MCMC to guide adaptive sampling of *linear* models (Yu et al. 2010, 2012, Yu & Li 2020)

Chapter 3

Generalizations of BART-BMA and BART-IS

Abstract

This chapter outlines extensions of the BART-BMA and BART-IS algorithms to more general settings, including binary outcomes, treatment effects for binary outcomes, censored outcomes, categorical outcomes, and count data. BART-IS and BART-BMA are readily extendable to model frameworks for which the marginal likelihood and posterior can be efficiently calculated or approximated. The examples discussed in this chapter make use of standard Quasi-Newton methods in combination with Laplace approximations.

As examples of how to apply the general approach, Logit-BART-BMA and Logit-BART-IS are described and shown to be competitive with existing tree-based methods on real-world binary classification datasets. In addition, Logit-BCF-IS (and Logit-BCF-BMA) give treatment effect estimates and intervals with accuracy comparable to the best performing methods on simulated datasets from a data analysis challenge. As a further example, Tobit-BART is introduced and implemented using the general BART-IS framework.

3.1 Introduction

3.1.1 BART for Generalized Linear Models

This chapter outlines how BART-BMA and BART-IS can be generalized to a variety of data settings, including binary outcomes, censored outcomes, count data, and multinomial response data. The general approach is applicable in settings in which a linear combination of variables can be replaced by a sum-of-tree model. As explained in chapter 2 of this thesis, a sum-of-trees is itself a representation of a linear combination of indicator variables for terminal nodes with coefficients equal to the terminal node mean parameters. This approach allows for non-linearity and complex interactions between variables, while also accounting for model uncertainty.

Recent advances in Bayesian methods have allowed for a large class of models to be approximated efficiently. The methods introduced in this paper are averages over generalized linear models with the linear combinations of covariates replaced by sums-of-trees. The approach may be applicable to a wider class of models, but this chapter will restrict attention to generalized linear models.

While the general algorithms are not restricted to a particular approximation method, a key candidate method that will be focused on in this chapter is Laplace approximation. Rue et al. (2009) introduce Integrated Nested Laplace Approximations (INLA), which are applicable to latent Gaussian models. Most structured Bayesian models take the form of latent Gaussian models, which are a special case of structured additive regression models.

In structured additive regression models, the outcome y_i is assumed to belong to an exponential family, where the mean μ_i is linked to a structured additive predictor η_i through a link-function $g(.)$, so that

$g(\mu_i) = \eta_i$. The structured additive predictor η_i takes the form:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \tag{3.1}$$

where the $\{f^{(j)}(.)\}$'s are unknown functions of the covariates u, the $\{\beta_k\}$'s represent the linear effects of covariates z and the ε_i's are unstructured terms. Latent Gaussian models apply a Gaussian prior to $\{f^{(j)}(.)\}$, $\{\beta_k\}$, and ε_i.

This paper will focus on averages of generalized linear models, which are of the form $\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki}$. The models being averaged over all use the same link function, and the linear combination of covariates $\sum_{k=1}^{n_\beta} \beta_k z_{ki}$ (or part of the linear combination) is replaced by a linear combination of indicator variables for inclusion in terminal nodes of sums-of-trees (as described in chapter 2).[1] Models that do not involve Gaussian priors may also be included in the general framework outlined in this paper, provided an efficient approximation is available.

The discussion above outlines the general applicability of the approach through the use of INLA (Rue et al. 2009). However, for simplicity of demonstration, this paper will focus on examples in which standard Laplace approximations are feasible, and therefore further description of INLA is omitted. Furthermore, the feasibility of generalization to a particular model depends on the computational speed of the approximation, and performance will depend on the accuracy of the approximation and the appropriateness of the link function.[2] The limited existing literature on the combination of INLA and Bayesian Model Averaging involves averaging over parameters in spatial econometric models (Gómez-Rubio et al. 2020, Gómez-Rubio & Rue 2018, Bivand et al. 2014, 2015). However, this literature does not discuss averaging over different sets of (non-linear functions of) covariates.

The General BART-IS algorithm involves random, data-independent draws of sum-of-tree models, and a marginal likelihood weighted average of these models. The General BART-BMA algorithm involves a model search algorithm that beings by constructing single tree models, and then appends trees to these models, and averages over the set of searched models that have highest posterior probability. The BART-BMA approach requires construction of residuals representing the unexplained part of η_i, which are used to construct trees to be appended to the models. However, the calculation of residuals, while straightforward in the case of Logit, is not always possible. In this sense BART-IS is more generalizable than BART-BMA, as BART-BMA requires model-specific adjustments to the model search algorithm.

A key requirement for this approach to be feasible is that the marginal likelihood can be efficiently calculated and the posterior distribution has a closed form or has a very efficient sampler. This requirement is satisfied in the case of Bayesian logistic regression with a standard Laplace approximation, which is used in this chapter as an illustrative example. Logit-BART-BMA and Logit-BART-IS involve averaging over models in which the binary outcome has success probability equal to the logistic function of a sum-of-tree function.

[1] Models that include $f(.)$ terms such as random effects models $f(u_i) = f_i$, dynamic models $f(u_t) = f_t$, and spatial models $f(u_s) = f_s$, may also be included in the general approach described in this chapter, but these models are not the focus of this chapter.

[2] It could be argued that the link function imposes a strong assumption, and therefore a moment-condition based approach such as Generalized Random Forests (Athey et al. 2019) or Orthogonal Random Forests (Oprescu et al. 2018) is more appropriate in some contexts.

3.1.2 Binary Classification Example and Literature Review

Single tree methods can readily be applied to binary outcome data. Trees in a random forest applied to binary outcomes produce predictions between zero and one because leaf estimates are averages of binary variables. However, sum-of-tree based methods such as BART are less directly applicable to binary outcome data because sums-of-trees can produce predictions outside the range $[0, 1]$ and ideally the statistical framework of BART should account for the fact that the outcomes are binary. Therefore BART-based models for binary outcomes (and other generalized linear models for different forms of outcome variable), rely on a choice of link function.

Sum-of-tree models, such as AdaBoost with decision trees as weak learners, often produce excellent results when applied to binary classification problems (Freund & Schapire 1995, Freund et al. 1996). An early example of a sum-of-tree model for binary outcomes placed in a statistical framework is the LogitBoost algorithm (Friedman et al. 2000). BART can be extended to binary outcome prediction by applying a probit or logit link function to a sum-of-tree model. Chipman et al. (2010) implement Probit-BART-MCMC using the data augmentation Markov Chain Monte Carlo approach of Albert & Chib (1993). Zhang & Härdle (2010) independently applied Probit-BART to credit risk modelling and found that it is competitive with other machine learning methods. Abu-Nimeh et al. (2008) also applied this approach to spam email detection. The **R** package **BART** implements Logit-BART using a computationally intensive MCMC algorithm based on the approach of Gramacy et al. (2012).

The performance of MCMC implementations of BART has been noted to be less impressive for binary outcomes than for continuous outcomes (Hill et al. 2020, Carnegie et al. 2015).[3] The algorithms in this paper provide alternatives to the MCMC implementations of BART for binary outcomes.

A number of recent papers have extended the applicability of BART. Examples include BART variations of multinomial Probit (Kindo, Wang & Peña 2016), quantile regression (Kindo, Wang, Hanson & Peña 2016), survival analysis (Sparapani et al. 2016), recurrent event analysis (Sparapani et al. 2018), and competing risks models (Sparapani et al. 2019). See Hill et al. (2020), Tan & Roy (2019) and Linero (2017) for review articles.

Murray (2017) proposes new priors and a data augmentation scheme that allow for an efficient MCMC sampler for BART-based methods outside the context of Gaussian models. The approach of Murray (2017) (Log-linear BART) is to model the log of the regression function as a sum-of-trees and apply a generalized inverse Gaussian prior distribution to the terminal node parameters. Log-linear BART is applicable to logistic regression, multinomial logistic regression, and Poisson regression among other models.

This paper provides alternatives to Log-linear BART that retain the standard BART priors and do not rely on MCMC. [4] [5] A BART-BMA framework provides efficient greedy algorithms that outputs a relatively small number of parsimonious models. A BART-IS framework is straightforward to implement and trivially parallelizable.[6] The simple BART-BMA and BART-IS approaches provide readily implementable benchmarks for more complicated schemes such as the MCMC-based methods.

[3] Dorie et al. (2019) note that performance can be improved by using cross-validation to choose hyperparameters.

[4] While the focus of this paper is implementation algorithms, I also provide options for alternative model priors on the tree structures, including the prior proposed by Quadrianto & Ghahramani (2014) and the spike and tree prior Rockova & van der Pas (2017), in the **R** packages **logitbartBMA** and **safeBart** . Code is available at https://github.com/EoghanONeill

[5] The approach introduced in this paper can be combined with alternative parameter priors, e.g. different terminal node priors and hierarchical priors, provided the marginal likelihood can be efficiently calculated and it is possible to sample efficiently from a given sum-of-tree model (in the set of models being averaged). This possibility is a topic for future research.

[6] See chapter 2 of this thesis for further discussion of the usefulness of the BART-BMA and BART-IS algorithms.

The methods discussed in this chapter are relevant to a range of economic applications. Binary classification algorithms can be applied to propensity score estimation, and also prediction problems such as credit default prediction and prediction of consumer purchases. Multinomial regression methods are extensively used in modelling discrete choice problems in econometrics. The methods introduced in this chapter provide a flexible machine learning approach that accounts for model uncertainty and potentially complex functional forms.

The remainder of the paper is structured as follows. Section 3.2 provides a brief review of BART. Section 3.3 outlines the general framework for extending BART-BMA and BART-IS to a wide range of model settings. Section 3.4 describes the binary classification methods Logit-BART-IS and Logit-BART-BMA and compares the performance of these algorithms to other methods using publicly available datasets. Section 3.5 describes methods for treatment effect estimation with binary outcomes Logit-BCF-IS and Logit-BCF-BMA, and compares these algorithms to other methods using data from the ACIC 2019 data challenge. Section 3.6 discusses further model settings to which the generalized BART-BMA and BART-IS algorithms can be applied, with Tobit-BART-IS as an illustrative example.[7] Section 3.7 concludes the paper.

3.2 Review of BART and BART-BMA

In this section, we describe BART (Chipman et al. 2010), BART-BMA (Hernández et al. 2018), and an approximate, sub-optimal approach to implementation of Probit-BART-BMA and Probit-BART-IS that can be as a benchmark for the more principled approach introduced later in this paper.

This section repeats the overview from chapter 2, and is included for completeness so that this chapter is self-contained.

3.2.1 Overview of BART

Description of BART Model and Priors

Suppose there are n observations, and the $n \times p$ matrix of explanatory variables, X, has i^{th} row $x_i = [x_{i1}, ..., x_{ip}]$. For the standard BART model $Y_i = \sum_{j=1}^{m} g(x_i; T_j, M_j) + \varepsilon_i$, where $g(x_i; T_j, M_j)$ is the output of a decision tree. T_j refers to decision tree $j = 1, ..., m$, where m is the total number of trees in the model. M_j are the terminal node parameters of T_j, and $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

For BART (Chipman et al. 2010), prior independence is assumed across trees T_j and across terminal node means $M_j = (\mu_{1j}...\mu_{b_j j})$ (where $1, ..., b_j$ indexes the terminal nodes of tree j). The form of the prior used by Chipman et al. (2010) is:

$$p(M_1, ..., M_m, T_1, ..., T_m, \sigma) \propto \left[\prod_j \left[\prod_k p(\mu_{kj}|T_j) \right] p(T_j) \right] p(\sigma)$$

In standard BART, $\mu_{kj}|T_j \overset{i.i.d}{\sim} N(0, \sigma_0^2)$ where $\sigma_0 = \frac{0.5}{e\sqrt{m}}$ and e is a user-specified hyper-parameter.

Chipman et al. (2010) set a regularization prior on the tree size and shape $p(T_j)$ to discourage any one tree from having undue influence over the sum of trees. The probability that a given node within a tree T_j is split into two child nodes is $\alpha(1 + d_h)^{-\beta}$, where d_h is the depth of (internal) node h and α and β are parameters

[7] To the best of my knowledge, this is the first example of a Tobit-BART regardless of the implementation. An interesting topic for future research would be an MCMC based implementation of Tobit-BART.

which determine the size and shape of T_j respectively. Thus $p(T_j) = \prod_{h=1}^{b_j-1} \alpha(1+d_h)^{-\beta} \prod_{k=1}^{b_j}(1-\alpha(1+d_k)^{-\beta})$, where h indexes the internal nodes of the tree T_j, and k indexes the terminal nodes.

Chipman et al. (2010) assume that the model precision σ^{-2} has a conjugate prior distribution $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$ with degrees of freedom v and scale λ. There are also priors on the splitting variables and splitting points in each tree. Chipman et al. (2010) use the uniform prior on available splitting variables, and the uniform prior on the discrete set of available splitting variables.

3.2.2 Overview of BART-BMA

BART-BMA applies the same priors as standard BART (section 3.2.1), except the variance of the terminal node parameters is proportional to the variance of the error term, $\mu_{ij}|T, \sigma \sim N(0, \frac{\sigma^2}{a})$, as suggested by Chipman et al. (1998).[8] Integration of the likelihood with respect to the μ parameters and σ results in a closed form expression proportional to the marginal likelihood.

The marginal likelihood can be derived as follows. Let $Y = (Y_1, ..., Y_n)$ be the outcome vector. For a given sum of trees model $\mathcal{T}$, the posterior distribution of Y is:

$$Y|\mathcal{T}, M, \sigma^{-2} \sim N(\sum_{j=1}^{m} J_j M_j, \sigma^2 I)$$

where J_j (which depends on the original matrix of covariates X) is an $n \times b_j$ binary matrix with the element in position (i, j) indicating the inclusion of observation $i = 1, ..., n$ in terminal node $k = 1, ..., b_j$ of tree j.

Let $W = [J_1 ... J_m]$ be an $n \times b$ matrix , where $b = \sum_{j=1}^{m} b_j$ and $\underline{\boldsymbol{\mu}} = (M_1^T ... M_m^T)^T$ be a vector of size b of terminal nodes assigned to trees $T_1, ..., T_m$. We can then write $W\underline{\boldsymbol{\mu}} = \sum_{j=1}^{m} J_j M_j$,[9] and therefore

$$Y|\underline{\boldsymbol{\mu}}, \sigma^{-2} \sim N(W\underline{\boldsymbol{\mu}}, \sigma^2 I)$$

which, with $\underline{\boldsymbol{\mu}} \sim N(0, \frac{\sigma^2}{a} I_b)$, where I_b is a $b \times b$ identity matrix, implies

$$p(Y) = MVST_v(0, \lambda(I_n + \frac{1}{a} W W^T))$$

$$= \frac{\Gamma(\frac{\nu+n}{2})(\lambda v)^{\frac{\nu+n}{2}}}{\Gamma(\frac{v}{2}) v^{\frac{n}{2}} \pi^{\frac{n}{2}} \lambda^{\frac{n}{2}} (\frac{1}{a})^{\frac{b}{2}} det\,(aI_b + W^T W)^{\frac{1}{2}}} \left[\lambda v + Y^T Y - Y^T W (aI_b + W^T W)^{-1} W^T Y\right]^{-\frac{\nu+n}{2}}$$

Then, noting that anything that does not depend on W or b will cancel out when calculating the model weights, we can calculate:

$$\propto \frac{1}{(\frac{1}{a})^{\frac{b}{2}} det\,(aI_b + W^T W)^{\frac{1}{2}}} \left[\lambda v + Y^T Y - Y^T W (aI_b + W^T W)^{-1} W^T Y\right]^{-\frac{\nu+n}{2}}$$

And the log of this expression is: $\frac{b}{2}\log(a) - \frac{1}{2}\log(det(M)) - \frac{\nu+n}{2}\log(\lambda v + Y^T Y - Y^T W M^{-1} W^T Y)$ where $M = aI_b + W^T W$.

[8] Moran et al. (2018) argue against the use conjugate priors in Bayesian linear regression. However, this issue will not be discussed in further detail in this paper. Nonetheless, it is worth noting that the methods introduced in this paper can be improved further by careful calibration of the a parameter, e.g. by cross-validation.

[9] $W\underline{\boldsymbol{\mu}} = \sum_{j=1}^{m} J_j M_j$ is analogous to $X\boldsymbol{\beta}$ in standard linear regression notation.

A deterministic model search algorithm first reduces the set of potential splitting variables by a changepoint detection algorithm, and then recursively adds splits to trees that are potentially to be appended to models in the set of currently selected sum of tree models. After a set of single tree models are selected, changepoints in the residuals are used as potential splitting variables for constructing the next set of trees to potentially append to the selected models. Then a new set of residuals is constructed for the new set of sum-of-two-tree models, changepoints are detected, and trees are appended to create a set of sum-of-three-tree models, and so on.

The set of models to be averaged over are those with posterior probability within some distance of the highest probability model found by the model search algorithm. i.e. For all proposed models, $\mathcal{T}_\ell$, indexed by ℓ, the algorithm obtains

$$p(Y|\mathcal{T}_\ell, X)p(\mathcal{T}_\ell) \propto p(\mathcal{T}_\ell|Y, X) = \frac{p(Y|\mathcal{T}_\ell, X)p(\mathcal{T}_\ell)}{p(\mathbf{y})}$$

And keeps the models such that

$$\arg\max_{\ell'}(\log(p(\mathcal{T}_{\ell'}|Y, X))) - \log(p(\mathcal{T}_\ell|Y, X)) \leq \log(o)$$

where o is Occam's window, and the minimum is over the set of all proposed models.

3.3 Framework for Generalization of BART-BMA and BART-IS

BART-BMA and BART-IS are applicable to a wide range of model settings in which a linear combination of covariates can be replaced by a sum-of-tree model. For example, for Logit-BART, the latent outcome can be modelled as a sum-of-trees instead of a standard linear model.

A key requirement for the computational feasibility of this general framework is that there should exist efficient methods for calculating the marginal likelihood and posterior predictions of the model of interest. A closed form for the posterior distribution or an efficient method for sampling from the posterior distribution is required for sampling any quantities of interest or producing credible intervals.

For models such as Logit and Tobit, it is possible to obtain an approximation to the marginal likelihood by a Laplace approximation about the Maximum a Posteriori parameter estimates, which can be obtained efficiently from a Quasi-Newton algorithm (Murphy 2012, Chib 1992). Similar approaches can be used for other models. Recently developed methods, including Integrated Nested Laplace Approximations (Rue et al. 2009), are applicable to a wide range of models including multinomial logit, Poisson regression, and models with hierarchical priors, e.g. mixed logit.[10]

Algorithm 1 and Algorithm 2 outline the general BART-BMA and BART-IS algorithms. The General BART-BMA algorithm begins by constructing latent outcome variable values for the training data (this is necessarily model-specific and arbitrary) and then applying a changepoint detection algorithm to obtain a set of potential splitting rules.[11] Single tree models are constructed using these splitting rules, as in standard BART-BMA (see chapter 2), but with marginal likelihood calculations that are specific to the generalized linear model and approximation method. Then a new set of residuals are calculated for each single-tree model in Occam's window, and changepoints are found for these new residuals. The new changepoints are used to construct new trees to be appended to the existing models, creating a set of sum-of-two tree models.

[10] However, there is a trade-off between accuracy of approximations and computational speed. In some cases it might not be computationally feasible to place a model in the BART-IS framework. This would require a level of experimentation with different approximation methods.

[11] Changepoint detection algorithms include Pruned Exact Linear Time Killick et al. (2012) and a simple grid-search.

Residuals are calculated for the sum-of-two tree models and a set of sum-of-three tree models are created, and so on. The General BART-IS algorithm is essentially the same as the algorithm presented in chapter 2, except the marginal likelihood and sampling from the mixture of posterior distributions are specific to the generalized linear model and approximation method. The random draws of trees are the same as in chapter 2,[12] and the General BART-IS is highly parallelizable as each iteration of the for-loop can be assigned to a different processor.

A limitation of the BART-BMA approach is that it requires the calculation of residuals and application of a changepoint detection algorithm to the residuals for the purpose of reducing the set of potential splitting variables. For some models, the residuals are of a latent outcome, and it is not clear how to proceed. In the case of Logit-BART-BMA, section 3.4 outlines how it is possible to make use of existing ideas for logit boosted tree methods (Friedman et al. 2000). However, there might not exist a straightforward and effective method for calculation of residuals for some models, and therefore the BART-IS approach, for which there is no calculation of residuals nor data-dependent search for splitting points, is more general. For General BART-BMA, the latent outcome is unknown for any observations in the training data, and the initialization is entirely arbitrary and model-specific. It is not guaranteed that there exists an initialization that leads to an effective model search for all models, and an entirely different model search algorithm without construction of latent outcome values or residuals may be more effective.

For averages of models with multiple latent outcomes (each modelled by a sum-of-trees) per model, the BART-BMA approach is infeasible[13] and the BART-IS approach remains feasible (although a larger number of models should be sampled). An example of such a model would be multinomial logistic regression with different sums-of-trees for the latent utility of each alternative.[14] However, the discussion in this paper will be restricted to settings where the same underlying variables (or sum-of-trees) are used for all latent variables.

A word of caution is required here. The performance of these methods is highly dependent on the appropriateness of the overall model specification (e.g. logit link function), the accuracy of the approximations of the marginal likelihood and posteriors. In the case of BART-BMA, the model search algorithm might not perform as well as for a simple linear model, and parameters such as the size of Occam's window and changepoint detection parameters may have to be tuned to control the trade-off between computational feasibility and breadth of the model search. BART-IS generally requires a large number of draws of models, and the feasibility of the approach is inversely related to the size of the model space and the computational time required to calculate the marginal likelihood.

[12] The samples of models can be made "offline", i.e. before any data is obtained, as in BART-IS and safe-Bayesian Random Forests (Quadrianto & Ghahramani 2014).

[13] A form of BART-BMA with considerable changes to the model search algorithm might be possible. This is beyond the scope of this paper.

[14] It is possible to share the same sum-of-tree structure, e.g. Linero et al. (2019), or sample separate sums-of-trees, e.g. Murray (2017).

Input: $n \times p$ matrix X

Response Y. Vector of binary, censored, categorical, or count data, or perhaps a more complicated (e.g. multivariate) outcome.

Output: Depends on the model. e.g. predicted outcomes or probabilities, parameter estimates.

Initialize: *Residuals*. Details depend on the model setting, e.g. for standard BART-BMA, begin with the vector of outcomes, and for Logit-BART-BMA begin with a transformation to the scale of the latent outcome (η_i in equation (3.1)). In general the residuals should be on the scale of the (possibly latent) variable that is directly modelled by a sum-of-trees.

Initialize *lowest_model_prob*, the minimum posterior probability of all models found so far.

Initialize: $L = 1$, Set the list of models $List\_ST$ to include a single tree model with no splits.

[Each round in the outer loop searches over possible additions of one tree to existing sum-of-tree models. (First round begins with single tree models)]

for $j \leftarrow 1$ *to* num_trees **do**

  [For each model ℓ in OW from the previous round, search for trees to add] **for** $\ell \leftarrow 1$ *to* L **do**

    if $count\_mu\_trees_\ell \leq m_\mu$ **then**

      1. **Find Good Splitting Rules**.

         Apply a changepoint detection algorithm to the residuals to reduce the number of potential splitting rules. This is model-specific, and may involve first applying some function to the residuals. See Logit-BART-BMA for an example.

      2. **Grow trees to append to sum-of-tree model**.

         Begin with a tree stump and grow trees recursively using splitting rules from step 1. Each time a split is considered, calculate the posterior model probability and check if the model is in OW. [This requires efficient calculation of the marginal likelihood].

         Add new models to temporary list $temp\_OW$ if in OW.

    end

    Make sum of trees models and update residuals

    Reset list of models in OW $List\_ST = temp\_OW$.

    Update *lowest_model_prob* to minimum posterior probability of models in $List\_ST$.

    Set $L = length(temp\_OW)$. Reset $temp_{OW}$ to list of length zero.

  end

end

Delete models in $list\_ST$ with log posterior probability more than $log(o)$ from *lowest_model_prob*.

The output is a model averaged prediction of an outcome/probability or parameter estimate.

Intervals can be obtained from either a closed form expression or probability-weighted sampling from each model in OW.

Algorithm 1: BART-BMA General Algorithm

Input: $n \times p$ matrix X

Response Y. Vector of binary, censored, categorical, or count data, or perhaps a more complicated (e.g. multivariate) outcome.

Output: Depends on the model. e.g. predicted outcomes or probabilities, parameter estimates.

Each round in the outer loop involves drawing a model from a model sampler. This loop is trivially parallelizable.

for $m \leftarrow 1$ *to num_models* **do**

1. Draw a model from the model sampler. This can be the sampler used by Quadrianto & Ghahramani (2014), the BART prior, or the spike and tree prior (Rockova & van der Pas 2017).

2. Obtain the model predictions and/or parameters that summarize the (possibly approximate) posterior distribution.

3. Obtain model weights. This requires efficient calculation of the marginal likelihood. If the model sampler is not the model prior, then multiply the marginal likelihood by the ratio of the model prior probability to the model sampler probability. For a safe-Bayesian approach, use the marginal likelihood to the power of a number between 0 and 1 (Quadrianto & Ghahramani 2014).

end

The output depends on the model and object of interest.

e.g. The predicted outcome or probability is a marginal likelihood weighted average of model predictions.

Parameter distributions and credible intervals can be obtained from model weighted samples from (possibly approximate) posterior distributions. In some cases a closed form gives an efficient alternative.

Algorithm 2: BART-IS General Algorithm

3.4 Example of General Algorithms Applied to Binary Outcome Data: Logit-BART-BMA and Logit-BART-IS

In this section, the general algorithms introduced in section 3.3 are applied to Logit-sum-of-tree models for binary outcome data. First, an outline is given for a simpler benchmark approach that does not make use of the more principled algorithm. Second, the binary outcome model and Laplace approximation method are summarized. The Logit-BART-BMA and Logit-BART-IS algorithms are detailed as specific examples of the general framework. Finally, the methods are applied to binary classification datasets.

3.4.1 A Benchmark Probit Approximation for BART-BMA and BART-IS

Single tree methods can readily be applied to binary outcome data. Trees in a random forest applied to binary outcomes produce predictions between zero and one because leaf estimates are averages of binary variables. However, sum-of-tree based methods such as BART are less directly applicable to binary outcome data because sums-of-trees can produce predictions outside the range $[0, 1]$ and ideally the statistical framework of BART should account for the fact that the outcomes are binary.

A simple extension of BART to Probit involves first converting the binary outcomes to the scale of the latent variable, replacing observations $y_i = 1$ with $y_i^* = 3.1$ and replacing observations $y_i = 0$ with $y_i^* = -3.1$. These latent variable values correspond to very high and very low probabilities of $y_i = 1$. Then standard BART-MCMC, BART-BMA, or BART-IS is applied to the data with y_i^* as the dependent variable. Finally, the normal CDF function is applied to the latent outcome predictions to obtain predicted probabilities, and applied to the latent outcome prediction intervals to obtain prediction intervals for the probabilities. This is the approach adopted by Hernández et al. (2018) and will be referred to in the remainder of this document as Approximate Probit-BART-BMA.[15] Similarly, this approach in combination with the BART-IS algorithm will be referred to as Approximate Probit-BART-IS.

However, the approach outlined above does not truly apply a binary outcome model to the data. A more rigorous approach would involve a likelihood that accounts for the probability that the actual outcome equals to zero or one, and not begin with arbitrary values for the latent outcome. The framework outlined in section 3.3 provides one possible method for implementing the more rigorous approach.

3.4.2 Model, Priors, and Notation for Logit-BART

Throughout this chapter, the notation is chosen to be similar to that used by Hernández et al. (2018). The prior for the terminal node parameters is $\mu_{k,j}|T, \sigma \sim N(0, \frac{1}{a})$ (where j, k denotes the j^{th} terminal node of the k^{th} tree in the sum-of-tree model), and unlike in BART-BMA for continuous outcomes, there is not a separate parameter for the variance of the error term (the variance of the error term is not separately identified).

Let $W = [J_1 ... J_m]$ be an $n \times b$ matrix , where $b = \sum_{j=1}^{m} b_j$, J_j is a binary matrix of size $n \times b_j$ with the element in the i^{th} row and j^{th} column denoting the inclusion of observation $i = 1, ..., n$ in terminal node $k = 1, ..., b_j$ of tree j. Let M_j be a vector $(\mu_{1,j}, ..., \mu_{b_j,j})$ of terminal node means for the j^{th} tree, and let $\underline{\boldsymbol{\mu}} = (M_1^T ... M_m^T)^T$ be a vector of size b of terminal node means assigned to trees $T_1, ..., T_m$. We can then write $W\underline{\boldsymbol{\mu}} = \sum_{j=1}^{m} J_j M_j$. The product $W\underline{\boldsymbol{\mu}} = \sum_{j=1}^{m} J_j M_j$ is analogous to $X\boldsymbol{\beta}$ in standard linear regression notation. Let W_i denote the i^{th} row of W.

The outcomes are binary, $y_i \in \{0, 1\}$. The probability of the outcome $y_i = 1$ is given by the logistic function, and will be denoted by p_i for convenience:

$$p_i = \Pr(y_i = 1|W_i, \underline{\boldsymbol{\mu}}) = \frac{1}{1 + e^{-\underline{\boldsymbol{\mu}}^T W_i^T}} = \frac{e^{W_i \underline{\boldsymbol{\mu}}}}{1 + e^{W_i \underline{\boldsymbol{\mu}}}}$$

where W_i denotes the i^{th} row of W. The likelihood is: $p(\boldsymbol{y}|W, \underline{\boldsymbol{\mu}}) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$.[16] The log-likelihood is $\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i) \log(1 - p_i)] = \boldsymbol{y}^T W \underline{\boldsymbol{\mu}} - \sum_{i=1}^{N} \log(1 + e^{-W_i \underline{\boldsymbol{\mu}}})$.

[15] The improvements to the BART-BMA algorithm described in chapter 2 of this thesis also apply to this approximate Probit-BART implementation.

[16] Note that W is defined by the sum-of-tree model $\mathcal{T}$. Conditioning on the model is excluded here for brevity.

3.4.3 Laplace Approximation

The prior $\underline{\boldsymbol{\mu}} \sim \mathcal{N}(0, \frac{1}{a} I_b)$ and the likelihood give an intractable posterior distribution. However, a Laplace approximation gives a normal posterior distribution for the terminal node parameters. An approximation of the posterior can be obtained by s second order Taylor expansion about the Maximum A Posteriori (MAP) estimate:

$$\underline{\boldsymbol{\mu}}_{MAP} = \arg\min_{\underline{\boldsymbol{\mu}}} -(\log p(\boldsymbol{y}|W, \underline{\boldsymbol{\mu}}) + \log p(\underline{\boldsymbol{\mu}}))$$

$$= \arg\min_{\underline{\boldsymbol{\mu}}} - \left[\boldsymbol{y}^T W \underline{\boldsymbol{\mu}} - \sum_{i=1}^{N} \log(1 + e^{-W_i \underline{\boldsymbol{\mu}}}) - \frac{1}{2} b \log(2\pi) + \frac{1}{2} b \log(a) - \frac{a}{2} \underline{\boldsymbol{\mu}}^T \underline{\boldsymbol{\mu}} \right]$$

The approximate distribution is:

$$p(\underline{\boldsymbol{\mu}}|\boldsymbol{y}, W) \approx \mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP}, H^{-1})$$

where H is the Hessian matrix of the negative log posterior (evaluated at the MAP).

$$H = W^T S W + a I_b$$

where $S = diag(p_i(1-p_i))$ is an $n \times n$ diagonal matrix with diagonal elements determined by the probabilities p_i obtained from the logistic function. The Hessian and the gradient of the negative posterior probability can be used to obtain an approximation of the MAP. The gradient is $\boldsymbol{g} = W^T(\boldsymbol{p} - \boldsymbol{y}) + a\underline{\boldsymbol{\mu}}$ where $\boldsymbol{p} = (p_1, ..., p_n)^T$. The MAP can be found by Newton's method or more efficient Quasi-Newton methods such as the limited memory BFGS (L-BFGS) algorithm.[17]

When averaging over the set of sum-of-tree models $\mathcal{T}_1, ..., \mathcal{T}_M$, the approximate distribution of the parameters is:

$$\underline{\boldsymbol{\mu}}|\boldsymbol{y} \sim \sum_{m=1}^{M} \mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H_{(m)}^{-1}) p(\mathcal{T}_m|\boldsymbol{y})$$

where $p(\mathcal{T}_m|\boldsymbol{y})$ is the posterior model probability,

$$p(\mathcal{T}_m|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mathcal{T}_m) p(\mathcal{T}_m)$$

where $p(\boldsymbol{y}|\mathcal{T}_m)$ is the marginal likelihood, which can be approximated using the Laplace approximation, as outlined in Appendix C.2, and $p(\mathcal{T}_m)$ is the prior model probability. The prior probability is the same as for BART-BMA for continuous outcomes and straightforward to calculate or, in the case of BART-IS, it does not need to be calculated.

The subsections below include details for estimating the posterior mean, calculating credible intervals, and calculating the marginal likelihood.

[17] See appendix C.1 for the standard Newton method for finding the minimum of the negative log of the posterior distribution. The implementations provided in the **R** packages **safeBart** and **logitbartbma** use L-BFGS.

Estimation of Posterior Predictive Mean Probability

The model averaged approximate posterior for coefficients is $\underline{\boldsymbol{\mu}}|\boldsymbol{y} \sim \sum_{m=1}^{M} \mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H_{(m)}^{-1}) p(\mathcal{T}_m|\boldsymbol{y})$. Using the logistic (sigmoid) function probability $\frac{e^{W_i \underline{\boldsymbol{\mu}}}}{1+e^{W_i \underline{\boldsymbol{\mu}}}}$, the model averaged posterior (predictive) probability is:

$$p(y_* = 1|\boldsymbol{x}_*, X, \boldsymbol{y}) = \sum_{m=1}^{M} \left(\int \frac{e^{W_{*,(m)} \underline{\boldsymbol{\mu}}_{(m)}}}{1 + e^{W_{*,(m)} \underline{\boldsymbol{\mu}}_{(m)}}} p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H_{(m)}^{-1}) d\underline{\boldsymbol{\mu}}_{(m)} \right) p(\mathcal{T}_m|\boldsymbol{y})$$

where y_* is the outcome for a new observation, X is the matrix of variables in the training data, $\boldsymbol{y}$ is the vector of outcomes in the training data, $\boldsymbol{x}_*$ is the covariate vector for the new observation, which is input to the sum-of-tree models to obtain row vectors for each model, $W_{*,(m)}$, $m = 1, ..., M$, consisting of binary variables to indicate inclusion in terminal nodes.

The integral in the above expression for $p(y_* = 1|\boldsymbol{x}_*, X, \boldsymbol{y})$ is intractable. Numerous approaches are possible for estimation of predictive probabilities, and calculation of the marginal likelihood and credible intervals.[18] The example below outlines a standard Laplace approximation with the probit function (normal CDF) used as an approximation to the logistic (sigmoid) function because this approach fast, straightforward to implement, and can be used to benchmark other approaches. Appendix C.4 outlines simple Monte Carlo alternatives.

Probit Approximation of Posterior Predictive Mean Probability

Machine learning methods often combine the Laplace approximation for logistic regression with a normal CDF approximation (Spiegelhalter & Lauritzen 1990, Bishop 2006, Murphy 2012). The logistic (sigmoid) function can be approximated by the normal CDF:

$$p(y_* = 1|\boldsymbol{x}_*, X, \boldsymbol{y}) = \sum_{m=1}^{M} \left(\int \frac{e^{W_{*,(m)} \underline{\boldsymbol{\mu}}_{(m)}}}{1 + e^{W_{*,(m)} \underline{\boldsymbol{\mu}}_{(m)}}} p(\underline{\boldsymbol{\mu}}_{(m)}|O_{MAP,(m)}, H_{(m)}^{-1}) d\underline{\boldsymbol{\mu}}_{(m)} \right) p(\mathcal{T}_m|\boldsymbol{y})$$

$$\approx \sum_{m=1}^{M} \left(\int \Phi(e^{W_{*,(m)} \underline{\boldsymbol{\mu}}_{(m)}}) p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H_{(m)}^{-1}) d\underline{\boldsymbol{\mu}}_{(m)} \right) p(\mathcal{T}_m|\boldsymbol{y})$$

The integrals in the above expression can be rewritten as one-dimensional integrals:

$$p(y_* = 1|\boldsymbol{x}_*, X, \boldsymbol{y}) \approx \sum_{m=1}^{M} \left(\int \Phi(\alpha_{(m)}) p(\alpha_{(m)}|\psi_{\alpha,(m)}, \sigma^2_{\alpha,(m)}) d\alpha_{(m)} \right) p(\mathcal{T}_m|\boldsymbol{y}) = \sum_{m=1}^{M} \Phi \left(\frac{\psi_{\alpha,(m)}}{\sqrt{1 + \sigma^2_{\alpha,(m)}}} \right) p(\mathcal{T}_m|\boldsymbol{y})$$

where $\psi_{\alpha,(m)} = W_{*,(m)} \underline{\boldsymbol{\mu}}_{MAP,(m)}$ and $\sigma^2_{\alpha,(m)} = W_{*,(m)} H_{(m)}^{-1} W_{*,(m)}^T$. For each model, the distribution of $\alpha_{(m)} = W_{*,(m)} \underline{\boldsymbol{\mu}}_{(m)}$ is $\mathcal{N}(\psi_{\alpha,(m)}, \sigma^2_{\alpha,(m)})$.

Often 1 is replaced by t^{-2} where $t^2 = \frac{\pi}{8}$ to give a closer approximation to the probability that would have been obtained from the logistic function (Spiegelhalter & Lauritzen 1990, Bishop 2006, Murphy 2012):

$$p(y_* = 1|\boldsymbol{x}_*, X, \boldsymbol{y}) \approx \sum_{m=1}^{M} \Phi \left(\frac{\psi_{\alpha,(m)}}{\sqrt{\frac{8}{\pi} + \sigma^2_{\alpha,(m)}}} \right) p(\mathcal{T}_m|\boldsymbol{y})$$

A number of alternative approaches exist for calculating the marginal likelihood and posterior mean.

[18] See Chopin et al. (2017) for a discussion

Appendix C.3 describes an approach for estimating the posterior mean that involves applying Laplace's method twice (Tierney & Kadane 1986). Calculation of credible intervals by root-finding or Monte Carlo draws from the posterior is straightforward and detailed in Appendices C.5 and C.4.2.

Alternative to Probit Approximation: Gibbs Sampler for Final Inference (Laplace Approximation for Marginal Likelihoods)

As in the original BART-BMA paper (Hernández et al. 2018), after the models are selected (or sampled in the case of BART-IS) it is possible to use a Gibbs sampler to take draws from each model, and draw from each model with probability equal to the posterior model probability. In the case of Logit-BART-BMA, this can be implemented by estimating the posterior model probability using a Laplace approximation (as outlined above), or some other method[19] and then taking "exact" draws (from the true model rather than an approximation) using a Gibbs sampler (Albert & Chib 1993). For each draw of model parameters, it is possible to calculate the quantity of interest. e.g. $\frac{e^{W_{*,(m)}\underline{\mu}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\mu}_{(m),s}}}$ (or differences in probabilities for treatment effects). Then the mean and quantiles of the values across samples can be used for predictions and credible intervals.

The Gibbs sampler described by Polson et al. (2013) is potentially well-suited to this purpose because it is fast and uniformly ergodic (Choi et al. 2013). Polson et al. (2013) note that their sampler "opens the door for exact Bayesian treatments of many modern-day machine-learning classification methods based on mixtures of logits ".

3.4.4 Logit-BART-BMA

The prior over the model space is the same as for standard BART-BMA and BART-IS.[20] The Logit-BART-BMA model search algorithm is a special case of Algorithm 1 and only differs from the standard BART-BMA algorithm in the calculation of residuals and application of a changepoint detection algorithm to the residuals to reduce the number of potential splitting rules. This section discusses a number of possible approaches to the calculation of residuals and the changepoint detection algorithm for Logit-BART-BMA.

There are a few potential methods for suggesting potential splitting rules in each round of the model search algorithm. The approach presented here is inspired by the LogitBoost algorithm (Friedman et al. 2000). Alternative approaches are detailed in Appendix C.6.

A variant of AdaBoost, LogitBoost (Friedman et al. 2000), involves fitting a base learner to be added to a sum of models (i.e. boosted models), to which the logistic function is then applied to obtain the probability. Each base learner minimizes the weighted sum of squares, i.e. applies weighted least squares, to the following variable:

$$z_i = \frac{y_i - p(x_i)}{p(x_i)(1-p(x_i))}$$

with weights $w_i = p(x_i)(1-p(x_i))$, where $p(x_i)$ is the individual-specific probability estimated in the previous round, initialized at $p(x_i) = 0.5$. [21]

Logit BART-BMA estimates the whole logit model at each step, and therefore z_i is only really relevant

[19] See Friel & Wyse (2012) for a review of possible methods.

[20] Alternative priors on the tree structures, provided in the **R** packages **logitbartBMA** and **safeBart** include the prior proposed by Quadrianto & Ghahramani (2014) and the spike and tree prior Rockova & van der Pas (2017). Code is available at https://github.com/EoghanONeill

[21] It is possible to apply the restriction $z_i \in [-3, 3]$ and also apply trimming or another method to avoid numerical instability issues when dividing by $p(x_i)(1-p(x_i))$ when $p(x_i)$ is close to zero or one.

to the initial stage in each round that involves applying the changepoint detection algorithm. The proposed approach here is to apply the changepoint detection algorithm to z_i with a weighted (sum of squares) cost function with weights w_i. [22]

The key idea, as with AdaBoost, is that the set of changepoints used in constructing new trees to be appended to the model, should place more weight on observations misclassified by the current model. However, unlike AdaBoost, there is no such adjustment made in the final criterion for the acceptance of the new trees because the entire sum-of-tree model is re-estimated when a new tree is appended to a model and the marginal likelihood based criterion is applied to the entire model.

This adjustment to changepoint detection is also applicable to the naive approximation to Probit-BART-BMA and Probit-BART-IS discussed in section 3.4.1. In approximate Probit-BART-BMA, it is also possible to fit the new tree using z_i as in LogitBoost. However, this is a topic for future research. [23]

3.4.5 Logit-BART-IS

Logit-BART-IS is a special case of the general framework given in Algorithm 2. Algorithm 3 outlines how to apply the efficient logit approximation methods described in section 3.4.3 in the general BART-IS framework. Logit-BART-BMA does not involve model search, and therefore does not involve initialization of latent outcome values or calculation of residuals.

Input: $n \times p$ matrix X

Response binary vector Y.

Output: Predictive probabilities, intervals for predictive probabilities.

Each round in the outer loop involves drawing a model from a model sampler. This loop is trivially parallelizable.

for $m \leftarrow 1$ *to num_models* **do**

1. Draw a model from the model sampler. This can be the sampler used by Quadrianto & Ghahramani (2014), the BART prior (Chipman et al. 2010), or the spike and tree prior (Rockova & van der Pas 2017).
2. Obtain MAP parameter values for a Laplace approximation as outlined in section 3.4.3. Obtain predicted probabilities as outlined in section 3.4.3.
3. Obtain model weights. The marginal likelihood is efficiently calculated as outlined in section C.2.

end

Model averaged predictions are calculated as outlined in section 3.4.3.

Credible intervals for the model averaged distribution are obtained as outlined in Appendix C.5.

Algorithm 3: Logit BART-IS Algorithm

[22] The weights w_i essentially account for second-order information. Other methods such as MART (Friedman & Meulman 2003) only use $y_i - p(x_i)$ in the tree building step (but use second order information when estimating terminal nodes values).

[23] This is more applicable to the original BART-BMA implementation of Hernández et al. (2018) that estimated each new tree separately using only residuals, and less applicable to the new BART-BMA implementation presented in chapter 2 of this thesis which estimates the whole model at each step. In this sense the new implementation of BART-BMA is more analogous to variations on AdaBoost algorithms that perform backfitting at each step.

3.4.6 Application to UCI Datasets

This section contains a comparison of Logit-BART-BMA and Logit-BART-IS against other methods using publicly available datasets from the widely used UCI Machine Learning Repository (Dua & Graff 2017). The chosen datasets are binary classification datasets relevant to economic applications. Table 3.1 contains a description of the data. Missing observations are removed from all datasets. The number of variables is the number remaining after removal of some variables (e.g. unique text strings), and transformation of some categorical variables into multiple binary variables so that tree-based methods are applicable with available software.

The algorithms compared are Logit-BART-IS,[24] Logit-BART-BMA, Approximate Probit BART-IS,[25] Approximate Probit BART-BMA, Probit-BART-MCMC, Logit-BART-MCMC,[26] linear logistic regression, and Random Forests.[27] Methods are evaluated using the Brier Score and Area Under the Curve (AUC).[28] The data is randomly divided into training and hold-out test data, and all methods are applied without parameter tuning.

Tables 3.1 to 3.6 show the binary classification results for a range of training sample sizes. Across many examples, the Logit-BART-IS and Logit-BART-BMA implementations are surprisingly competitive with the MCMC implementations given the small number of trees in each model and relatively small number of sampled models for BART-IS and very small number of models in Occam's window in BART-BMA. For a number of datasets, the more principled general framework with Laplace approximations provides a notable improvement over the probit transformation approach described in section 3.4.1. The results demonstrate that the general BART approach introduced in this paper produces the intended result of estimates that are similar to those produced by MCMC BART implementations.[29] Logit-BART-BMA results are only presented for the training samples of up to 2000 observations due to the computational time required for some datasets under default parameters.[30]

The similar performance of Logit-BART-MCMC and Probit-BART-MCMC is unsurprising. The fact that there is no consistently best performing model across all sample sizes and datasets (although Probit-BART-MCMC is the method that slightly outperforms other methods across the most datasets) indicates that all methods produce similar estimates and the ranking of methods may be influenced by random variation in the data and splitting into test and training data. It is possible that for some datasets, MCMC does not deliver notable improvements over simple BMA or IS based approaches, while for other datasets there may

[24] Logit-BART-IS was implemented with only 5 trees per model, and a total of 20,000 sampled models. This is a small number of draws relative to the number of models drawn for BART-IS with continuous outcomes in the second chapter of this thesis. Each model takes more computational time than a linear model, therefore some compromise must be made on computational speed. However, the results are surprisingly competitive with the MCMC implementations considering the small number of samples. Therefore a topic for future research would be whether the results are more accurate with a larger set of samples, perhaps using parallelization over a larger number of cores for computational feasibility.

[25] Approximate Probit BART-IS was implemented with only 10 trees per model and a total of only 1000 sampled models. The results are surprisingly competitive given the small number of samples.

[26] Probit-BART-MCMC and Logit-BART-MCMC were both implemented using the **R** package **BART** with 5000 burn-in draws and 10,000 post-burn-in draws. Each model sampled by Probit-BART-MCMC has the default number of 50 trees, and each model sampled by Logit-BART-MCMC has the default number of trees of 200.

[27] Random Forests were implemented using the **R** package **ranger** and 10,000 trees. All other parameters were set to the default values.

[28] The Brier score is defined as $\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{p}_i)^2$ where N is the number of samples in the holdout data and $\hat{p}_i$ is the predicted probability that $y_i = 1$. The area under the Receiver Operating Characteristic curve is calculated using the **R** package **ROCR**.

[29] It is not expected that the Logit-BART-BMA or Logit-BART-IS results give notably more accurate estimates than the MCMC based implementations as ultimately these are alternative implementations for essentially the same model framework.

[30] The computational requirements for the Logit-BART-BMA search algorithm are likely to be sensitive to model search parameters, model prior parameters, and choice of changepoint detection algorithm. The optimal choice of parameters may differ across datasets, and differ to standard BART-BMA for continuous outcomes.

be potential for more substantial gains from MCMC.[31]

The results for the `Census Income` dataset indicate that the IS based approach does not perform as well as the MCMC approach, and no approaches perform markedly better than standard logistic regression. It is possible that there is a strong linear relationship between one of the covariates and the dependent variable, and therefore the logistic regression model performs well. This may also explain the poor results for Logit-BART-IS when applied to this dataset because an implementation that makes use of fewer trees per sum-of-tree model is likely to be less precise at capturing linear functions of covariates.[32]

[31] See Chopin et al. (2017) for a similar discussion regarding sampling of parameters in a single logistic regression model.

[32] Probit-BART-MCMC was implemented with 50 trees per model, Logit-BART-MCMC was implemented with 200 trees per model, and Logit-BART-IS and Logit-BART-BMA were implemented with 5 trees per model.

Dataset name	Description (from UCI repository)	Number of variables	Number of Observations	Reference
`Shopper`	Online Shoppers Purchasing Intention Dataset Data Set.	74	12,330	Sakar et al. (2019)
`Bank Marketing`	The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).	51	4521	Moro et al. (2014)
`Insurance`	This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company (caravan insurance in the Netherlands). The data consists of 86 variables and includes product usage data and socio-demographic data.	133	5821	Van Der Putten & van Someren (2000)
`Credit Cards`	Prediction of customer default in Taiwan.	33	30,000	Yeh & Lien (2009)
`Credit Screening`	Examples represent positive and negative instances of people who were and were not granted credit by a Japanese company that grants credit.	46	653	None
`German Credit`	Statlog (German Credit) Data Set. This dataset classifies people described by a set of attributes as good or bad credit risks.	61	1000	None
`Australian Credit`	Statlog (Australian Credit Approval) Data Set. This file concerns credit card applications.	42	690	Quinlan (1987)
`Census Income`	Predict whether income exceeds \$50K/yr based on census data. Extraction was done by Barry Becker from the 1994 US Census database.	64	30,162 training, 15,060 testing	Kohavi (1996)

Table 3.1: UCI Dataset descriptions

UCI Binary Outcome Data Results

500 Training Observations

	Shopper		Bank Marketing		Insurance		Credit Cards	
Method	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC
Logit-BART-IS	**0.079**	0.903	0.081	0.865	0.057	0.661	0.145	0.751
Logit-BART-BMA	0.083*	0.868*	0.084	0.841	0.058	0.653	0.143	0.727
Approx-Probit-BART-IS	0.145	0.661	0.102	0.715	0.060	0.591	0.175	0.698
Approx-Probit-BART-BMA	0.091	0.866	0.090	0.861	0.065	0.603	0.172	0.719
Probit-BART-MCMC	**0.079**	**0.906**	0.080	0.870	**0.056**	**0.688**	0.141	0.751
Logit-BART-MCMC	0.081	0.905	0.082	0.854	**0.056**	0.679	**0.140**	**0.753**
Logistic Regression	0.149	0.692	0.559	0.485	0.104	0.580	0.164	0.688
RF	0.085	0.897	**0.077**	**0.883**	0.059	0.591	0.145	0.745
Holdout sample size	11,830		4021		5321		29,500	

* indicates where the PELT algorithm with unweighted residuals was used for reducing the number of splitting points.

Table 3.2: UCI Binary Classification Datasets, training sample size = 500

500 Training Observations

	Credit Screening		German Credit		Australian Credit		Census Income	
Method	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC
Logit-BART-IS	0.097	0.930	0.178	0.737	0.147	0.873	0.128	0.859
Logit-BART-BMA	0.095	0.930	0.187	0.713	**0.101**	**0.927**	0.124	0.862
Approx-Probit-BART-IS	0.210	0.811	0.188	0.736	0.224	0.749	0.155	0.768
Approx-Probit-BART-BMA	0.102	0.912	0.214	0.712	0.163	0.888	0.135	0.855
Probit-BART-MCMC	0.096	0.924	0.164	0.777	0.142	0.875	**0.113**	**0.888**
Logit-BART-MCMC	0.097	0.919	**0.162**	**0.778**	0.147	0.871	0.118	0.879
Logistic Regression	0.129	0.857	0.168	0.775	0.158	0.858	0.140	0.845
RF	**0.092**	**0.942**	0.167	0.772	0.131	0.887	**0.113**	0.886
Holdout sample size	153		500		190		15,060	

Table 3.3: UCI Binary Classification Datasets, training sample size = 500

1000 Training Observations

	Shopper		Bank Marketing		Insurance		Credit Cards		Census Income	
Method	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC
Logit-BART-IS	**0.077**	**0.915**	0.080	0.849	0.055	**0.751**	**0.140**	**0.756**	0.123	0.864
Logit-BART-BMA	0.100	0.791	0.080	0.842	0.056	0.711	0.142	0.735	0.119	0.871
Approx-Probit-BART-IS	0.144	0.680	0.108	0.752	0.060	0.659	0.178	0.714	0.197	0.763
Approx-Probit-BART-BMA	0.088	0.898	0.085	0.846	0.060	0.635	0.182	0.692	0.118	0.888
Probit-BART-MCMC	0.078	0.909	**0.075**	0.879	**0.054**	0.744	0.141	0.751	**0.106**	**0.899**
Logit-BART-MCMC	0.079	0.908	0.077	0.874	0.055	0.733	**0.140**	0.754	0.109	0.895
Logistic Regression	0.160	0.749	0.080	0.848	0.078	0.583	0.150	0.706	0.220	0.641
RF	0.080	0.908	0.076	**0.885**	0.058	0.630	0.144	0.740	0.109	0.896
Holdout sample size	11,330		3521		4821		29,000		15,060	

Table 3.4: UCI Binary Classification Datasets, training sample size = 1000

2000 Training Observations

	Shopper		Bank Marketing		Insurance		Credit Cards		Census Income	
Method	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC
Logit-BART-IS	0.075	0.912	0.085	0.848	**0.050**	0.717	0.141	0.744	0.124	0.871
Logit-BART-BMA	0.097	0.793	0.076	0.860	0.052	0.722	0.140	0.739	0.116	0.879
Approx-Probit-BART-IS	0.144	0.698	0.103	0.689	0.055	0.641	0.172	0.701	0.185	0.758
Approx-Probit-BART-BMA	0.085	0.906	0.081	0.846	0.054	0.707	0.165	0.711	0.121	0.885
Probit-BART-MCMC	**0.075**	**0.917**	**0.072**	0.896	**0.050**	**0.753**	**0.137**	**0.766**	**0.102**	**0.909**
Logit-BART-MCMC	0.076	0.915	0.073	0.895	**0.050**	0.751	**0.137**	**0.766**	**0.102**	0.908
Logistic Regression	0.135	0.705	0.080	0.864	0.058	0.673	0.148	0.711	0.111	0.888
RF	0.076	**0.917**	**0.072**	**0.899**	0.052	0.718	0.140	0.764	0.104	0.905
Holdout sample size	10,330		2521		2821		28,000		15,060	

Table 3.5: UCI Binary Classification Datasets, training sample size = 2000

10,000 Training Observations

	Credit Cards		Census Income	
Method	Brier	AUC	Brier	AUC
Logit-BART-IS	0.138	0.751	0.125	0.854
Approx-Probit-BART-IS	0.171	0.682	0.193	0.711
Probit-BART-MCMC	**0.135**	**0.776**	**0.098**	**0.914**
Logit-BART-MCMC	0.136	0.778	0.100	0.912
Logistic Regression	0.144	0.721	0.105	0.902
RF	0.136	0.770	0.100	0.910
Holdout sample size	20,000		15,060	

Table 3.6: UCI Binary Classification Datasets, training sample size = 10,000

3.5 Example Application of General Algorithms to Treatment Effect Estimation For Binary Outcomes

Subsection 3.5.1 outlines how the methods introduced in section 3.4 can be applied to treatment effect estimation for binary outcomes. Subsection 3.5.2 outlines BMA and an alternative parameterization of Logit-BART for treatment effect estimation. Subsection 3.5.3 compares the accuracy of these methods to existing implementations using simulated data.

3.5.1 Treatment Effect Estimation with Logit-BART-BMA and Logit-BART-IS

Often a policy maker is interested not only in prediction, but in the effect of the allocation of an individual or other unit of interest to "treatment" (Kleinberg et al. 2015). The object of interest in such a scenario is the treatment effect, which is defined as the difference in potential outcomes $Y_i(1) - Y_i(0)$, where $Y_i(1)$ is the potential outcome if individual i is allocated to treatment and $Y_i(0)$ is the potential outcome if individual i is allocated to the control group (Neyman 1923, Rubin 1974). The fundamental problem of causal inference is that we do not observe the causal effect for any individual, i (Holland 1986).

The estimand of interest is the Individual Treatment Effect (ITE)

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]. \tag{3.2}$$

Whereas the ATE can be estimated by a difference in means $\bar{y}_t - \bar{y}_c$, where $\bar{y}_t$ ($\bar{y}_c$) is the mean of the outcome variable for the treated (control) group, the CATE can be thought of as a subpopulation average treatment effect.[33] [34] The CATE is identified under unconfoundedness, i.e. $Y_i(1), Y_i(0) \perp T_i|X_i$, and overlap, i.e. $0 < \Pr(T_i = 1|X_i = x) < 1 \; \forall \; x$, where T_i denotes the treatment indicator variable.

BART has been shown to be a highly effective method for treatment effect estimation (Hill 2011, Green & Kern 2012, Dorie et al. 2019, Hahn et al. 2019, Wendling et al. 2018). The standard approach to treatment effect estimation using BART is in the S-Learner framework of treatment effect meta-algorithms (Künzel et al. 2019). The treatment variable is included as a potential splitting variable in the same way as all the other covariates. Treatment effect estimates are obtained from the difference in predictions from the trained model when treatment is set to 1 and set to 0, i.e. the estimates ITE is $\hat{f}(1, x_i) - \hat{f}(0, x_i)$ where $\hat{f}$ is the prediction function obtained from an average of sum-of-tree models and the arguments are the treatment dummy variable and all other covariates, x_i. Confounding can be mitigated by including the estimated propensity score as a potential splitting variable (Hahn et al. 2017). See the second chapter of this thesis for further details.

Logit-BART-BMA and Logit-BART-IS can be applied to treatment effect estimation for binary outcomes using the usual S-Learner approach. The technical details for calculation of predictions and prediction intervals are included in appendix C.7.

[33] In instances where we condition on x being in some subset of the covariate space, i.e. $x \in \mathbb{A} \subset \mathbb{X}$, and $\tau_{\mathbb{A}} = \mathbb{E}[Y_i(1) - Y_i(0)|x \in \mathbb{A}]$, we also refer to this as the CATE (with suitably re-defined covariates).

[34] Another estimand is the average treatment effect conditional upon observed covariates $\bar{\tau} = \frac{1}{N}\sum_{i=1}^{N} \tau(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x_i]$. Imbens & Rubin (2015) refer to this as the conditional average treatment effect, but we shall use the above definition of the CATE.

3.5.2 Logit-BCF-BMA and Logit-BCF-IS

Bayesian Causal Forests (BCF) is a method for treatment effect estimation (Hahn et al. 2020). See section 3.5.1 for an overview of treatment effects and the potential outcome framework. BCF is re-parameterization of BART that allows for an independent prior to be placed on τ and also include the estimated propensity score, $\hat{\pi}_i$, as a potential splitting variable.

$$f(x_i, z_i) = \mu(x_i, \hat{\pi}_i) + \tau(x_i)z_i$$

where $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ are both sums of trees. See the second chapter of this thesis for further details. This section outlines how the general BART-BMA and BART-IS frameworks can be used to implement Bayesian Causal Forests for binary dependent variables with a logistic function of the sum-of-tree model:

$$\Pr(y_i = 1|x_i, \hat{\pi}_i, z_i) = sig(\mu(x_i, \hat{\pi}_i) + \tau(x_i)z_i)$$

where sig is a sigmoid (logistic) function. The vectors of μ and τ parameters, $\underline{\boldsymbol{\mu}}$ and $\underline{\boldsymbol{\tau}}$ have prior distributions $\underline{\boldsymbol{\mu}} \sim N(0, \frac{1}{a_\mu} I_{b_\mu})$ and $\underline{\boldsymbol{\tau}} \sim N(0, \frac{1}{a_\tau} I_{b_\tau})$ respectively. A similar formulation of BCF for binary outcomes is used by Starling et al. (2020) with a Probit link function and targeted smoothing.[35] Starling et al. (2020) estimate relative risk, however the focus here will be difference in probabilities for comparability with existing treatment effect estimation methods.

However, a limitation of this approach, relative to standard Bayesian Causal Forests for continuous outcomes, is that the treatment effect, $sig(\mu(x_i, \hat{\pi}_i) + \tau(x_i)) - sig(\mu(x_i, \hat{\pi}_i))$, depends not only on the sum-of-trees $\tau(x_i)$, but also on $\mu(x_i, \hat{\pi}_i)$, and therefore this re-parameterization does not provide a framework in which the regularization of the treatment effect estimates is wholly specified through the prior on $\tau(x_i)$. A similar issue has previously been noted by Starling et al. (2020) in the estimation of relative risk. Shrinkage of τ does not imply shrinkage to homogeneous relative risk. Starling et al. (2020) refer to heterogeneity in relative risk arising due to heterogeneity in baseline risk as *structural heterogeneity*. Therefore, ideally the scale of the priors for μ and τ should be set by careful prior elicitation (Starling et al. 2020).

For Logit-BCF-BMA and Logit-BCF-IS the algorithm for obtaining the MAP and Laplace approximations is slightly different to Logit-BART-BMA and Logit-BART-IS because the μ and τ terminal nodes are regularized by different parameters a_μ and a_τ respectively. The posterior mean and interval calculations are the same as for BART-BMA and BART-IS ITEs and the CATE (means and intervals), except $W^{tr}_{(m)}$ and $W^{c}_{(m)}$ are replaced by $[W_{(\mu,m)}W_{(\tau,m)}]$ and $[W_{(\mu,m)}\mathbf{0}]$ respectively.[36] [37] The details for the calculation of the MAP by standard quasi-Newton methods are given in Appendix C.8.

3.5.3 Application to ACIC Data Challenge

The annual Atlantic Causal Inference Conference (ACIC) has run a data analysis competition for treatment effect estimation methods. BART and BCF have performed well in this competition (Dorie et al. 2019, Hahn et al. 2019).

[35] Starling et al. (2020) implement Probit BCF using MCMC.

[36] $\mathbf{0}$ is a matrix of zeros of the same dimensions as $W_{(\tau,m)}$

[37] As in the case of BART-BMA and BART-IS, a viable alternative may be to apply a Gibbs sampler for draws from each model in the mixture.

Table 3.7 presents a comparison between BCF-MCMC,[38], BART-MCMC,[39], Causal Forests,[40] Probit-BART-MCMC,[41] Probit-BART-cause,[42] Logit-BART-IS,[43] and Logit-BCF-IS[44] applied to the publicly available data from the 2019 ACIC Data Challenge.[45] The results are restricted to the 1200 datasets in the low-dimensional category with less than 1000 observations and a binary dependent variable.[46] The RMSE and coverage are calculated using the true population ATE.

The standard Probit-BART-MCMC implementation produces the most accurate ATE estimates, however this method involved a large number of draws and was quite slow. Logit-BCF-IS produces impressive results given that each model contains only a small number of trees. The confidence intervals produced by Logit-BCF-IS are much wider than those produced by other methods. This may suggest that a larger number CATE samples is required from the mixture of 5,000 models.

	ATE		
Method	RMSE	Coverage	Length
BCF-MCMC	0.0486	0.850	0.174
BART-MCMC	0.0465	0.821	0.153
CF	0.0477	0.863	0.175
Probit-BART-MCMC	0.0427	0.879	0.169
BART-cause	0.0423	0.935	0.199
Logit-BART-IS	0.0555	0.813	0.199
Logit-BCF-IS	0.0452	0.913	0.292

Table 3.7: Results for ACIC Data Challenge low-dimensional datasets with less than 1000 observations and a binary dependent variable.

[38]BCF was implemented using the **R** package **bcf** function for continuous outcomes because currently the software does not provide options for logit or probit based implementations. The number of burn-in draws was set to 2000 and the number of post-burn-in draws aws set to 2000. Each model contained the default number of 200 μ trees and 50 τ trees. All other parameters were set t ot he default values.

[39]BART-MCMC was implemented with 100 burn-in draws and 1000 post-burn-in draws. Each model contained the default number of 200 trees. All other parameters were set to their default value.

[40]Causal forests were implemented with the **R** package **grf**. The number of trees was set to 4000.

[41]Probit-BART was implemented using the **BART** package in **R**. The number of model draws was set to the default value of 1000 post-burn-in draws with 100 burn-in draws. The number of tree in each model was set to the default number of 50, and all other parameters were set to default values.

[42]BART-cause is an alternative MCMC implementation of BART for average treatment effect estimation available in the **R** package **bartCause**. This was implemented with 4,000 post-burn-in samples, 1000 burn-in samples, and 1 separate chains. The the rest of the parameters are set to the defaults (see the **dbarts** package function **bart2** for more details), with 75 trees per model.

[43]Logit-BART-IS was implemented with 5,000 sampled models, only 5 trees per model and 10,000 CATE samples (from the mixture of sampled models) for calculation of CATE intervals.

[44]Logit-BCF-IS was implemented with 5,000 sampled models, 5 μ trees and 5 τ trees per model and 10,000 CATE samples (from the mixture of sampled models) for calculation of CATE intervals.

[45]Results are not presented for BCF-BMA or BART-BMA, because the current implementations can require a large quantity of RAM, and this can lead to errors/crashes.

[46]The current implementations of BART-IS and BCF-IS are slow when applied to datasets with many observations. The methods presented in this chapter are designed for data with a binary dependent variable. See chapter 3 of this thesis for the results for ACIC 2019 datasets with continuous outcomes.

3.6 Example of General-BART-BMA and General-BART-IS for Censored Outcome Data

3.6.1 Tobit BART-BMA and Tobit BART-IS

The example in this subsection is an average of Bayesian Tobit models, with variables described by sums-of-trees. The tree structures have the standard BART prior.[47] The terminal node parameters have a normal prior distribution and there is an inverse gamma prior on the variance of the error term.[48]

$$\tau^2 = \sigma^{-2} \sim \Gamma(\frac{\nu}{2}, \frac{\nu\lambda}{2}) \ , \ \underline{\boldsymbol{\mu}} \sim N(0, \frac{\sigma^2}{a}) \ , \ or \ \underline{\boldsymbol{\mu}} \sim N(0, \frac{1}{a}\tau^{-2})$$

and the convenient Tobin reparameterization is $(\underline{\boldsymbol{\mu}}, \tau^2) \rightarrow (\boldsymbol{\alpha} = \underline{\boldsymbol{\mu}}\tau, \tau = (\tau^2)^{\frac{1}{2}})$. This gives

$$\alpha = \tau\beta \sim N(0, \frac{1}{a}I)$$

Let the covariate matrix, W be the set of binary variables indicating inclusion of observations in terminal nodes. The standard Tobit model framework is

$$y_i^* = row_i(W)\underline{\boldsymbol{\mu}} + \varepsilon_i \ , \ \varepsilon \sim i.i.d. \ N(0, \tau^{-2})$$

$$y_i = \max\{y_i^*, 0\} \ , \ i = 1, ..., n$$

See appendix C.9 for details on how to implement the Tobit model using standard Laplace approximations. Chib (1992) outlines a number of approaches for implementation of Bayesian Tobit models, including Laplace approximations (fully exponential Laplace approximations, as outlined by Tierney & Kadane (1986)) and a Gibbs sampler.

An average of Tobit sum-of-tree models is obtainable by application of the general BART-BMA or BART-IS algorithms outlined in section 3.3 in combination with one of a number of potenrial Tobit approximation methods, including:

1. Use a Laplace approximation for the marginal likelihood and posterior distributions.
2. Use a Laplace approximation for the marginal likelihood, and then apply a Gibbs sampler for each model in the mixture. [49]

A sum-of-tree Tobit model, based on gradient boosting, is used by Sigrist (2018) to predict defaults on loans made to Small and Medium Sized enterprises. Gradient-boosted Tobit outperforms Logit, Tobit, and a number of machine learning methods. BART can be viewed as a Bayesian alternative to gradient boosted trees as it involves sum-of-tree models. Therefore it is desirable to investigate the performance of a Tobit-BART implementation at predicting censored outcomes.

The example below is based on the simulations described by Sigrist (2018). The goal is prediction of censored outcomes out-of-sample. As in Sigrist (2018), the competing methods are be Tobit and binary

[47] I have provided options in the **safeBart** package for Tobit-BART-IS with draws from the Quadrianto & Ghahramani (2014) prior and the spike and tree prior (Rockova & van der Pas 2017).

[48] Chib (1992) used an uninformative prior for Bayesian Tobit. The normal prior is preferred here for the terminal nodes because this is the prior used by standard BART, and it is desirable to regularize the terminal node parameters.

[49] Chib (1992) describes a Gibbs sampler for Tobit. Perhaps an alternative based on the sampler of Polson et al. (2013) is applicable.

classification methods logistic regression, Logit-BART-IS,[50] Probit-BART-MCMC, Logit-BART-MCMC,[51] and Random Forests.[52] The performance measures are the Brier Score and Area Under the Curve (AUC) for out of sample predictive probabilities of censored outcomes.

There are 30 uniformly distributed covariates, $X_1, ..., X_{30} \sim Unif(-1, 1)$, the latent outcome Y^*, and observed outcome Y are determined by the following data generating process:

$$Y^* = \sum_{k=1}^{f} 0.3(X_k)_+ + \sum_{k=1}^{3} \sum_{j=k+1}^{4} (X_k X_j)_+ + \varepsilon \ , \ \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$Y = min(2.84, Y^*)$$

where $(x)_+ = max(x, 0)$, and $\sigma_\varepsilon = 0.7$.[53] The results are presented in Table 3.8. While Tobit-BART-IS outperforms the other methods, the results are somewhat underwhelming. One possible explanation for this is that the outcome should be transformed and the prior on the terminal node parameters should be carefully calibrated so that coefficients are regularized towards zero or predictive probabilities are regularized towards the training sample proportion of censored outcomes.[54]

Method	Brier	AUC
Tobit-BART-IS	**0.046**	**0.776**
Tobit	0.049	0.613
Logit-BART-IS	0.047	0.740
Probit-BART-MCMC	**0.046**	0.750
Logit-BART-MCMC	0.047	0.721
Logistic Regression	0.051	0.610
RF	0.047	0.758

Table 3.8: Results for Tobit-BART-IS simulation study.

3.7 Conclusion

3.7.1 Summary

This chapter outlines a generalization of BART to a wide range of model settings. This approach builds on the algorithms introduced in chapter 2 and existing methods for approximate inference and calculation of model evidence. As an example, the approach is applied to the implementation of Logit-BART. The approach is validated by the fact that Logit-BART-IS and Logit-BART-BMA produce similar results to existing MCMC implementations of Probit-BART and Logit-BART.

Depending on computational resources and the speed of approximate inference methods such as Laplace approximations, the new methods may provide fast alternatives to MCMC-based approaches. The general BART-IS algorithm is highly parallelizable.

[50] Logit-BART-IS and Tobit-BART-IS were implemented with 20,000 draws and 5 trees per model.

[51] Probit-BART-MCMC and Logit-BART-MCMC were implemented using the **BART** package in **R** using 5000 burn-in draws and 10,000 post burn-in draws.

[52] Random Forests was implemented using the **ranger** package in **R** with 10,000 trees.

[53] The upper bound and standard deviation are those chosen by Sigrist (2018). Unlike the simulations presented by Sigrist (2018), the simulations presented here do not involve data censoring being determined by a latent variable that has a different error term to the error term for the observed outcome.

[54] One possibility would be to add an intercept to the sum-of-tree model, demean the outcome, and set a, ν and λ such that the prior predicts observations to lie in the training data range with high probability.

The methods outlined in this chapter have some limitations. There is no guarantee that the BART-BMA search algorithm will be particularly effective in searching the model space, and in some Logit-BART-BMA examples the current implementation is prohibitively slow if model search parameters are not appropriately adjusted. The BART-IS model sampler only takes a small sample from the large model space, and does not adapt to find models with higher posterior probability.[55] Therefore it is possible that none of the sampled models are similar to the "true" model.

Despite the potential limitations, the algorithms described in this chapter, particularly BART-IS, are of practical use to researchers seeking a quick and dirty approach to implementation of generalizations of BART. Simple BART-IS implementations can provide useful benchmarks for testing the accuracy of new MCMC-based implementations of similar BART model frameworks.

3.7.2 Future Research: Multinomial Logit, Poisson Regression, and Other Generalizations

Poisson regression and multinomial logit can be implemented with standard Laplace approximations (Madigan et al. 2005, Cawley et al. 2007, Silverman et al. 2019). Integrated Nested Laplace Approximations (Rue et al. 2009) can be used to implement a range of models including multinomial logit[56] and Poisson regression (and allows for hierarchical priors, e.g. mixed logit) and provides accurate calculations of the marginal likelihood. The general framework introduced in section 3.3 can therefore be extended to a wide variety of settings.[57]

However, a key requirement is that the calculation of the marginal likelihood and posterior inference are computationally efficient. The speed of the BART-IS or BART-BMA based implementations will depend on the choice of methods. The construction of residuals for the changepoint detection algorithm in BART-BMA is not straightforward for all model settings, and the arbitrary model search algorithm is not guaranteed to perform well outside of the linear regression context for which it was originally designed. Therefore the BART-IS framework is more generalizable and is recommended above the BART-BMA framework for application of BART to a wider class of models.

A possible approach for multinomial logit BART is to use Integrated Nested Laplace Approximations to calculate the marginal likelihoods, and then use a variation of the sampler described by Polson et al. (2013). [58]

[55] However, as noted in chapter 2, a Bayesian Adaptive Sampling approach to sampling of BART models is an interesting topic for future research (Clyde et al. 2011). However, it is not obvious how to proceed in constructing such a sampler.

[56] Multinomial logit is implementable using the multinomial-Poisson transform (Baker 1994).

[57] See Barber et al. (2016) for some general asymptotic results on the use of the marginal likelihood for model selection.

[58] Linderman et al. (2015) describe this sampler for multinomial logit within a larger model. Similarly multinomial-Logit-BCF-BMA and multinomial-Logit-BCF-IS are possible extensions and these methods would produce estimates of treatment effects on probabilities of categories.

Bibliography

Abu-Nimeh, S., Nappa, D., Wang, X. & Nair, S. (2008), Bayesian additive regression trees-based spam detection for enhanced email privacy, *in* '2008 Third International Conference on Availability, Reliability and Security', IEEE, pp. 1044–1051.

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., Stuart, A. et al. (2017), 'Importance sampling: Intrinsic dimension and computational cost', *Statistical Science* **32**(3), 405–431.

Agarwal, R., Ranjan, P. & Chipman, H. (2014), 'A new bayesian ensemble of trees approach for land cover classification of satellite imagery', *Canadian Journal of Remote Sensing* **39**(6), 507–520.

Alaa, A. M. & van der Schaar, M. (2018), 'Bayesian nonparametric causal inference: Information rates and learning algorithms', *IEEE Journal of Selected Topics in Signal Processing* **12**(5), 1031–1046.

Alaa, A. & Van Der Schaar, M. (2019), Validating causal inference models via influence functions, *in* 'International Conference on Machine Learning', pp. 191–201.

Albert, J. H. & Chib, S. (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American statistical Association* **88**(422), 669–679.

Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. (2010), 'Permutation importance: a corrected feature importance measure', *Bioinformatics* **26**(10), 1340–1347.

Arlot, S. & Genuer, R. (2014), 'Analysis of purely random forests bias', *arXiv preprint arXiv:1407.3939* .

Atan, O., Jordon, J. & van der Schaar, M. (2018), Deep-treat: Learning optimal personalized treatments from observational data using neural networks, *in* 'Thirty-Second AAAI Conference on Artificial Intelligence'.

Athey, S. (2018), The impact of machine learning on economics, *in* 'The economics of artificial intelligence: An agenda', University of Chicago Press.

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360.

Athey, S. & Imbens, G. W. (2015), 'Machine learning methods for estimating heterogeneous causal effects', *stat* **1050**(5).

Athey, S. & Imbens, G. W. (2017), 'The econometrics of randomized experiments', *Handbook of Economic Field Experiments* .

Athey, S., Tibshirani, J., Wager, S. et al. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178.

Bacher, A. (2016), 'A new bijection on m-dyck paths with application to random sampling', *arXiv preprint arXiv:1603.06290* .

Baker, S. G. (1994), 'The multinomial-poisson transformation', *Journal of the Royal Statistical Society: Series D (The Statistician)* **43**(4), 495–504.

Balandat, M. (2016), 'New tools for econometric analysis of high-frequency time series data-application to demand-side management in electricity markets'.

Barber, R. F., Drton, M. & Tan, K. M. (2016), Laplace approximation in high-dimensional bayesian regression, *in* 'Statistical Analysis for High-Dimensional Data', Springer, pp. 15–36.

Bargagli-Stoffi, F. J., De-Witte, K. & Gnecco, G. (2019), 'Heterogeneous causal effects with imperfect compliance: a novel bayesian machine learning approach', *arXiv preprint arXiv:1905.12707* .

Barro, R. J. (1996*a*), 'Democracy and growth', *Journal of economic growth* **1**(1), 1–27.

Barro, R. J. (1996*b*), Determinants of economic growth: A cross-country empirical study, Technical report, National Bureau of Economic Research.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**(2), 608–650.

Bertrand, M., Crépon, B., Marguerie, A. & Premand, P. (2017), 'Contemporaneous and post-program impacts of a public works program: Evidence from côte d'ivoire'.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, springer.

Bivand, R., Gómez-Rubio, V., Rue, H. et al. (2015), 'Spatial data analysis with r-inla with some extensions', *Journal of Statistical Software* **63**(i20).

Bivand, R. S., Gómez-Rubio, V. & Rue, H. (2014), 'Approximate bayesian inference for spatial econometrics models', *Spatial Statistics* **9**, 146–165.

Bleich, J., Kapelner, A., George, E. I. & Jensen, S. T. (2014), 'Variable selection for bart: An application to gene regulation', *The Annals of Applied Statistics* pp. 1750–1781.

Boatman, J. A., Vock, D. M. & Koopmeiners, J. S. (2020), 'Borrowing from supplemental sources to estimate causal effects from a primary data source', *arXiv preprint arXiv:2003.09680* .

Bollinger, B. & Hartmann, W. R. (2015), Welfare effects of home automation technology with dynamic pricing, Technical report.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D. & Do, K.-A. (2010), 'Bayesian ensemble methods for survival prediction in gene expression data', *Bioinformatics* **27**(3), 359–367.

Breiman, L. (1997), Arcing the edge, Technical report, Technical Report 486, Statistics Department, University of California at Berkeley.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Brock, W. A. & Durlauf, S. N. (2001), 'What have we learned from a decade of empirical research on growth? growth empirics and reality', *The World Bank Economic Review* **15**(2), 229–272.

Buntine, W. (1992), 'Learning classification trees', *Statistics and computing* **2**(2), 63–73.

Carnegie, N., Harada, M., Dorie, V. & Hill, J. (2015), 'treatsens: sensitivity analysis for causal inference', *R package* .

Carvalho, C., Feller, A., Murray, J., Woody, S. & Yeager, D. (2019), 'Assessing treatment effect variation in observational studies: Results from a data challenge', *arXiv preprint arXiv:1907.07592* .

Castillo, I. & Rockova, V. (2019), 'Multiscale analysis of bayesian cart', *arXiv preprint arXiv:1910.07635* .

Cawley, G. C., Talbot, N. L. & Girolami, M. (2007), Sparse multinomial logistic regression via bayesian l1 regularisation, *in* 'Advances in neural information processing systems', pp. 209–216.

CER (2011), Electricity smart metering customer behaviour trials (cbt) findings report, Technical report, Commision for Energy Regulation.

Chakraborty, S. (2016), Bayesian additive regression tree for seemingly unrelated regression with automatic tree selection, *in* 'Handbook of Statistics', Vol. 35, Elsevier, pp. 229–251.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. et al. (2017), Double/debiased machine learning for treatment and causal parameters, Technical report.

Chernozhukov, V., Demirer, M., Duflo, E. & Fernandez-Val, I. (2017), 'Generic machine learning inference on heterogenous treatment effects in randomized experiments', *arXiv preprint arXiv:1712.04802* .

Chib, S. (1992), 'Bayes inference in the tobit censored regression model', *Journal of Econometrics* **51**(1-2), 79–99.

Chipman, H. A., George, E. I. & McCulloch, R. E. (1998), 'Bayesian cart model search', *Journal of the American Statistical Association* **93**(443), 935–948.

Chipman, H. A., George, E. I., McCulloch, R. E. et al. (2010), 'Bart: Bayesian additive regression trees', *The Annals of Applied Statistics* **4**(1), 266–298.

Choi, H. M., Hobert, J. P. et al. (2013), 'The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic', *Electronic Journal of Statistics* **7**, 2054–2064.

Chopin, N., Ridgway, J. et al. (2017), 'Leave pima indians alone: binary regression as a benchmark for bayesian computation', *Statistical Science* **32**(1), 64–87.

Clyde, M. A. & Ghosh, J. (2012), 'Finite population estimators in stochastic search variable selection', *Biometrika* **99**(4), 981–988.

Clyde, M. A., Ghosh, J. & Littman, M. L. (2011), 'Bayesian adaptive sampling for variable selection and model averaging', *Journal of Computational and Graphical Statistics* **20**(1), 80–101.

Clyde, M., Desimone, H. & Parmigiani, G. (1996), 'Prediction via orthogonalized model mixing', *Journal of the American Statistical Association* **91**(435), 1197–1208.

CSE (2012), "beyond average consumption" - development of a framework for assessing impact of policy proposals on different consumer groups, Final report to ofgem, Centre for Sustainable Energy.

Cutler, A. & Zhao, G. (2001), 'Pert-perfect random tree ensembles', *Computing Science and Statistics* **33**, 490–497.

Davis, J. & Heller, S. B. (2017*a*), Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs, Technical report, National Bureau of Economic Research.

Davis, J. M. & Heller, S. B. (2017*b*), 'Using causal forests to predict treatment heterogeneity: An application to summer jobs', *American Economic Review* **107**(5), 546–550.

DECC (2012), Demand side response in the domestic sector - a literature review of major trials, Technical report, Frontier Economics and Sustainability First, London.

Demiriz, A., Bennett, K. P. & Shawe-Taylor, J. (2002), 'Linear programming boosting via column generation', *Machine Learning* **46**(1-3), 225–254.

Deshpande, S. K., Bai, R., Balocchi, C. & Starling, J. E. (2020), 'Vc-bart: Bayesian trees for varying coefficients', *arXiv preprint arXiv:2003.06416* .

Di Cosmo, V., Lyons, S. & Nolan, A. (2014), 'Estimating the impact of time-of-use pricing on irish electricity demand', *The Energy Journal* **35**(3).

Di Cosmo, V. & O'Hora, D. (2017), 'Nudging electricity consumption using tou pricing and feedback: evidence from irish households', *Journal of Economic Psychology* .

Dobra, A., Eicher, T. S. & Lenkoski, A. (2010), 'Modeling uncertainty in macroeconomic growth determinants using gaussian graphical models', *Statistical Methodology* **7**(3), 292–306.

Doppelhofer, G., Hansen, O.-P. & Weeks, M. (2016), 'Determinants of long-term economic growth redux: A measurement error model averaging (mema) approach', *NHH Dept. of Economics Discussion Paper* (19).

Doppelhofer, G. & Weeks, M. (2009), 'Jointness of growth determinants', *Journal of Applied Econometrics* **24**(2), 209–244.

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D. et al. (2019), 'Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition', *Statistical Science* **34**(1), 43–68.

Du, J. & Linero, A. (2019), Incorporating grouping information into bayesian decision tree ensembles, *in* 'International Conference on Machine Learning', pp. 1686–1695.

Du, J. & Linero, A. R. (2018), 'Interaction detection with bayesian decision tree ensembles', *arXiv preprint arXiv:1809.08524* .

Dua, D. & Graff, C. (2017), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

Durlauf, S. N., Kourtellos, A. & Tan, C. M. (2012), 'Is god in the details? a reexamination of the role of religion in economic growth', *Journal of Applied Econometrics* **27**(7), 1059–1075.

Eicher, T. S., Papageorgiou, C. & Raftery, A. E. (2011), 'Default priors and predictive performance in bayesian model averaging, with application to growth determinants', *Journal of Applied Econometrics* **26**(1), 30–55.

Entezari, R., Craiu, R. V. & Rosenthal, J. S. (2018), 'Likelihood inflating sampling algorithm', *Canadian Journal of Statistics* **46**(1), 147–175.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Farrell, M. H., Liang, T. & Misra, S. (2018), 'Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands', *arXiv preprint arXiv:1809.09953* .

Faruqui, A., Sergici, S. & Palmer, J. (2010), 'The impact of dynamic pricing on low income customers', *Institute for Electric Efficiency Whitepaper* .

Fernandez, C., Ley, E. & Steel, M. F. (2001*a*), 'Benchmark priors for bayesian model averaging', *Journal of Econometrics* **100**(2), 381–427.

Fernandez, C., Ley, E. & Steel, M. F. (2001*b*), 'Model uncertainty in cross-country growth regressions', *Journal of applied Econometrics* **16**(5), 563–576.

Freund, Y. & Schapire, R. E. (1995), A desicion-theoretic generalization of on-line learning and an application to boosting, *in* 'European conference on computational learning theory', Springer, pp. 23–37.

Freund, Y., Schapire, R. E. et al. (1996), Experiments with a new boosting algorithm, Citeseer.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* pp. 1189–1232.

Friedman, J. H. & Meulman, J. J. (2003), 'Multiple additive regression trees with application in epidemiology', *Statistics in medicine* **22**(9), 1365–1381.

Friedman, J. H., Popescu, B. E. et al. (2003), 'Importance sampled learning ensembles', *Journal of Machine Learning Research* **94305**, 1–32.

Friedman, J. H., Popescu, B. E. et al. (2008), 'Predictive learning via rule ensembles', *The Annals of Applied Statistics* **2**(3), 916–954.

Friedman, J. H. et al. (1991), 'Multivariate adaptive regression splines', *The annals of statistics* **19**(1), 1–67.

Friedman, J., Hastie, T. & Tibshirani, R. (2009), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2000), 'Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)', *The annals of statistics* **28**(2), 337–407.

Friel, N. & Wyse, J. (2012), 'Estimating the evidence–a review', *Statistica Neerlandica* **66**(3), 288–308.

Garcia, M. G., Medeiros, M. C. & Vasconcelos, G. F. (2017), 'Real-time inflation forecasting with high-dimensional models: The case of brazil', *International Journal of Forecasting* **33**(3), 679–693.

George, E. I. & McCulloch, R. E. (1995), 'Stochastic search variable selection', *Markov chain Monte Carlo in practice* **68**, 203–214.

George, E., Laud, P., Logan, B., McCulloch, R. & Sparapani, R. (2018), 'Fully nonparametric bayesian additive regression trees', *arXiv preprint arXiv:1807.00068* .

Geurts, P., Ernst, D. & Wehenkel, L. (2006), 'Extremely randomized trees', *Machine learning* **63**(1), 3–42.

Gómez-Rubio, V., Bivand, R. S. & Rue, H. (2020), 'Bayesian model averaging with the integrated nested laplace approximation', *Econometrics* **8**(2), 23.

Gómez-Rubio, V. & Rue, H. (2018), 'Markov chain monte carlo with the integrated nested laplace approximation', *Statistics and Computing* **28**(5), 1033–1051.

Gramacy, R. B., Polson, N. G. et al. (2012), 'Simulation-based regularized logistic regression', *Bayesian Analysis* **7**(3), 567–590.

Green, D. P. & Kern, H. L. (2012), 'Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees', *Public opinion quarterly* **76**(3), 491–511.

Grünwald, P. (2012), The safe bayesian, *in* 'International Conference on Algorithmic Learning Theory', Springer, pp. 169–183.

Hahn, P. R., Carvalho, C. M., Puelz, D., He, J. et al. (2018), 'Regularization and confounding in linear regression for treatment effect estimation', *Bayesian Analysis* **13**(1), 163–182.

Hahn, P. R., Dorie, V. & Murray, J. S. (2019), 'Atlantic causal inference conference (acic) data analysis challenge 2017', *arXiv preprint arXiv:1905.09515* .

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2017), 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects'.

Hahn, P. R., Murray, J. S., Carvalho, C. M. et al. (2020), 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects', *Bayesian Analysis* .

Harding, M. & Lamarche, C. (2016), 'Empowering consumers through data and smart technology: Experimental evidence on the consequences of time-of-use electricity pricing policies', *Journal of Policy Analysis and Management* **35**(4), 906–931.

He, J., Yalov, S. & Hahn, P. R. (2018), 'Accelerated bayesian additive regression trees', *arXiv preprint arXiv:1810.02215* .

He, J., Yalov, S., Murray, J. & Hahn, P. R. (2019), Stochastic tree ensembles for regularized supervised learning, Technical report, Technical report.

Heaton, M. J. & Scott, J. G. (2010), 'Bayesian computation and the linear model', *Frontiers of Statistical Decision Making and Bayesian Analysis* pp. 527–545.

Henderson, N. C., Louis, T. A., Rosner, G. L. & Varadhan, R. (2017), 'Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models', *arXiv preprint arXiv:1706.06611* .

Hernández, B., Raftery, A. E., Pennington, S. R. & Parnell, A. C. (2018), 'Bayesian additive regression trees using bayesian model averaging', *Statistics and Computing* **28**(4), 869–890.

Hill, J. L. (2011), 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

Hill, J., Linero, A. & Murray, J. (2020), 'Bayesian additive regression trees: A review and look forward', *Annual Review of Statistics and Its Application* **7**.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial', *Statistical science* pp. 382–401.

Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American statistical Association* **81**(396), 945–960.

Huber, F. & Rossini, L. (2020), 'Inference in bayesian additive vector autoregressive tree models', *arXiv preprint arXiv:2006.16333* .

Imai, K., Ratkovic, M. et al. (2013), 'Estimating treatment effect heterogeneity in randomized program evaluation', *The Annals of Applied Statistics* **7**(1), 443–470.

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Karl, A. & Lenkoski, A. (2012), 'Instrumental variable bayesian model averaging via conditional bayes factors', *arXiv preprint arXiv:1202.5846* .

Killick, R., Fearnhead, P. & Eckley, I. A. (2012), 'Optimal detection of changepoints with a linear computational cost', *Journal of the American Statistical Association* **107**(500), 1590–1598.

Kindo, B. P., Wang, H., Hanson, T. & Peña, E. A. (2016), 'Bayesian quantile additive regression trees', *arXiv preprint arXiv:1607.02676* .

Kindo, B. P., Wang, H. & Peña, E. A. (2016), 'Multinomial probit bayesian additive regression trees', *Stat* **5**(1), 119–131.

Kindo, B., Wang, H. & Pena, E. (2013), 'Mbact-multiclass bayesian additive classification trees', *stat* .

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review* **105**(5), 491–95.

Knaus, M., Lechner, M. & Strittmatter, A. (2018), 'Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence'.

Kohavi, R. (1996), Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid., *in* 'Kdd', Vol. 96, pp. 202–207.

Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. (2019), 'Metalearners for estimating heterogeneous treatment effects using machine learning', *Proceedings of the national academy of sciences* **116**(10), 4156–4165.

Kwok, S. W. & Carter, C. (1990), Multiple decision trees, *in* 'Machine Intelligence and Pattern Recognition', Vol. 9, Elsevier, pp. 327–335.

Lakshminarayanan, B., Roy, D. & Teh, Y. W. (2013), Top-down particle filtering for bayesian decision trees, *in* 'International Conference on Machine Learning', pp. 280–288.

Lakshminarayanan, B., Roy, D. & Teh, Y. W. (2015), Particle gibbs for bayesian additive regression trees, *in* 'Artificial Intelligence and Statistics', pp. 553–561.

Lechner, M. (2019), 'Modified causal forests for estimating heterogeneous causal effects'.

Lenkoski, A., Eicher, T. S. & Raftery, A. E. (2014), 'Two-stage bayesian model averaging in endogenous variable models', *Econometric reviews* **33**(1-4), 122–151.

Leon-Gonzalez, R. & Montolio, D. (2015), 'Endogeneity and panel data in growth regressions: A bayesian model averaging approach', *Journal of Macroeconomics* **46**, 23–39.

Levine, R. & Renelt, D. (1992), 'A sensitivity analysis of cross-country growth regressions', *The American economic review* pp. 942–963.

Linderman, S., Johnson, M. J. & Adams, R. P. (2015), Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation, *in* 'Advances in Neural Information Processing Systems', pp. 3456–3464.

Linero, A. R. (2017), 'A review of tree-based bayesian methods', *Communications for Statistical Applications and Methods* **24**(6).

Linero, A. R. (2018), 'Bayesian regression trees for high-dimensional prediction and variable selection', *Journal of the American Statistical Association* pp. 1–11.

Linero, A. R., Sinha, D. & Lipsitz, S. R. (2019), 'Semiparametric mixed-scale models using shared bayesian forests', *Biometrics* .

Linero, A. R. & Yang, Y. (2017), 'Bayesian regression tree ensembles that adapt to smoothness and sparsity', *arXiv preprint arXiv:1707.09461* .

Liu, Y. & Rockova, V. (2020), 'Variable selection via thompson sampling', *arXiv preprint arXiv:2007.00187* .

Liu, Y., Rocková, V. & Wang, Y. (2018), 'Abc variable selection with bayesian forests', *arXiv preprint arXiv:1806.02304* .

Logan, B. R., Sparapani, R., McCulloch, R. E. & Laud, P. W. (2019), 'Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees', *Statistical methods in medical research* **28**(4), 1079–1093.

Lu, M., Sadiq, S., Feaster, D. J. & Ishwaran, H. (2018), 'Estimating individual treatment effect in observational data using random forest methods', *Journal of Computational and Graphical Statistics* **27**(1), 209–219.

Madigan, D., Genkin, A., Lewis, D. D. & Fradkin, D. (2005), Bayesian multinomial logistic regression for author identification, *in* 'AIP conference proceedings', Vol. 803, American Institute of Physics, pp. 509–516.

Madigan, D. & Raftery, A. E. (1994), 'Model selection and accounting for model uncertainty in graphical models using occam's window', *Journal of the American Statistical Association* **89**(428), 1535–1546.

Madigan, D., York, J. & Allard, D. (1995), 'Bayesian graphical models for discrete data', *International Statistical Review/Revue Internationale de Statistique* pp. 215–232.

Minka, T. (2000), Bayesian linear regression, Technical report, Citeseer.

Moral-Benito, E. (2016), 'Growth empirics in panel data under model uncertainty and weak exogeneity', *Journal of Applied Econometrics* **31**(3), 584–602.

Moran, G. E., Ročková, V., George, E. I. et al. (2018), 'Variance prior forms for high-dimensional bayesian variable selection', *Bayesian Analysis* pp. 1091–1119.

Moro, S., Cortez, P. & Rita, P. (2014), 'A data-driven approach to predict the success of bank telemarketing', *Decision Support Systems* **62**, 22–31.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT press.

Murray, J. S. (2017), 'Log-linear bayesian additive regression trees for categorical and count responses', *arXiv preprint arXiv:1701.01503* .

Nembrini, S. (2019), 'On what to permute in test-based approaches for variable importance measures in random forests', *Bioinformatics* **35**(15), 2701–2705.

Neyman, J. (1923), 'Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted)', *Stat Sci* **5**, 463–472.

Nie, X. & Wager, S. (2017*a*), 'Learning objectives for treatment effect estimation', *arXiv preprint arXiv:1712.04912* .

Nie, X. & Wager, S. (2017*b*), 'Quasi-oracle estimation of heterogeneous treatment effects', *arXiv preprint arXiv:1712.04912* .

Ofgem (2013), 'Consumer vulnerability strategy'.
URL: *https://www.ofgem.gov.uk/ofgem-publications/75550/consumer-vulnerability-strategy.pdf*

Oprescu, M., Syrgkanis, V. & Wu, Z. S. (2018), 'Orthogonal random forest for causal inference', *arXiv preprint arXiv:1806.03467* .

Polson, N. G., Scott, J. G. & Windle, J. (2013), 'Bayesian inference for logistic models using pólya–gamma latent variables', *Journal of the American statistical Association* **108**(504), 1339–1349.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T. & Tibshirani, R. (2017), 'Some methods for heterogeneous treatment effect estimation in high-dimensions', *arXiv preprint arXiv:1707.00102* .

Pratola, M., Chipman, H., George, E. & McCulloch, R. (2017), 'Heteroscedastic bart using multiplicative regression trees', *arXiv preprint arXiv:1709.07542* .

Pratola, M. T., Chipman, H. A., Gattiker, J. R., Higdon, D. M., McCulloch, R. & Rust, W. N. (2014), 'Parallel bayesian additive regression trees', *Journal of Computational and Graphical Statistics* **23**(3), 830–852.

Pratola, M. T. et al. (2016), 'Efficient metropolis–hastings proposal mechanisms for bayesian regression tree models', *Bayesian analysis* **11**(3), 885–911.

Prest, B. C. (2017), Peaking interest: How awareness drives the effectiveness of time-of-use electricity pricing, *in* 'Riding the Energy Cycles, 35th USAEE/IAEE North American Conference, Nov 12-15, 2017', International Association for Energy Economics.

Preston, I., White, V. & Sturtevant, E. (2013), 'The hardest hit: Going beyond the mean', *a Centre for Sustainable Energy report for Consumer Futures. Available here: http://www. consumerfutures. org. uk/files/2013/05/The-hardest-hit. pdf* .

Quadrianto, N. & Ghahramani, Z. (2014), 'A very simple safe-bayesian random forest', *IEEE transactions on pattern analysis and machine intelligence* **37**(6), 1297–1303.

Quinlan, J. R. (1987), 'Simplifying decision trees', *International journal of man-machine studies* **27**(3), 221–234.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**(437), 179–191.

Redell, N. (2020), *forecastML: Time Series Forecasting with Machine Learning Methods*. R package version 0.9.0.
URL: *https://CRAN.R-project.org/package=forecastML*

Rocková, V. & Saha, E. (2018), 'On theory for bart', *arXiv preprint arXiv:1810.00787* .

Rockova, V. & van der Pas, S. (2017), 'Posterior concentration for bayesian regression trees and their ensembles', *arXiv preprint arXiv:1708.08734* .

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations', *Journal of the royal statistical society: Series b (statistical methodology)* **71**(2), 319–392.

Saito, Y. & Yasui, S. (2019), 'Counterfactual cross-validation: Effective causal model selection from observational data', *arXiv preprint arXiv:1909.05299* .

Sakar, C. O., Polat, S. O., Katircioglu, M. & Kastro, Y. (2019), 'Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks', *Neural Computing and Applications* **31**(10), 6893–6908.

Sala-i Martin, X., Doppelhofer, G. & Miller, R. I. (2004), 'Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach', *American economic review* pp. 813–835.

Sala-i Martin, X. X. (1997), I just ran four million regressions, Technical report, National Bureau of Economic Research.

Santos, P. H. F. d. & Lopes, H. F. (2018), 'Tree-based bayesian treatment effect analysis', *arXiv preprint arXiv:1808.09507* .

Schofield, J., Carmichael, R., amd M. Woolf, S. T., Bilton, M. & Strbac, G. (2014), Residential consumer responsiveness to time-varying pricing, Report a3 for the low carbon london lcnf project, Imperial College London.

Schuler, A., Baiocchi, M., Tibshirani, R. & Shah, N. (2018), 'A comparison of methods for model selection when estimating individual treatment effects', *arXiv preprint arXiv:1804.05146* .

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. & McCulloch, R. E. (2016), 'Bayes and big data: The consensus monte carlo algorithm', *International Journal of Management Science and Engineering Management* **11**(2), 78–88.

Shalit, U., Johansson, F. D. & Sontag, D. (2017), Estimating individual treatment effect: generalization bounds and algorithms, *in* 'Proceedings of the 34th International Conference on Machine Learning-Volume 70', JMLR. org, pp. 3076–3085.

Shi, C., Blei, D. M. & Veitch, V. (2019), 'Adapting neural networks for the estimation of treatment effects', *arXiv preprint arXiv:1906.02120* .

Sidebotham, L. (2015), 'Customer-led network revolution project closedown report', *Customer Led Network Revolution Project Closedown Report. Northern Powergrid, Newcastle upon Tyne* .

Sigrist, F. (2018), 'Gradient and newton boosting for classification and regression', *arXiv preprint arXiv:1808.03064* .

Silverman, J. D., Roche, K., Holmes, Z. C., David, L. A. & Mukherjee, S. (2019), 'Bayesian multinomial logistic normal models through marginally latent matrix-t processes', *arXiv preprint arXiv:1903.11695* .

Sparapani, R. A., Logan, B. R., McCulloch, R. E. & Laud, P. W. (2016), 'Nonparametric survival analysis using bayesian additive regression trees (bart)', *Statistics in medicine* **35**(16), 2741–2753.

Sparapani, R. A., Rein, L. E., Tarima, S. S., Jackson, T. A. & Meurer, J. R. (2018), 'Non-parametric recurrent events analysis with bart and an application to the hospital admissions of patients with diabetes', *Biostatistics* .

Sparapani, R., Logan, B. R., McCulloch, R. E. & Laud, P. W. (2019), 'Nonparametric competing risks analysis using bayesian additive regression trees', *Statistical methods in medical research* p. 0962280218822140.

Spiegelhalter, D. J. & Lauritzen, S. L. (1990), 'Sequential updating of conditional probabilities on directed graphical structures', *Networks* **20**(5), 579–605.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K. & Scott, J. G. (2018), 'Bart with targeted smoothing: An analysis of patient-specific stillbirth risk', *arXiv preprint arXiv:1805.07656* .

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R., Carvalho, C. M. & Scott, J. G. (2020), 'Targeted smooth bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation', *arXiv preprint arXiv:1905.09405* .

Steel, M. F. (2017), 'Model averaging and its use in economics', *arXiv preprint arXiv:1709.08221* .

Stewart, L. (1987), 'Hierarchical bayesian analysis using monte carlo integration: computing posterior distributions when there are many possible models', *Journal of the Royal Statistical Society: Series D (The Statistician)* **36**(2-3), 211–219.

Stock, J. H. & Watson, M. W. (2002), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business & Economic Statistics* **20**(2), 147–162.

Strobl, C. (2008), *Statistical issues in machine learning: Towards reliable split selection and variable importance measures*, Cuvillier Verlag.

Taddy, M., Chen, C.-S., Yu, J. & Wyle, M. (2015), 'Bayesian and empirical bayesian forests', *arXiv preprint arXiv:1502.02312* .

Tan, Y. V., Flannagan, C. A. & Elliott, M. R. (2016), 'Predicting human-driving behavior to help driverless vehicles drive: random intercept bayesian additive regression trees', *arXiv preprint arXiv:1609.07464* .

Tan, Y. V., Flannagan, C. A. & Elliott, M. R. (2018), '" robust-squared" imputation models using bart', *arXiv preprint arXiv:1801.03147* .

Tan, Y. V. & Roy, J. (2019), 'Bayesian additive regression trees and the general bart model', *arXiv preprint arXiv:1901.07504* .

Tian, L., Alizadeh, A. A., Gentles, A. J. & Tibshirani, R. (2014), 'A simple method for estimating interactions between a treatment and a large number of covariates', *Journal of the American Statistical Association* **109**(508), 1517–1532.

Tierney, L. & Kadane, J. B. (1986), 'Accurate approximations for posterior moments and marginal densities', *Journal of the american statistical association* **81**(393), 82–86.

Van Der Putten, P. & van Someren, M. (2000), Coil challenge 2000: The insurance company case, Technical report, Technical Report 2000–09, Leiden Institute of Advanced Computer Science ....

Volinsky, C. T., Madigan, D., Raftery, A. E. & Kronmal, R. A. (1997), 'Bayesian model averaging in proportional hazard models: assessing the risk of a stroke', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(4), 433–448.

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

Wang, G.-w., Zhang, C.-x. & Yin, Q.-y. (2019), 'Rs-bart: a novel technique to boost the prediction ability of bayesian additive regression trees', *Chinese Journal of Engineering Mathematics* **36**(4), 461.

Weisberg, H. I. & Pontes, V. P. (2015), 'Post hoc subgroups in clinical trials: Anathema or analytics?', *Clinical trials* **12**(4), 357–364.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. & Gallego, B. (2018), 'Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases', *Statistics in medicine* **37**(23), 3309–3324.

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

Yang, Y., Cheng, G. & Dunson, D. B. (2015), 'Semiparametric bernstein-von mises theorem: Second order studies', *arXiv preprint arXiv:1503.04493* .

Yao, Y., Vehtari, A., Simpson, D., Gelman, A. et al. (2018), 'Using stacking to average bayesian predictive distributions', *Bayesian Analysis* .

Yeh, I.-C. & Lien, C.-h. (2009), 'The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients', *Expert Systems with Applications* **36**(2), 2473–2480.

Yoon, J., Jordon, J. & van der Schaar, M. (2018), 'Ganite: Estimation of individualized treatment effects using generative adversarial nets'.

Yu, Q. & Li, B. (2020), 'Model-guided adaptive sampling for bayesian model selection', *Journal of the Korean Statistical Society* pp. 1–19.

Yu, Q., Li, B., Fang, Z. & Peng, L. (2010), 'An adaptive sampling scheme guided by bart—with an application to predict processor performance', *Canadian Journal of Statistics* **38**(1), 136–152.

Yu, Q., Li, B., Fang, Z. & Peng, L. (2012), 'Model guided adaptive design and analysis in computer experiment', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **5**(5), 399–409.

Zeldow, B., Re III, V. L., Roy, J. et al. (2019), 'A semiparametric modeling approach using bayesian additive regression trees with an application to evaluate heterogeneous treatment effects', *The Annals of Applied Statistics* **13**(3), 1989–2010.

Zhang, J. L. & Härdle, W. K. (2010), 'The bayesian additive classification tree applied to credit risk modelling', *Computational Statistics & Data Analysis* **54**(5), 1197–1205.

Zhou, T., Daniels, M. J. & Müller, P. (2019), 'A semiparametric bayesian approach to dropout in longitudinal studies with auxiliary covariates', *Journal of Computational and Graphical Statistics* (just-accepted), 1–32.

Appendices

Appendix A

First Chapter Appendix

A.1 Simulation Study - Variable Importance Permutation Test

We present a simulation study investigating the extent to which p-values for a permutation-based variable importance test are influenced by the bias of the variable importance measure towards continuous variables and categorical variables with more categories. This study is designed in a similar way to that used by Strobl (2008) for investigating the bias of random forest variable importance measures.

First, we generate the following covariates and treatment indicator: $X_1 \sim N(0,1)$, $X_2 \sim Cat(2)$, $X_3 \sim Cat(4)$, $X_4 \sim Cat(10)$, $X_5 \sim Cat(20)$, $treatment \sim Cat(2)$, where $Cat(k)$ denotes a categorical distribution with k categories of equal probability. Then we consider simulations of the outcome under the following three model designs:

For design 1, none of the covariates affect the outcome, and the outcome is normally distributed: $Y \sim N(0,1)$ For design 2 and 3, the dependent variable is defined in a similar way to a simulation study carried out by Athey & Imbens (2016):

$$Y = \eta(X) + \frac{1}{2}(2 \times treatment - 1) \times \kappa(X) + \epsilon \tag{A.1}$$

where $\epsilon \sim N(0,1)$. For design 2 the functions are $\eta(X) = 0$ and $\kappa(X) = X_2$, and for design 3 the functions are $\eta(X) = \frac{1}{2}X_1 + X_2$ and $\kappa(X) = X_2$.

We simulate these designs 100 times, with 500 observations per simulation, and for each simulation we permute the dependent variable 100 times and obtain p-values, and then present boxplots of the p-values for each variable.[1] The boxplots of variable importances obtained using the unpermuted dependent variable are shown in Figure A.1. The boxplots for the p-values are shown in Figure A.2. The boxes give the lower quartile, median, and upper quartiles across repeated simulations. The whiskers give the most extreme data points that are no more than 1.5 times the interquartile range from the box. The circles denote outliers.

Note that the results in Figures A.1 and A.2 should be interpreted differently. The variable importances in Figure A.1 are not used in a test of significance, but rather in a comparison of importance across variables. In contrast, Figure A.2 is clearer and correctly indicates that the binary variable is significant in designs 2 and 3. This is an argument in favour of the permutation test.

Although for design 1 none of the variables affects the outcome, in Figure A.1a X_1 has greater variable importance than X_2, because of the aforementioned bias towards continuous variables.

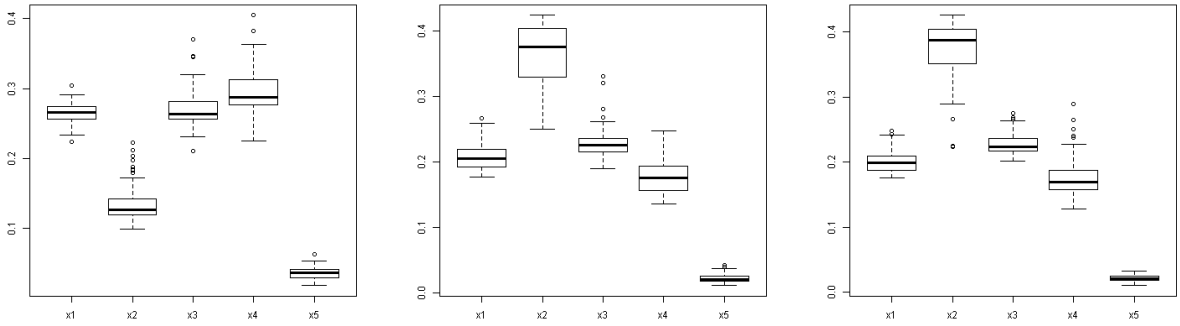
For categorical variables X_3, X_4, and X_5, all with more categories than X_2, there are two factors influencing the bias of the variable importance measure. As the number of categories increases, there are more potential splits on the variable of interest, because there is a binary variable for each category. This explains why X_3 has greater variable importance than X_2 in Figure A.1a. On the other hand, considering the case

[1] The parameters for the causal forest are: Number of trees = 5000, bootstrap sample fraction = 0.5, number of potential splitting variables random selected at each split = number of variables divided by 3 and rounded down, minimum node size = 5.

of a variable with a large number of categories, X_5, there will be relatively few observations allocated to any one category, and therefore a split on one of the X_5 categories is unlikely to lead to a large improvement in the splitting criterion. Therefore the variable importance measures for X_5 are small.

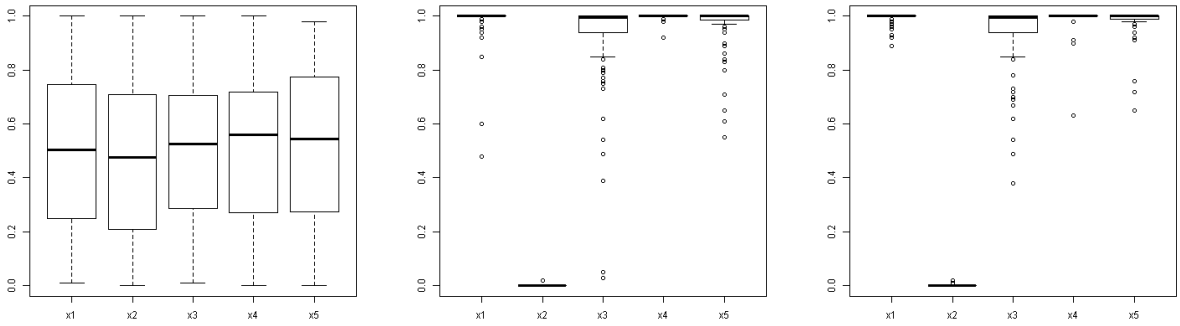
The p-values in Figure A.2 appear to be unaffected by these biases. In Figure A.2a, none of the variables tend to have significant p-values, reflecting the fact that none of the variables has any influence on the outcome.

In Figures A.2b and A.2c, X_2 is correctly identified as the important variable. Although Figures A.1b and A.1c also indicate that X_2 is the most important variable, there are also misleading differences in the importances of the other variables. However, in Figures A.2b and A.2c, the variables X_1, X_3, X_4, and X_5 tend to have similar, insignificant p-values.



(a) Design 1 var. imp. (b) Design 2 var. imp. (c) Design 3 var. imp.

Figure A.1: Boxplots of simulation study variable importances, 100 permutations, 100 iterations



(a) Design 1 p-values (b) Design 2 p-values (c) Design 3 p-values

Figure A.2: Boxplots of simulation study p-values, 100 permutations, 100 iterations

A.2 Classification Analysis Tables

Table A.1: Classification Analysis (CLAN): Age variable averages for most and least peak demand responsive households

Variable	25% most responsive	25% least responsive	Difference
18-25	0	0	0
26-35	0.095 (0.043, 0.147)	0.080 (0.032, 0.128)	-0.015 (-0.083, 0.058)
36-45	0.230 (0.156, 0.305)	0.152 (0.088, 0.216)	-0.078 (-0.178, 0.017)
46-55	0.349 (0.265, 0.434)	0.168 (0.102, 0.234)	-0.181 (-0.288, -0.074)
56-65	0.206 (0.135, 0.278)	0.192 (0.122, 0.262)	-0.014 (-0.115, 0.084)
65+	0.103 (0.049, 0.157)	0.392 (0.305, 0.479)	0.289 (0.190, 0.392)
Refused	0.016	0.008	-0.008

Note: For some binary variables with few non-zero values, confidence interval could not be obtained because for some sample splits there was not sufficient variation within quartiles for a confidence interval to be calculated.

Table A.2: Classification Analysis (CLAN): Class variable averages for most and least peak demand responsive households

Variable	25% most responsive	25% least responsive	Difference
Upper middle and middle	0.190 (0.121, 0.260)	0.056	0.135 (0.057, 0.217)
Lower middle	0.294 (0.213, 0.374)	0.216 (0.143, 0.289)	0.070 (-0.037, 0.181)
Skilled working	0.190 (0.121, 0.260)	0.128 (0.069, 0.187)	0.055 (-0.032, 0.148)
Working and non-working	0.302 (0.220, 0.383)	0.560 (0.472, 0.648)	-0.258 (-0.377, -0.139)
Farmers	0.016	0.024	0.008
Refused	0.008	0.008	0.000

Note: For some binary variables with few non-zero values, confidence interval could not be obtained because for some sample splits there was not sufficient variation within quartiles for a confidence interval to be calculated.

Table A.3: Classification Analysis (CLAN): Employment variable averages for most and least peak demand responsive households

Variable	25% most responsive	25% least responsive	Difference
Employee	0.0563 (0.0476, 0.0651)	0.328 (0.245, 0.411)	0.235 (0.113, 0.355)
Self-emp (with employees)	0.095 (0.043, 0.147)	0.008	0.087 (0.032, 0.141)
Self-emp (with no employees)	0.071 (0.026, 0.117)	0.032	0.039 (-0.015, 0.097)
Unemployed (seeking work)	0.048 (0.010, 0.085)	0.072 (0.026, 0.118)	-0.024 (-0.084, 0.035)
Unemployed (not seeking work)	0.040	0.040	-0.008 (-0.055, 0.043)
Retired	0.159 (0.094, 0.223)	0.496 (0.497, 0.585)	-0.337 (-0.445, -0.227)
Carer	0.024	0.008	0.008 (-0.022, 0.038)

Note: For some binary variables with few non-zero values, confidence interval could not be obtained because for some sample splits there was not sufficient variation within quartiles for a confidence interval to be calculated.

Appendix B

Second Chapter Appendix

B.1 Potential Variations on BART-BMA

The closed form for the predictive distribution suggests a number of possible improvements and variations on BART-BMA.

- The a parameter can be set by a full Bayesian approach, Empirical Bayes approach, cross-validation, or other methods. For comparison with the original results obtained by Hernández et al. (2018), this paper uses the arbitrary value $a = 3$ throughout.
- The BART-BMA predictions are a probability weighted average of ridge regressions. Methods for fast estimation of ridge regressions can be applied for improved computational speed.
- Different priors can be applied to the terminal node parameters and the error variance. This was discussed to an extent in the original single Bayesian tree context by Chipman et al. (1998). For example, data-informed priors can be applied to the error variance, as outlined in the context of standard Bayesian linear regression by Sala-i Martin et al. (2004). There is an extensive literature concerning the use of the g-prior in Bayesian Model Averaging (Eicher et al. 2011).
- Different model weights can be applied, for example, weights can be based on in-sample sum-of-squared errors. This was discussed to an extent by Chipman et al. (1998). For BMA of linear regressions, Sala-i Martin et al. (2004) suggest model weights equal to $p(M_j)n^{-k/2}SSE_j^{n/2}$, where M_j is the model and n is the number of observations. Another option is a BIC approximation, $p(M_j)[n\ln(\frac{1}{n}SSE_j) + k\ln(n)]$. This approach may improve computational speed.
- BART-BMA outputs a relatively small number of Bayesian linear regressions. The covariates are indicator variables for terminal nodes. In principle, any Bayesian model combination method can be applied to the set of selected models. Bayesian Stacking (Yao et al. 2018) might give a more accurate predictive distribution.

B.2 Comparison of Computational Times, Friedman Simulations

This appendix includes computational times for the Friedman simulations described in section 2.5.1. Table B.1 presents the computational times in seconds. The BART-IS results are for 1,000,000 draws of models each containing 30 trees. The BART and DART MCMC results are for 10,000 samples plus 1000 burn-in draws of models each containing 200 trees. It can be observed that Random forests with the default number of draws of 500 trees are much faster than BART based methods. BART-BMA is faster than other BART-based methods, particularly for the 100 and 1000 variable simulations.[1] BART-IS is slower than BART-MCMC,

[1]The speed of the BART-BMA algorithm can be improved further by using the Pruned Exact Linear Time (PELT) change-point detection algorithm (Killick et al. 2012). This is particularly recommended when the number of observations is large.

however this depends on the number of draws and the number of processors available. BART-IS with 250,000 draws would give comparable or faster results than BART-MCMC. Alternatively, if the number of processors were scaled up from 7 to 30, BART-IS would have comparable speed to BART-MCMC.[2] The slow-down of BART-IS as the number of variables increases is surprising given that the size of the drawn models does not change. This suggests that the random sampling of splitting variables and construction of terminal node indicator variables slows down with the number of variables in the dataset. This may be sensitive to the choice of model sampler. There are a number of possible approaches for speeding up BART-IS. If the increase in computational time is only due to sampling of splitting variables from a discrete uniform categorical distribution, then offline sampling of models removes this problem entirely.

Method	BART-BMA	BART-IS	RF	BART-MCMC	DART
100 variables	3.19	870.65	0.19	218.40	226.25
1000 variables	32.66	1164.39	0.48	230.84	259.54
5000 variables	163.23	1332.77	1.62	356.64	426.96
10000 variables	337.15	1617.75	3.54	498.90	608.50
15000 variables	520.92	1754.45	6.94	626.51	814.61

Table B.1: Computational times, in seconds, for Friedman data simulations.

B.3 Multivariate BART-IS

For multivariate BART-IS, options include the use of the same tree structures for different outcomes (similar to shared Bayesian Forests (Linero et al. 2019)), or different tree structures for each outcome (as in BART for Seemingly Unrelated Regression (Chakraborty 2016)). [3]

Let the vector of d outcomes for individual i be denoted by $\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,d})^T$. Then, if we impose the same tree structure on the model for all outcomes, we have

$$\boldsymbol{y}_i = \begin{bmatrix} O_1^T \\ O_2^T \\ \vdots \\ O_d^T \end{bmatrix} (W)_i^T + \begin{bmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,d} \end{bmatrix}$$

Where $O_1, ..., O_d$ are distinct terminal node coefficient vectors for each outcome, $(W)_i$ is the i^{th} row of the W matrix of terminal node indicator variables, and $\varepsilon_{i,j}$ is the error for individual i, outcome j.

Alternatively, one can allow for distinct sets of tree structures for each outcome, with corresponding matrices of terminal node indicator variables $W_1, ..., W_d$. It is also possible for splits in each sum-of-tree model to be constructed from different sets of potential splitting variables.

However, in simulated examples, BART with the grid-search algorithm tends to outperform BART with the PELT algorithm in terms of accuracy of predictions.

[2] However, a fairer comparison would also make use of a parallelized version of BART, either using multiple chains or the approach described by Pratola et al. (2014). Nonetheless, since BART-IS is in principle more parallelizable than BART-MCMC, there should exist some number of processors such that BART-IS is faster than BART-MCMC.

[3] Code has not yet been written for Multivariate BART-IS. This appendix only outlines the idea.

This gives the following model:

$$\boldsymbol{y}_i = \begin{bmatrix} O_1^T & 0 & 0 & \dots & 0 \\ 0 & O_2^T & 0 & \dots & 0 \\ 0 & 0 & O_3^T & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 0 & O_d^T \end{bmatrix} \begin{bmatrix} (W_1)_i^T \\ (W_2)_i^T \\ \vdots \\ (W_d)_i^T \end{bmatrix} + \begin{bmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,d} \end{bmatrix}$$

The tree drawing process for any tree is the same as in univariate BART-IS. There exist priors that give a closed form for the marginal likelihood and posterior predictive distribution (Minka 2000).

B.4 Importance Sampling Implementation of Semiparametric BART

Zeldow et al. (2019) outline semiparametric BART, which is essentially an average of models, each consisting of a sum-of-trees plus a linear model. Zeldow et al. (2019) present the approach in the context of treatment effect estimation, but it can also be applied to prediction. It is straightforward to average over models that are defined by the addition of a linear combination of covariates and a sums-of-trees. First, define a set of covariates that can be included in the linear model, then define prior inclusion probabilities for these covariates [such that the prior is independent of the prior over the sum-of-trees], and priors on the coefficients of included covariates. The prior for coefficients should allow for conjugacy of the whole model, for example, the prior variance of coefficients can be set equal to a scalar multiple of the variance of the error terms.

Then, for each sampled model, we sample the variables included in the linear part of the model by a set of Bernoulli draws, and this gives a covariate matrix X. Then draw the sum-of-trees part of the model as in standard BART-IS, which gives a matrix W and define the overall model matrix as $[X\ W]$.

The resulting models are standard Bayesian linear regressions, and the marginal likelihoods and predictive distributions have closed forms. Importance sampling of BART plus a linear model can be viewed as a combination of BART-IS and the implementation of BMA of Bayesian linear regressions used by Sala-i Martin et al. (2004).

B.5 Spike-and-Tree Prior

B.5.1 Definition of Spike-and-Tree Prior

Results are presented in this paper for BART-BMA with the spike and tree prior described by Rockova & van der Pas (2017) (as an alternative to the standard BART splitting prior). The prior is defined by $\pi(\mathcal{S}|q) = \frac{1}{\binom{p}{q}}$ for $\alpha, q, c > 0$. This prior can be implemented by taking a Bernoulli draw for inclusion of each variable, with a conjugate beta prior distribution on the splitting probability. (i.e. the number of splitting variables can be given a beta-binomial distribution). A drawn variable is used at least once in the tree.

A Poisson prior is placed on the number of terminal nodes, $\pi(k) = \frac{\lambda_0^k}{(e^{\lambda_0}-1)k!}$, $k = 1, 2, ...$ for some $\lambda_0 \in \mathbb{R}$. However, this should be $\pi(k|q)$ and truncated from the left and to the right so that $q \leq k \leq n-1$, where right truncation only occurs with a date-informed prior that requires every terminal node contains at least one observation.

Then, given $q, \mathcal{S}, k$, assign a uniform prior over valid tree topologies $\mathcal{T} = \{\Omega_k\}_{k=1}^{K} \in \mathcal{V}_c^k$. A valid tree topology must have some minimum number of training observations in each terminal node. The prior probability of a valid tree is $\pi(\mathcal{T}|\mathcal{S}, k) = \frac{1}{\Delta(\mathcal{V}_s^k)}\mathbb{I}(\mathcal{T} \in \mathcal{V}_s^k)$. The number of possible valid tree constructions is $S(k-1,q)q!(n-1)!/(n-k)!$, where $S(k-1,q)$ is a Sterling number of the second kind. This can be used to account for duplications of the same tree by multiple possible tree constructions in the BART-BMA algorithm. The number of valid tree diagrams is equal to $C_{k-1}q!S(k-1,q)\binom{n-1}{k-1}$, where C_{k-1} is the $k-1^{th}$ Catalan number.

B.5.2 Sampling from the Spike-and-Tree Prior

1. Bernoulli draws on the set of included variables. Obtain a set of variables, $\mathcal{S}$, with $|\mathcal{S}| = q$.

2. Draw number of terminal nodes, k from a Poisson distribution truncated on the left (if we require that the tree splits on each variable in $|\mathcal{S}|$ at least once) and right such that $q \leq K-1 \leq n$, i.e. $q+1 \leq K \leq n+1$.

3. Draw a tree structure with the specified number of terminal nodes uniformly at random. This is an efficient algorithm created by Bacher (2016). This gives a representation of the tree structure.

4a. (If using a data-independent prior) Take a standard uniform draw for each split point. Then loop through splitting points, and adjust splitting points within the corresponding sub-tree that split on the same variable again such that it is possible for observations to fall in any terminal node.

4b. (If using the data-dependent prior) Draw a set of splitting points from the $n-1$ possible splits of the variables. Here, the splits are splits of the n observations (i.e. still in one dimension, we haven't allocated splits to the variables yet. Each split "point" just specifies the number of observations that are to the left of that split. Note that for each of these split "points" there is a possible split on each variable).

   Fill in the splits in the tree. Apply the following algorithm:

   While there are split points remaining:

   (a) Take the lowest remaining split point.

   (b) Allocate it to the leftmost remaining internal node.

   (c) Remove the split point and internal node.

5. For each internal node, randomly draw a splitting variable from $\mathcal{S}$. There will be one split point on the chosen variable that results in the number of observations to the left allocated to that split in step 5.

   If we want to apply the condition that each of the $|S|$ potential splitting variables must be used at least once, then we can first draw from all possible variables $(K-1)-|S|$ times with replacement, but then start restricting the number of possible draws, i.e. draw $|S|$ times without replacement. Then randomly shuffle the splitting variables among the splitting points. [An alternative would be any algorithm that creates random (ordered) partitions of the $K-1$ splitting points among the $|\mathcal{S}|$ splitting variables.]

B.6 BCF-BMA Algorithm

.

Input: $n \times p$ matrix X with continuous response variable Y
Output: RMSE, Credible interval for $\hat{Y}$, after burn-in updates for σ
Initialise: $Tree\_Response = Y\_scaled$;
Initialise: $lowest\_BIC$, $L = 1$, Set of $\mathcal{T} = List\_ST$ = a tree stump
Initialise: $count\_mu\_trees_\ell = 1$, $count\_tau\_trees_\ell = 1$
for $j \leftarrow 1$ *to* $m_\mu + m_\tau$ **do**
  for $\ell \leftarrow 1$ *to* L **do**
    if $count\_mu\_trees_\ell \leq m_\mu$ **then**
      1. **Find Good Splitting Rules**. Run greedy search to find $numcp\%_\mu$ best split rules for each current sum of trees $\mathcal{T}_{\mu\ell}$ in $\mathcal{T}_\ell$ in Occam's window, using the partial residuals of $\mathcal{T}_\ell$ as $Tree\_response$.
      2. **Grow greedy trees based on their partial residuals to append to current sum of trees model** $\mathcal{T}_{\mu\ell}$. Set new proposal tree T^* to stump

      for $H \leftarrow 1$ *to* $max\_tree\_depth_\mu$ **do**
        for $i \leftarrow 1$ *to number of terminal nodes in* T^* **do**
          for $d \leftarrow 1$ *to* $num\_split\_rules_\mu$ **do**
            Grow proposal tree T^* using splitting rule d from list of splitting rules found in part 1. Append T^* to $\mathcal{T}_{\mu\ell}$ to make new sum of trees model $\mathcal{T}_\ell^*$. **if** *Sum of trees* $\mathcal{T}_\ell^*$ *is in Occam's window* **then**
              Append T^* to $\mathcal{T}_{\mu\ell}$ and save new sum of trees model to temporary list $temp_{OW}$, and save new values of $count\_mu\_trees := count\_mu\_trees_\ell + 1$, and $count\_tau\_trees := count\_tau\_trees_\ell$ for each element of $temp_{OW}$ in lists $temp\_count\_mu\_list$ and $temp\_count\_tau\_list$.
            end
          end
        end
      end
    end
    if $j \leq count\_tau\_trees_\ell$ **then**
      1. **Find Good Splitting Rules**. Run greedy search to find $numcp\%_\tau$ best split rules for each current sum of trees $\mathcal{T}_{\tau\ell}$ in $\mathcal{T}_\ell$ in Occam's window, using the partial residuals of $\mathcal{T}_\ell$ *for treated individuals only* as $Tree\_response$.
      2. **Grow greedy trees based on their *treated individuals'* partial residuals to append to current sum of trees** $\mathcal{T}_{\tau\ell}$. Set new proposal tree T^* to stump

      for $H \leftarrow 1$ *to* $max\_tree\_depth_\tau$ **do**
        for $i \leftarrow 1$ *to number of terminal nodes in* T^* **do**
          for $d \leftarrow 1$ *to* $num\_split\_rules_\tau$ **do**
            Grow proposal tree T^* using splitting rule d from list of splitting rules found in part 1. Append T^* to $\mathcal{T}_{\tau\ell}$ to make new sum of trees model $\mathcal{T}_\ell^*$. **if** *Sum of trees* $\mathcal{T}_\ell^*$ *is in Occam's window* **then**
              Append T^* to $\mathcal{T}_{\tau\ell}$ and add new sum of trees model to temporary list $temp_{OW}$, and save a new value of $count\_mu\_trees := count\_mu\_trees_\ell$, and $count\_tau\_trees := count\_tau\_trees_\ell + 1$ for each element of $temp_{OW}$ in lists $temp\_count\_mu\_list$ and $temp\_count\_tau\_list$.
            end
          end
        end
      end
    end
    Make sum of trees models and update residuals
    List of sum of trees models to grow further $List\_ST = temp_{OW}$
    List of all sum of trees models to date $sum\_trees\_in\_OccamsWindow += temp_{OW}$
    Lists of counts of mu trees and tau trees in all sum of tree models to date $count\_mu\_trees += temp\_count\_mu\_list$, $count\_tau\_trees += temp\_count\_tau\_list$.
    Update $lowest\_BIC = min(sum\_trees\_in\_OccansWindow)$
    Set $L = length(temp_{OW})$
    Set $length(temp_{OW}) = 0$
  end
end
Get total list of L sum of trees in Occam's window by deleting models from $sum\_trees\_in\_OccamsWindow$ list whose BIC is greater than $\log(o)$ from $lowest\_BIC$
$\hat{\tau}$ = **Sum of weighted predictions** $\hat{\tau}_\ell$ **over all** L **sum of trees models in Occam's window**
For prediction intervals, obtain quantiles by a root finding algorithm (or implement a post hoc Gibbs Sampler for each sum of trees accepted in Occam's window)
return:
Credible intervals for $\hat{\tau}$; Sum of trees in Occam's window;
Posterior probability of each sum of trees model.

Algorithm 4: BCF-BMA Algorithm

Figure B.1: Example results for BCF and BCF-BMA

B.7 BCF Applied to CER Smart Meter Trial Data

Figure B.1 investigates the issue noted in section 2.7.1 concerning the unrealistic demand response estimates produced by a standard causal forest. The ITE estimates are presented for a Bayesian Causal Forests with 2000 draws and varying numbers of μ trees and τ trees, and also for BCF with 4000 draws. It can be seen in figure B.1 that the extent to which the algorithm produces unrealistic estimates is highly sensitive to the number of trees and the number of draws.

Appendix C

Third Chapter Appendix

C.1 Standard Newton-Raphson algorithm for finding the MAP of Bayesian Logistic Regression

---

Require parameter value, e.g. $a = 0.01$

Initialize $\underline{\boldsymbol{\mu}} = \mathbf{0}$

repeat

$p_i = \frac{1}{1+e^{-W_i \underline{\mu}}}$ for $i = 1, ..., n$

$S = diag(p_i(1 - p_i))$

$\boldsymbol{g} = W^T(\boldsymbol{p} - \boldsymbol{y}) + a\underline{\boldsymbol{\mu}}$

$H = W^T SW + aI_b$

$\underline{\boldsymbol{\mu}}_{new} = \underline{\boldsymbol{\mu}}_{old} - H^{-1}\boldsymbol{g}$

until *convergence*;

Algorithm 5: Newton's method for obtaining the mode (MAP) of the posterior approximation

C.2 Marginal Likelihood Approximation

The Laplace Approximation gives the following approximation for the marginal likelihood (using the normalization constant of the multivariate Gaussian distribution)

$$p(\boldsymbol{y}|\mathcal{T}^{(m)}) \approx e^{\log(p(\underline{\boldsymbol{\mu}}_{MAP,(m)}, \boldsymbol{y}))}(2\pi)^{b^{(m)}/2}|H_{MAP,(m)}|^{-1/2}$$

$$= (\frac{1}{a})^{-\frac{b^{(m)}}{2}} \prod_{i=1}^{N} \left[\left(\frac{e^{row_i(W_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}}}{1 + e^{row_i(W_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}}} \right)^{y_i} \left(\frac{1}{1 + e^{row_i(W_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}}} \right)^{1-y_i} \right] |H_{MAP,(m)}|^{-1/2}$$

and the log of the marginal likelihood is approximated by:

$$\frac{b^{(m)}}{2}\log(a) + \sum_{i=1}^{N} \left[y_i row_i(W_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)} - \log\left(1 + e^{row_i(W_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}}\right)\right] - \frac{1}{2}log(|H_{MAP,(m)}|)$$

C.3 Applying Laplace's Method Approximation Twice to Approximate Posterior Mean Probability

Tierney & Kadane (1986) describe an approach for approximating the posterior mean of any smooth unimodal

function of the parameters, $g(\theta)$.[1] This involves the observation that the posterior mean can be approximated by first applying Laplace's method to find the mode of the integral in the numerator of the posterior mean of the function.

$$\mathbb{E}[g] = \mathbb{E}[g(\theta|X)] = \frac{\int g(\theta)e^{\mathcal{L}(\theta)}\pi(\theta)d\theta}{\int e^{\mathcal{L}(\theta)}\pi(\theta)d\theta}$$

where $\mathcal{L}$ is the log likelihood function.

First, the MAP of the posterior for θ is obtained by Newton's method. Then Laplace's method is used to obtain an approximation for the denominator integral.

Then this is combined with a Laplace approximation for the integral in the numerator.

Let $L = \log(\pi(\theta)) + \frac{\mathcal{L}(\theta)}{n}$ and $L^* = \log(g(\theta)) + \log(\pi(\theta)) + \frac{\mathcal{L}(\theta)}{n}$. Then

$$\mathbb{E}[g] = \mathbb{E}[g(\theta|X)] = \frac{\int e^{nL^*}d\theta}{\int e^{nL}d\theta}$$

Let $\hat{\theta} = \theta_{MAP}$ be the mode of L. Similarly let $\hat{\theta}^*$ be the mode of L^*. Then, taking the ratio of the two Laplace approximations gives:

$$\hat{E}_n[g] = \left(\frac{\det(H^{*-1})}{\det(H^{-1})}\right)^{1/2} \exp\{n(L^*(\hat{\theta}^*) - L^*(\hat{\theta}))\}$$

where H and H^* are the negatives of the Hessians of L and L^* respectively (i.e. the Hessians of the negative log likelihood). The error is of order $\mathcal{O}(n^{-2})$.

This can in principle be applied to the logit model with $g(\underline{\boldsymbol{\mu}}_{(m)}) = \frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m)}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m)}}}$

C.4 Outline of Monte Carlo Approximation for Logit-BART-BMA and Logit-BART-IS

C.4.1 Monte Carlo Approximation of Posterior Predictive Mean Probability

Two possible Monte Carlo approximation approaches are:

1. Approximate each integral separately, and then average by the model posterior probability. i.e. For each model, obtain a large number S of samples of $\underline{\boldsymbol{\mu}}_{(m),1}, ..., \underline{\boldsymbol{\mu}}_{(m),S}$ from the approximate distribution $\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})$ and estimate the posterior predicted probability of $y_* = 1$ for model m as $\frac{1}{S}\sum_{s=1}^{S} \frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}$ and then the model averaged probability is:

$$\sum_{m=1}^{M} p(\mathcal{T}_m|\boldsymbol{y})\frac{1}{S}\sum_{s=1}^{S}\frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}$$

2. Take a large number, S, of samples from the mixture of multivariate normal distributions $\underline{\boldsymbol{\mu}}|\boldsymbol{y} \sim \sum_{m=1}^{M}\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})p(\mathcal{T}_m|\boldsymbol{y})$. Note that this involves sampling from each model's normal approximation with probability $p(\mathcal{T}_m|\boldsymbol{y})$, and for each model the sampled vector $\underline{\boldsymbol{\mu}}$ has a different interpretation and

[1] The function is also required to be nonzero, and preferably positive, but it is possible to add a large constant or take the negative

can have different dimensions because the sum-of-tree structures differ across models. Then the estimate is

$$\frac{1}{S}\sum_{s=1}^{S}\frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}$$

C.4.2 Monte Carlo Approximation of Credible Intervals for Posterior Predictive Probability

Take a large number, S, samples from the mixture of multivariate normal distributions

$$\underline{\boldsymbol{\mu}}|\boldsymbol{y} \sim \sum_{m=1}^{M}\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H_{(m)}^{-1})p(\mathcal{T}_m|\boldsymbol{y})$$

Note that this involves sampling from each model's normal approximation with probability $p(\mathcal{T}_m|\boldsymbol{y})$, and for each model the sampled vector $\underline{\boldsymbol{\mu}}$ has a different interpretation and can have different dimensions because the sum-of-tree structures differ across models.

For each draw, calculate the probability $\frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}$. Then obtain the desired quantiles of the S probabilities. e.g. For a 95% interval, find L_b and U_b such that $\frac{1}{S}\sum_{s=1}^{S}\mathbb{I}\left(\frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}} < L_b\right) = 0.025$ and $\frac{1}{S}\sum_{s=1}^{S}\mathbb{I}\left(\frac{e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{W_{*,(m)}\underline{\boldsymbol{\mu}}_{(m),s}}} < U_b\right) = 0.975$.

C.5 Root-finding Approximation of Credible Intervals for Posterior Predictive Probability

Using the sigmoid (logistic) function, $sig()$, we require L_b such that

$$\sum_{m=1}^{M}\left(\int_{-\infty}^{\infty}\mathbb{I}\left(sig(\alpha_{(m)}) < L_b\right)p(\alpha_{(m)}|\psi_{\alpha,(m)},\sigma^2_{\alpha,(m)})d\alpha_{(m)}\right)p(\mathcal{T}_m|\boldsymbol{y})$$

$$=\sum_{m=1}^{M}\Phi\left(\frac{sig^{-1}(L_b)-\psi_{\alpha,(m)}}{\sigma_{\alpha,(m)}}\right)p(\mathcal{T}_m|\boldsymbol{y}) = 0.025$$

A simple root finding algorithm (e.g. bisection) can be used to find $c = sig^{-1}(L_b)$. Then L_b is obtained from $sig(c)$. Similarly, for the upper bound, we require U_b such that $\sum_{m=1}^{M}\Phi\left(\frac{sig^{-1}(U_b)-\psi_{\alpha,(m)}}{\sigma_{\alpha,(m)}}\right)p(\mathcal{T}_m|\boldsymbol{y}) = 0.975$, which can be obtained by a root-finding algorithm.

C.6 Alternative methods for constructing Logit-BART-BMA Residuals

C.6.1 Constructing Residuals using Predicted Probabilities or MAP Estimates

Three simple methods for calculation of residuals are::

- A naive approach resulting from an unedited model search algorithm using the residuals $y_i - \widehat{Pr}(y_i = 1)$ where $\widehat{Pr}(y_i = 1)$ is the estimated probability from the model. However, the tree will be appended to

a sum-of-trees modelling the latent outcome, which is a continuous variable that is not restricted to be between 0 and 1.

- Residuals can be calculated for the latent outcome U by beginning with $U_i = 3.1$ if $y_i = 1$ and $U = -3.1$ if $y_i = 0$ (or a similar number that gives a probability close to 1 or 0) and for each model obtaining $U_i - row_i(W)\underline{\boldsymbol{\mu}}_{MAP}$ as the residual to be used in the changepoint detection algorithm in the next round, where $row_i(W)\underline{\boldsymbol{\mu}}_{MAP}$ is the MAP prediction of the latent outcome y^*.
- An even less computationally burdensome approach would be to only search for potential splits before the first round of the algorithm. This involves applying the changepoint detection algorithm to the latent outcome U defined by $U_i = 3.1$ if $y_i = 1$ and $U = -3.1$ if $y_i = 0$. Then keep these potential split points for all future rounds of the algorithm (i.e. do not apply the changepoint algorithm again).

C.6.2 Arbitrary fixed grid of splits, without residuals

The changepoint detection algorithm can be replaced by an alternative method for reducing the number of potential splitting points.

- Propose an arbitrary deterministic grid of splitting points, possibly after applying a Probability Integral Transform using the Empirical Distribution Function of the residuals, and proceed to use these splits in the rest of the algorithm without applying a changepoint detection algorithm. This is likely to be very slow, particularly for high-dimensional data, unless the set of potential splits for each variable is severely restricted, which may compromise the ability of these methods to find models close to the true data generating process.
- Alternatively, the grid of points for each variable can be found by first applying some other tree-based method or search algorithm. For example, one could use standard BART-BMA or the simple Probit-BART-BMA and save the splitting points to use in Logit-BART-BMA. [2]

C.7 Technical Details for Logit-BART-BMA and Logit-BART-IS Treatment Effect Estimation

C.7.1 Estimation of Mean of Posterior Distribution of Individual Treatment Effects

Beginning with the Laplace approximations of posterior distributions of the terminal node coefficients outlined in section 3.4.3,
$\underline{\boldsymbol{\mu}}_{(m)}|\boldsymbol{y}, \mathcal{T}_m \sim \mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})$, the goal is to estimate the expected difference in the the probability $y_* = 1$ for an individual with and without treatment, i.e. $T_* = 1$ and $T_* = 0$, conditioning on the same set of values for other covariates x_* in both cases. i.e. Estimate $\mathbb{E}[y_*|x_*, T_* = 1] - \mathbb{E}[y_*|x_*, T_* = 0]$.

When treatment is a splitting variable in the sum-of-tree model, the terminal nodes that individual i is allocated to when we set $T_i = 1$ can be different to the terminal nodes for $T_i = 0$. Therefore the variables indicating inclusion in terminal nodes will be different in these two scenarios. Denote these two sets of indicator variables as $row_i(W^{tr})$ and $row_i(W^c)$ for allocation to treatment and control respectively. Then

[2] i.e. Save all the splits that were used or suggested in a first step BART-BMA (as if the outcome were continuous) or approximate Probit-BART-BMA

for a new observation, with original covariate vector x_*, we estimate the expected difference in probabilities for $row_*(W^{tr})$ and $row_*(W^c)$. Therefore the expected treatment effect is:

$$\sum_{m=1}^{M}\left[\int\left(sig(row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})-sig(row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})\right)p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

this can also be separated into two integrals

$$\sum_{m=1}^{M}\left[\int sig(row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}-\int sig(row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

where *sig* denotes the sigmoid function (i.e. logistic). We consider two possible approaches: Monte Carlo and Probit approximation.

Monte Carlo Approximation of Expected ITE

Two possible approaches to Monte Carlo Approximation of the Expected ITE are:

1. Approximate each integral and then average by the model posterior probability. i.e. For each model, obtain a large number S of samples of $\underline{\boldsymbol{\mu}}_{(m),1},...,\underline{\boldsymbol{\mu}}_{(m),S}$ from the approximate distribution $\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})$ and estimate the difference in probabilities for model m. Then the model averaged difference in probabilities (treated minus untreated) is:

$$\sum_{m=1}^{M}p(\mathcal{T}_m|\boldsymbol{y})\frac{1}{S}\sum_{s=1}^{S}\left[\frac{e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}-\frac{e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}\right]$$

2. Take a large number, S, samples from the mixture of multivariate normal distributions $\underline{\boldsymbol{\mu}}|\boldsymbol{y}\sim\sum_{m=1}^{M}\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})p(\mathcal{T}_m|\boldsymbol{y})$. Note that this involves sampling from each model's normal approximation with probability $p(\mathcal{T}_m|\boldsymbol{y})$, and for each model the sampled vector $\underline{\boldsymbol{\mu}}$ has a different interpretation and can have different dimensions because the sum-of-tree structures differ across models. Then the estimate is

$$\frac{1}{S}\sum_{s=1}^{S}\left[\frac{e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}-\frac{e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}\right]$$

Probit Approximation of Expected ITE

The sigmoid (logistic) function can be approximated by a normal CDF:

$$\sum_{m=1}^{M}\left[\int sig(row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}-\int sig(row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

$$\approx\sum_{m=1}^{M}\left[\int\Phi(row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}-\int\Phi(row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)},H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

Let $\psi_{\alpha,(m,tr)}=row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}$ and $\sigma^2_{\alpha,(m,tr)}=row_*(W^{tr}_{(m)})H^{-1}_{(m)}row_*(W^{tr}_{(m)})^T$ and $\psi_{\alpha,(m,c)}=$

$row_*(W^c_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}$ and $\sigma^2_{\alpha,(m,c)} = row_*(W^c_{(m)})H^{-1}_{(m)}row_*(W^c_{(m)})^T$. Then $\alpha_{(m,tr)} = row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)} \sim \mathcal{N}(\psi_{\alpha,(m,tr)}, \sigma^2_{\alpha,(m,tr)})$ and $\alpha_{(m,c)} = row_*(W^c_{(m)})\underline{\boldsymbol{\mu}}_{(m)} \sim \mathcal{N}(\psi_{\alpha,(m,c)}, \sigma^2_{\alpha,(m,c)})$

Then the integrals can be rewritten as one dimensional integrals, and the expected ITE is:[3]

$$\sum_{m=1}^{M}\left[\int \Phi(\alpha_{(m,tr)})p(\alpha_{(m,tr)}|\psi_{\alpha,(m,tr)},\sigma^2_{\alpha,(m,tr)})d\alpha_{(m,tr)} - \int \Phi(\alpha_{(m,c)})p(\alpha_{(m,c)}|\psi_{\alpha,(m,c)},\sigma^2_{\alpha,(m,c)})d\alpha_{(m,c)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

$$= \sum_{m=1}^{M}\left[\Phi\left(\frac{\psi_{\alpha,(m,tr)}}{\sqrt{1+\sigma^2_{\alpha,(m,tr)}}}\right) - \Phi\left(\frac{\psi_{\alpha,(m,c)}}{\sqrt{1+\sigma^2_{\alpha,(m,c)}}}\right)\right]p(\mathcal{T}_m|\boldsymbol{y})$$

Monte Carlo Approximation of ITE Intervals

Take a large number, S, of samples from the mixture of multivariate normal distributions $\underline{\boldsymbol{\mu}}|\boldsymbol{y} \sim \sum_{m=1}^{M}\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})p(\mathcal{T}_m|\boldsymbol{y})$. Note that this involves sampling from each model's normal approximation with probability $p(\mathcal{T}_m|\boldsymbol{y})$, and for each model the sampled vector $\underline{\boldsymbol{\mu}}$ has a different interpretation and can have different dimensions because the sum-of-tree structures differ across models.

For each sample, s calculate the difference in probabilities under treatment and control group allocation (i.e. for $W^{tr}_{(m)}$ and $W^c_{(m)}$), and then find the relevant quantiles.

For example, for a 95% interval, find L_b such that

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{I}\left\{\frac{e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} < L_b\right\} = 0.025$$

and find U_b such that $\frac{1}{S}\sum_{s=1}^{S}\mathbb{I}\left\{\frac{e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} < U_b\right\} = 0.975.$

Monte Carlo Approximation of ITE Interval, reducing the dimension of the integral

Unlike in section C.5, the interval for the ITE does not have an obvious closed form obtainable from a Probit approximation. However, it is still possible to reduce the dimension of the integral, such that when the integral is approximated by Monte Carlo methods, draws can be made from univariate or bivariate normal distributions (instead of potentially much higher dimensional draws of $\underline{\boldsymbol{\mu}}_{(m)}$.

The goal is to find L_b defined in the following formula

$$\sum_{m=1}^{M}\left[\int_{-\infty}^{\infty}\mathbb{I}\left(sig(\alpha_{(m,tr)}) - sig(\alpha_{(m,c)}) < L_b\right)p(\alpha_{(m,tr)},\alpha_{(m,c)}|\psi_{\alpha,(m,tr)},\sigma^2_{\alpha,(m,tr)},\psi_{\alpha,(m,c)},\sigma^2_{\alpha,(m,c)})d\alpha_{(m,tr)}d\alpha_{(m,c)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

where the variables and parameters are defined as in section C.5. Note that $(\alpha_{(m,tr)},\alpha_{(m,tr)})$ has the following bivariate normal distribution:

$$\begin{bmatrix}\alpha_{(m,tr)}\\ \alpha_{(m,c)}\end{bmatrix} = \begin{bmatrix}row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)}\\ row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)}\end{bmatrix}$$

[3] For closer approximations to logistic probabilities, this can be replaced by $\sum_{m=1}^{M}\left[\Phi\left(\frac{\psi_{\alpha,(m,tr)}}{\sqrt{\frac{8}{\pi}+\sigma^2_{\alpha,(m,tr)}}}\right) - \Phi\left(\frac{\psi_{\alpha,(m,c)}}{\sqrt{\frac{8}{\pi}+\sigma^2_{\alpha,(m,c)}}}\right)\right]p(\mathcal{T}_m|\boldsymbol{y})$.

$$\sim \mathcal{N}\left(\begin{bmatrix} row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)} \\ row_*(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)} \end{bmatrix} , \begin{bmatrix} row_*(W^{tr}_{(m)}) \\ row_*(W^{c}_{(m)}) \end{bmatrix} H^{-1} \begin{bmatrix} row_*(W^{tr}_{(m)})^T & row_*(W^{c}_{(m)})^T \end{bmatrix}\right)$$

It is possible to take S draws from the model weighted average of bivariate normal distributions (i.e. draw from each model's bivariate normal distribution with probability equal to the posterior model probability), and for each draw, s, calculate $sig(\alpha_{(m,tr),s}) - sig(\alpha_{(m,c),s})$ and then take obtain the desired quantiles of the draws.

However, it is also possible to reduce the integrals to one-dimensional integrals.

Note that the conditional distribution of $\alpha_{(m,tr)}|\alpha_{(m,c)}$ is

$$\alpha_{(m,tr)}|\alpha_{(m,c)} \sim$$
$$\mathcal{N}\left(\mathbb{E}[\alpha_{(m,tr)}|\alpha_{(m,c)}], (1-\rho^2)\sigma^2_{\alpha,(m,tr)}\right)$$

where $\rho = (row_*(W^c_{(m)})H^{-1}row_*(W^{tr}_{(m)})^T$

$\mathbb{E}[\alpha_{(m,tr)}|\alpha_{(m,c)}] = row_*(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)} + \rho\sqrt{\frac{\sigma_{\alpha,(m,tr)}}{\sigma_{\alpha,(m,c)}}}(\alpha_{(m,c)} - row_*(W^c_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)})$

Then the integral of interest can be re-written as

$$\sum_{m=1}^{M}\left[\int_{-\infty}^{\infty}\Phi\left(\frac{sig^{-1}(L_b + sig(\alpha_{(m,c)})) - \mathbb{E}[\alpha_{(m,tr)}|\alpha_{(m,c)}]}{\sqrt{(1-\rho^2)\sigma^2_{\alpha,(m,tr)}}}\right)\phi(\frac{\alpha_{(m,c)} - \psi_{\alpha,(m,c)}}{\sigma_{\alpha,(m,c)}})d\alpha_{(m,c)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

The *sig* function in the above integrals can be replaced by the normal CDF of a probit approximation if the computation is faster.

If an entirely deterministic algorithm is desired, deterministic numerical methods can probably be used to evaluate the univariate integrals in the above expression, however, this would have to be used in combination with a root finding algorithm, and in each iteration of the algorithm the integrals will have to be re-calculated. The integrals could probably be calculated using Monte Carlo methods, but again would have to be recalculated for each iteration of the root finding algorithm.

Therefore, the optimal approach may be to draw from the mixture of bivariate normal distributions, and obtain quantiles of calculated quantiles (the standard Monte Carlo approach, albeit with the dimension of the draws reduced to 2).

C.7.2 Estimation of Mean of Posterior Distribution of Conditional Average Treatment Effects

Now consider the Conditional Average Treatment Effect, i.e. $\frac{1}{N}\sum_{i=1}^{N}[\mathbb{E}[y_i|x_i, T_i = 1] - \mathbb{E}[y_i|x_i, T_i = 1]]$.

$$\sum_{m=1}^{M}\left[\int \frac{1}{N}\sum_{i=1}^{N}\left[\left(sig(row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)}) - sig(row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})\right)\right]p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right]p(\mathcal{T}_m|\boldsymbol{y})$$

this can also be separated into two integrals

$$\sum_{m=1}^{M}\left[\int \frac{1}{N}\sum_{i=1}^{N} sig(row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)} - \int \frac{1}{N}\sum_{i=1}^{N} sig(row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right] p(\mathcal{T}_m|\boldsymbol{y})$$

where *sig* denotes the sigmoid function (i.e. logistic). Note that $row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)}$ can be estimated for $i = 1, ..., N$ in one matrix calculation $W^{tr}_{(m)}\underline{\boldsymbol{\mu}}_{(m)}$.

Monte Carlo Approximation of Expected CATE

[This is essentially the same as for ITEs]

Two possible approaches to Monte Carlo Approximation of the Expected CATE are:

1. It is possible to approximate each integral and then average by the model posterior probability. i.e. For each model, obtain a large number S of samples of $\underline{\boldsymbol{\mu}}_{(m),1}, ..., \underline{\boldsymbol{\mu}}_{(m),S}$ from the approximate distribution $\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})$ and estimate the difference in probabilities for model m .Then model averaged difference in probabilities (treated minus untreated) is:

$$\sum_{m=1}^{M} p(\mathcal{T}_m|\boldsymbol{y})\frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}\right]$$

2. Take a large number, S, samples from the mixture of multivariate normal distributions $\underline{\boldsymbol{\mu}}|\boldsymbol{y} \sim \sum_{m=1}^{M}\mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})p(\mathcal{T}_m|\boldsymbol{y})$. Note that this involves sampling from each model's normal approximation with probability $p(\mathcal{T}_m|\boldsymbol{y})$, and for each model the sampled vector $\underline{\boldsymbol{\mu}}$ has a different interpretation and can have different dimensions because the sum-of-tree structures differ across models. Then the estimate is

$$\frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}\right]$$

Probit Approximation of Expected CATE

The sigmoid (logistic) function can be approximated by a normal CDF:

$$\sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\left[\int sig(row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)} - \int sig(row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right] p(\mathcal{T}_m|\boldsymbol{y})$$

$$\approx \sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\left[\int \Phi(row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)} - \int \Phi(row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m)})p(\underline{\boldsymbol{\mu}}_{(m)}|\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)})d\underline{\boldsymbol{\mu}}_{(m)}\right] p(\mathcal{T}_m|\boldsymbol{y})$$

Let $\psi_{\alpha,i,(m,tr)} = row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}$ and $\sigma^2_{\alpha,i,(m,tr)} = row_i(W^{tr}_{(m)})H^{-1}_{(m)}row_i(W^{tr}_{(m)})^T$ and $\psi_{\alpha,i,(m,c)} = row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{MAP,(m)}$ and $\sigma^2_{\alpha,i,(m,c)} = row_i(W^{tr}_{(m)})H^{-1}_{(m)}row_i(W^{c}_{(m)})^T$. Then $\alpha_{i,(m,tr)} = row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m)} \sim$

$\mathcal{N}(\psi_{\alpha,i,(m,tr)}, \sigma^2_{\alpha,i,(m,tr)})$ and $\alpha_{i,(m,c)} = row_i(W^c_{(m)})\underline{\boldsymbol{\mu}}_{(m)} \sim \mathcal{N}(\psi_{\alpha,i,(m,c)}, \sigma^2_{\alpha,i,(m,c)})$ Then the integrals can be rewritten as one dimensional integrals, and the expected ITE is:

$$\sum_{m=1}^{M} \frac{1}{N} \sum_{i=1}^{N} \left[\int \Phi(\alpha_{i,(m,tr)}) p(\alpha_{i,(m,tr)} | \psi_{\alpha,i,(m,tr)}, \sigma^2_{\alpha,i,(m,tr)}) d\alpha_{i,(m,tr)} - \int \Phi(\alpha_{i,(m,c)}) p(\alpha_{i,(m,c)} | \psi_{\alpha,i,(m,c)}, \sigma^2_{\alpha,i,(m,c)}) d\alpha_{i,(m,c)} \right] p(\mathcal{T}_m | \boldsymbol{y})$$

$$= \sum_{m=1}^{M} \frac{1}{N} \sum_{i=1}^{N} \left[\Phi\left(\frac{\psi_{\alpha,i,(m,tr)}}{\sqrt{1+\sigma^2_{\alpha,i,(m,tr)}}} \right) - \Phi\left(\frac{\psi_{\alpha,i,(m,c)}}{\sqrt{1+\sigma^2_{\alpha,i,(m,c)}}} \right) \right] p(\mathcal{T}_m | \boldsymbol{y})$$

This is equal to the arithmetic average of the ITE estimates.[4]

C.7.3 Credible Intervals for CATE Posterior Distribution

Monte Carlo Approximation of CATE Intervals

Take a large number, S, samples from the mixture of multivariate normal distributions $\underline{\boldsymbol{\mu}} | \boldsymbol{y} \sim \sum_{m=1}^{M} \mathcal{N}(\underline{\boldsymbol{\mu}}_{MAP,(m)}, H^{-1}_{(m)}) p(\mathcal{T}_m | \boldsymbol{y})$. Note that this involves sampling from each model's normal approximation with probability $p(\mathcal{T}_m | \boldsymbol{y})$, and for each model the sampled vector $\underline{\boldsymbol{\mu}}$ has a different interpretation and can have different dimensions because the sum-of-tree structures differ across models.

For each sample, s calculate the average (over $i = 1, ..., N$) difference in probabilities under treatment and control group allocation (i.e. for $W^{tr}_{(m)}$ and $W^{c}_{(m)}$), and then find the relevant quantiles. i.e. calculate $\frac{1}{N} \sum_{i=1}^{N} \left(\frac{e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} \right)$ for each draw and find the quantiles.

For example, for a 95% interval, find L_b such that

$$\frac{1}{S} \sum_{s=1}^{S} \mathbb{I} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} \right) < L_b \right] = 0.025$$

and find U_b such that

$$\frac{1}{S} \sum_{s=1}^{S} \mathbb{I} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{tr}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} - \frac{e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}}{1+e^{row_i(W^{c}_{(m)})\underline{\boldsymbol{\mu}}_{(m),s}}} \right) < U_b \right] = 0.975$$

Approximation of CATE Intervals, reducing the dimension of the integral

The dimension reduction can not be applied to the same extent as in the ITE case because $\frac{1}{N} \sum_{i=1}^{N} sig(\alpha_{i,(m,tr)}) - sig(\alpha_{i,(m,c)})$ depends on $2N$ parameters given by $\alpha_{i,(m,tr)}$ and $\alpha_{i,(m,c)})$ for $i = 1, ..., N$.

$$\sum_{m=1}^{M} \left[\int_{-\infty}^{\infty} \mathbb{I} \left(\frac{1}{N} \sum_{i=1}^{N} sig(\alpha_{i,(m,tr)}) - sig(\alpha_{i,(m,c)}) < L_b \right) p(\boldsymbol{\alpha}_{(m)} | \boldsymbol{\psi}_{(m)}, \boldsymbol{\sigma}^{\mathbf{2}}_{(m)}) d\boldsymbol{\alpha}_{(m)} \right] p(\mathcal{T}_m | \boldsymbol{y})$$

[4] For closer approximations to logistic probabilities, this can be replaced by

$$= \sum_{m=1}^{M} \frac{1}{N} \sum_{i=1}^{N} \left[\Phi\left(\frac{\psi_{\alpha,i,(m,tr)}}{\sqrt{\frac{8}{\pi}+\sigma^2_{\alpha,i,(m,tr)}}} \right) - \Phi\left(\frac{\psi_{\alpha,i,(m,c)}}{\sqrt{\frac{8}{\pi}+\sigma^2_{\alpha,i,(m,c)}}} \right) \right] p(\mathcal{T}_m | \boldsymbol{y})$$

where $\boldsymbol{\alpha}_{(m)}$ is a $2N \times 1$ vector if the $\alpha_{i,(m,tr)}$ and $\alpha_{i,(m,c)}$) for $i = 1, ..., N$ and similarly $\boldsymbol{\psi}_{(m)}$ and $\boldsymbol{\sigma}^{\boldsymbol{2}}_{(m)}$ are vectors of the (approximate) means and variances of the elements of $\boldsymbol{\alpha}_{(m)}$. $\boldsymbol{\alpha}_{(m)}$ is multivariate normal, and it is possible to draw from each $\boldsymbol{\alpha}_{(m)}$ to evaluate all M integrals by Monte Carlo, or to draw from the model weighted mixture distribution of the $\boldsymbol{\alpha}_{(m)}$ (i.e. the mixture of multivariate normals). However, this may be generally of a higher dimension than $\underline{\boldsymbol{\mu}}_{(m)}$, depending on the data and selected models. Furthermore, extra calculations are required to obtain the means, variances, and covariances of the elements of $\boldsymbol{\alpha}_{(m)}$. Therefore this might not be computationally more efficient.

C.8 Finding the MAP for Logit BCF

Let the vector of all terminal node parameters be denoted by $\underline{\boldsymbol{\theta}} = [\underline{\boldsymbol{\mu}}^T \underline{\boldsymbol{\tau}}^T]^T$. The Laplace approximation involves a second order Taylor expansion about the Maximum A Posteriori (MAP) estimate

$$\underline{\boldsymbol{\theta}}_{MAP} = \arg\min_{\underline{\boldsymbol{\theta}}} -(\log p(\boldsymbol{y}|W, \underline{\boldsymbol{\theta}}) + \log p(\underline{\boldsymbol{\theta}}))$$

$$= \arg\min_{\underline{\boldsymbol{\theta}}} - \left[\boldsymbol{y}^T W \underline{\boldsymbol{\theta}} - \sum_{i=1}^{N} \log(1 + e^{-W_i \underline{\boldsymbol{\theta}}}) - 0.5b \log(2\pi) + \frac{1}{2} b_\mu \log(a_\mu) + \frac{1}{2} b_\tau \log(a_\tau) - \frac{a_\mu}{2} \underline{\boldsymbol{\mu}}^T \underline{\boldsymbol{\mu}} - \frac{a_\tau}{2} \underline{\boldsymbol{\tau}}^T \underline{\boldsymbol{\tau}} \right]$$

(where b_μ and b_τ are the numbers of terminal nodes in the sums-of-trees represented by $\mu(x)$ and $\tau(x)$ respectively) gives the approximation of the posterior:

$$p(\underline{\boldsymbol{\theta}}|\boldsymbol{y}, W) \approx \mathcal{N}(\underline{\boldsymbol{\theta}}_{MAP}, H^{-1})$$

where H is the Hessian matrix of the negative log posterior (evaluated at the MAP).

$$H = W^T S W + A$$

where A is a diagonal matrix with the first b_μ diagonal elements equal to a_μ and the final b_τ elements equal to a_τ, and $S = diag(p_i(p_i))$ is an $n \times n$ diagonal matrix with diagonal elements determined by the probabilities p_i obtained from the logistic function.

The Hessian and the gradient of the negative posterior probability can be used to obtain an approximation of the MAP. The gradient is:

$$\boldsymbol{g} = W^T(\boldsymbol{p} - \boldsymbol{y}) + \begin{bmatrix} a_\mu \underline{\boldsymbol{\mu}} \\ a_\tau \underline{\boldsymbol{\tau}} \end{bmatrix}$$

where $\boldsymbol{p} = (p_1, ..., p_n)^T$, and $\underline{\boldsymbol{\mu}}$ and $\underline{\boldsymbol{\tau}}$ are the terminal nodes of the sums-of-trees $\mu(x)$ and $\tau(x)$ respectively.

C.9 Tobit-BART-IS Implementation Details

C.9.1 Tobit Posterior and gradients with standard semi-conjugate priors

Chib (1992) used an uninformative prior for Bayesian Tobit. However, here we use the standard BART prior on the terminal node parameters and inverse gamma prior on the variance of the error term.

$$\tau^2 = \sigma^{-2} \sim \Gamma(\frac{\nu}{2}, \frac{\nu\lambda}{2})$$

$$\underline{\boldsymbol{\mu}} \sim N(0, \frac{\sigma^2}{a}) \ , \ or \ \underline{\boldsymbol{\mu}} \sim N(0, \frac{1}{a}\tau^{-2})$$

and the convenient Tobin reparameterization is $(\underline{\boldsymbol{\mu}}, \tau^2) \rightarrow (\boldsymbol{\alpha} = \underline{\boldsymbol{\mu}}\tau, \tau = (\tau^2)^{\frac{1}{2}})$. This gives

$$\alpha = \tau\beta \sim N(0, \frac{1}{a}I)$$

The standard Tobit model framework is

$$y_i^* = row_i(W)\underline{\boldsymbol{\mu}} + \varepsilon_i \ , \ \varepsilon \sim i.i.dN(0, \tau^{-2})$$

$$y_i = \max\{y_i^*, 0\} \ , \ i = 1, ..., n$$

The likelihood is:

$$\ell(\underline{\boldsymbol{\mu}}, \tau^2) = \left[\prod_{i \in C} 1 - \Phi(W_i\underline{\boldsymbol{\mu}}\tau)\right](2\pi)^{-\frac{n_1}{2}}(\tau^2)^{\frac{n_1}{2}}e^{-\tau^2\|y_1 - X_1\underline{\boldsymbol{\mu}}\|^2/2} = \ell_0(\underline{\boldsymbol{\mu}}, \tau^2)\ell_1(\underline{\boldsymbol{\mu}}, \tau^2)$$

or, reparameterized, the likelihood is

$$\ell(\boldsymbol{\alpha}, \tau) = \left[\prod_{i \in C} 1 - \Phi(W_i\boldsymbol{\alpha})\right](2\pi)^{-\frac{n_1}{2}}(\tau^2)^{\frac{n_1}{2}}e^{-\|\tau y_1 - W_1\boldsymbol{\alpha}\|^2/2} = \ell_0(\boldsymbol{\alpha}, \tau)\ell_1(\boldsymbol{\alpha}, \tau)$$

where $c = \{j \ : \ y_j = 0, j = 1, ..., n\}$ (i.e. the set of observations for which the outcome is zero), n_1 is the number of observations for which the outcome is nonzero, y_1 is an $n_1 \times 1$ vector of nonzero outcomes, W_1 is an $n_1 \times b$ matrix of terminal node indicator variables corresponding to nonzero outcomes (y_1). $\|.\|$ is the Euclidean norm.

The log posterior is:

$$\tilde{L}(\underline{\boldsymbol{\mu}}, \tau^2) = \sum_{i \in C} \log[1 - \Phi(row_i(W\underline{\boldsymbol{\mu}}\tau))] - \frac{n_1}{2}\log(2\pi) + \frac{n_1}{2}\log(\tau^2) - \frac{\tau^2}{2}\|y_1 - W_1\underline{\boldsymbol{\mu}}\|^2$$

$$-\frac{b}{2}\log(2\pi) + \frac{b}{2}\log(a\tau) - \frac{a\tau}{2}\underline{\boldsymbol{\mu}}^T\underline{\boldsymbol{\mu}} + \frac{\nu}{2}[\log(2) - \log(\nu\lambda)] - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \left(\frac{\nu}{2} - 1\right)\log(\tau^2) - \frac{\tau^2 2}{\nu\lambda}$$

the reparameterized log posterior is

$$\tilde{L}(\boldsymbol{\alpha}, \tau) = \sum_{i \in C} \log[1 - \Phi(row_i(W\boldsymbol{\alpha}))] - \frac{n_1}{2}\log(2\pi) + \frac{n_1}{2}\log(\tau^2) - \frac{1}{2}(\tau y_1 - W_1\boldsymbol{\alpha})^T(\tau y_1 - W_1\boldsymbol{\alpha})$$

$$-\frac{b}{2}\log(2\pi) + \frac{b}{2}\log(a) - \frac{a}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \frac{\nu}{2}[\log(2) - \log(\nu\lambda)] - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \left(\frac{\nu}{2} - 1\right)\log(\tau^2) - \frac{\tau^2 2}{\nu\lambda}$$

$$\tilde{L}_\alpha = -W_0^T A_0 + W_1^T(\tau Y_1 - W_1\boldsymbol{\alpha}) - a\boldsymbol{\alpha}$$

$$\tilde{L}_\tau = \frac{n_1}{\tau} - Y_1^T(\tau Y_1 - W_1\boldsymbol{\alpha}) + \frac{2(\frac{\nu}{2}-1)}{\tau} - \frac{4\tau}{\nu\lambda}$$

And the Hessian matrix is:

$$\begin{bmatrix} -W_0^T B_0 W_0 - W_1^T W_1 - aI_b & W_1^T Y_1 \\ Y_1^T W_1 & -\frac{n_1}{\tau^2} - Y_1^T Y_1 - \frac{2(\frac{\nu}{2}-1)}{\tau^2} - \frac{4}{\nu\lambda} \end{bmatrix}$$

where $A_0 = vec(\lambda_i)$, $B_0 = diag(\lambda_i(\lambda_i - W_i\boldsymbol{\alpha}))$, $\lambda_i = \frac{\phi(W_i\boldsymbol{\alpha})}{1-\Phi(W_i\boldsymbol{\alpha})}$

The negative of the gradient and the negative of the Hessian above can be used to obtain the MAP by Newton's algorithm (minimizing the negative of the log posterior). Algorithm 6 outline's Newton's method for minimizing the negative log-likelihood

---

Require parameter value, e.g. $a = 0.01$

Initialize $\begin{bmatrix} \boldsymbol{\alpha} \\ \tau \end{bmatrix} = \begin{bmatrix} \mathbf{0}_b \\ 1 \end{bmatrix}$, where $\mathbf{0}_b$ is a zero vector of length b.

repeat

- $\lambda_i = \frac{\phi(W_i\boldsymbol{\alpha})}{1-\Phi(W_i\boldsymbol{\alpha})}$ for $i \in C$
- $A_0 = diag(\lambda_i)$
- $B_0 = diag(\lambda_i(\lambda_i - W_i\boldsymbol{\alpha}))$
- $\boldsymbol{g} = -\begin{bmatrix} W_0^T A_0 + W_1^T(\tau Y_1 - W_1\boldsymbol{\alpha}) - a\boldsymbol{\alpha} \\ \frac{n_1}{\tau} - Y_1^T(\tau Y_1 - W_1\boldsymbol{\alpha}) + \frac{2(\frac{\nu}{2}-1)}{\tau} - \frac{4\tau}{\nu\lambda} \end{bmatrix}$
- $H = -\begin{bmatrix} -W_0^T B_0 W_0 - W_1^T W_1 - aI_b & W_1^T Y_1 \\ Y_1^T W_1 & -\frac{n_1}{\tau^2} - Y_1^T Y_1 - \frac{2(\frac{\nu}{2}-1)}{\tau^2} - \frac{4}{\nu\lambda} \end{bmatrix}$
- $\underline{\boldsymbol{\mu}}_{new} = \underline{\boldsymbol{\mu}}_{old} - H^{-1}\boldsymbol{g}$

until *convergence*;

Algorithm 6: Newton's method for obtaining the mode (MAP) of the Tobit parameters

Alternatively, a quasi-Newton algorithm, such as the L-BFGS algorithm can be applied. The standard Laplace approximation for the marginal likelihood is:

$$p(\boldsymbol{y}|W_m, \mathcal{T}_{(m)}) = e^{\tilde{L}(\boldsymbol{\alpha}_{MAP}, \tau_{MAP})}(2\pi)^{b/2}|H_{MAP}|^{-1/2}$$

where H_{MAP} is the Hessian matrix of the negative log likelihood evaluated at the MAP parameter values. The log of the marginal likelihood approximation is:

$$log(p(\boldsymbol{y}|W_m, \mathcal{T}_{(m)})) = \tilde{L}(\boldsymbol{\alpha}_{MAP}, \tau_{MAP}) + \frac{b}{2}\log(2\pi) - (\frac{1}{2})log(|H_{MAP}|)$$

A more accurate approximation can be obtained using the double Laplace approximation methods of Tierney & Kadane (1986), as outlined by Chib (1992).

The Laplace approximation gives a multivariate normal approximation for the posterior distribution of the parameters:

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \tau \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\alpha}_{MAP} \\ \tau_{MAP} \end{bmatrix} \ , \ H_{MAP}^{-1} \right)$$

and the approximate marginal posterior distribution for $\boldsymbol{\alpha}$ is:

$$\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\alpha}_{MAP}, H_{\boldsymbol{\alpha},MAP})$$

where $H_{\boldsymbol{\alpha},MAP} = W_0^T B_0 W_0 + W_1^T W_1 + aI_b$ is the submatrix of the Hessian of the negative log likelihood corresponding to $\boldsymbol{\alpha}$ evaluated at the MAP parameter values.

The posterior predictive mean probability that the outcome $y_\star$ is equal to one is:

$$p(y_\star = 1 | row_\star(W), \mathcal{T}_{(m)}) = \int [1 - \Phi(row_\star(W)\boldsymbol{\alpha})] p(\boldsymbol{\alpha} | row_\star(W), \mathcal{T}_{(m)}) d\boldsymbol{\alpha}$$

where $row_\star(W)$ is the row vector of terminal node indicator variables for the new observation. The integral can be re-written as a one-dimensional integral by considering $\psi = row_\star(W)\boldsymbol{\alpha}$, $\psi_{MAP} = row_\star(W)\boldsymbol{\alpha}_{MAP}$, and $\sigma_\psi^2 = row_\star(W) H_{\boldsymbol{\alpha},MAP} row_\star(W)^T$, which is approximately normally distributed $\psi \sim \mathcal{N}(\psi_{MAP}, \sigma_\psi^2)$.

$$p(y_\star = 1 | row_\star(W), \mathcal{T}_{(m)}) = \int [1 - \Phi(\psi)] p(\psi | \psi_{MAP}, \sigma_\psi^2) d\psi$$

$$= 1 - \int \Phi(\psi) p(\psi | \psi_{MAP}, \sigma_\psi^2) d\psi = 1 - \Phi\left(\frac{\psi_{MAP}}{1 + \sigma_\psi^2} \right)$$

and the average over models $\mathcal{T}_{(1)}, ..., \mathcal{T}_{(M)}$ is:

$$p(y_\star = 1) = 1 - \frac{1}{M} \sum_{m=1}^{M} \Phi\left(\frac{\psi_{MAP,(m)}}{1 + \sigma_{\psi,(m)}^2} \right) p(\mathcal{T}_{(m)} | \boldsymbol{y}, \boldsymbol{X})$$

where $\psi_{MAP,(m)}$ and $\sigma_{\psi,(m)}$ are calculated using $\boldsymbol{\alpha}_{MAP,(m)}$ and $H_{\boldsymbol{\alpha},MAP,(m)}$, i.e. the MAP parameter values and Hessian evaluated at the MAP values for model (m).

Intervals for the predictive probability that $y_\star = 1$ can be obtained as follows. If the lower confidence probability is $lower\_prob = 0.025$ (i.e. for a 85% interval), then the lower bound for the predictive probability L_b satisfies:

$$\frac{1}{M} \sum_{i=1}^{M} \int_{-\infty}^{\infty} \mathbb{I}\{1 - \Phi(\psi) < L_b\} p(\psi | \psi_{MAP}, \sigma_\psi^2) d\psi \, p(\mathcal{T}_{(m)} | \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{M} \sum_{i=1}^{M} \int_{\Phi^{-1}(1-L_b)}^{\infty} p(\psi | \psi_{MAP}, \sigma_\psi^2) d\psi \, p(\mathcal{T}_{(m)} | \boldsymbol{y}, \boldsymbol{X})$$

$$= 1 - \frac{1}{M} \sum_{i=1}^{M} \int_{-\infty}^{\Phi^{-1}(1-L_b)} p(\psi | \psi_{MAP}, \sigma_\psi^2) d\psi \, p(\mathcal{T}_{(m)} | \boldsymbol{y}, \boldsymbol{X})$$

$$= 1 - \frac{1}{M} \sum_{i=1}^{M} \Phi\left(\frac{\Phi^{-1}(1 - L_b) - \psi_{MAP}}{\sigma_\psi} \right) \ p(\mathcal{T}_{(m)} | \boldsymbol{y}, \boldsymbol{X}) = lower\_prob$$

or equivalently

$$\frac{1}{M}\sum_{i=1}^{M}\Phi\left(\frac{\Phi^{-1}(1-L_b)-\psi_{MAP}}{\sigma_\psi}\right)\ p(\mathcal{T}_{(m)}|\boldsymbol{y},\boldsymbol{X})=1-lower\_prob$$

and similarly the upper bound U_b is the number such that $1-\frac{1}{M}\sum_{i=1}^{M}\Phi\left(\frac{\Phi^{-1}(1-U_b)-\psi_{MAP}}{\sigma_\psi}\right)\ p(\mathcal{T}_{(m)}|\boldsymbol{y},\boldsymbol{X})=1-upper\_prob$. Therefore L_b and U_b can be obtained by a root-finding algorithm (e.g. bisection).

Alternatively, Monte Carlo draws can be made from the mixture $\boldsymbol{\alpha}\sim\frac{1}{M}\sum_{i=1}^{M}\mathcal{N}(\boldsymbol{\alpha}_{MAP,(m)},H_{\boldsymbol{\alpha},MAP,(m)})$, and for each draw the probability $[1-\Phi(row_\star(W)\boldsymbol{\alpha})]$ can be calculated. Then the mean and quantiles across many Monte Carlo draws can be used for the predictive probability and interval for the predictive probability.