1 Introduction

Mediation is the process through which a treatment A affects an outcome Y through an intermediate variable (or mediator) M. A mediator M is a variable that occurs in time between the treatment A and outcome Y. It is on the causal pathway between the treatment and outcome, and hence may be affected by the treatment and may also affect the outcome itself. The main goal of mediation analysis is to disentangle the total effect of the treatment on the outcome into direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001). The indirect effect represents the effect of the treatment on the outcome operating through the mediator, while the direct effect represents the effect of all other causal mechanisms which do not operate through the mediator. include a diagram?

Consider an example where researchers are interested in whether smoking affects medical expenditures (further described in Section 6). Smoking is a binary treatment A and medical expenditures is the outcome of interest Y. Researchers may not be interested in only the direct effect of smoking on medical expenditures, but also in the indirect effect of smoking on medical expenditures through the pathway of overall perceived health (M). Perhaps smoking itself does not directly cause higher medical expenditures, but rather smoking indirectly causes higher medical expenditures because it decreases patients' overall perceived health. In this example, overall perceived health is considered the mediator.

Mediation analysis has been widely studied across many scientific fields including epidemiology, medicine, economics, and the social sciences (MacKinnon and Dwyer, 1993; Rubin, 2004; MacKinnon, 2008; J. M. Albert, 2008; Imai, Keele, and Tingley, 2010; VanderWeele, 2016). Early literature has focused largely on structural equation models (SEMs) to quantify mediation effects in terms of model coefficients (Wright, 1921). In particular, linear structural equation models (LSEMs) have been widely used (Baron and Kenny, 1986; MacKinnon and Dwyer, 1993; MacKinnon, 2008). However, these models have their drawbacks in that the mediation effects are defined with respect to a particular parametric model, often linear systems. Thus, the identification and causal interpretations of these effects become tied to the choice and correct specification of such a parametric model (Imai, Keele, and Tingley, 2010). However, Imai, Keele, and Tingley, 2010 later showed that under sequential ignorability, the average causal mediation effects can be identified nonparametrically, allowing for a general estimation procedure under nonlinear conditions. This nonparametric sequential ignorability assumption states that (i) the treatment is independent of all potential values of the outcome and mediator conditional on the covariates and (ii) the observed mediator is independent of all potential outcomes conditional on the observed treatment and covariates. These identification conditions allow researchers to causally interpret mediation effects regardless of the specific models used.

Since the early literature, mediation has been firmly grounded in the language of causal inference in terms of the potential outcomes framework (Imai, Keele, and Yamamoto, 2010; VanderWeele, 2016), formalizing notions of mediation effects regardless of the specific statistical model used. The potential outcomes framework quantifies causal effects in terms of hypothetical treatments or interventions. That is, the difference in the outcome were the mediator and/or treatment changed from the value realized under the treatment to that under the control. Such a framework allows for a causal interpretation of direct and indirect effects without complications of parametric assumptions. In this paper, we use the potential outcomes framework for causal mediation analysis from a Bayesian nonparametric point of view.

In particular, we use Bayesian additive regression trees (BART, Chipman, George, and McCulloch, 2010) as our underlying model of choice. BART has been seen to perform particularly well in causal inference, successfully inferring heterogeneous and average treatment effects (Hill, 2011; Wendling et al., 2018; Dorie et al., 2019). For outcome Y_i, binary treatment A_i, and control variables X_i for observation i, consider an unknown function μ that predicts Y_i

$$Y_i(a) = \mu(X_i, a) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{1}$$

The treatment effect of receiving the treatment versus control can thus be denoted by

$$\tau(X_i) = \mu(X_i, 1) - \mu(X_i, 0). \tag{2}$$

Given BART's strong predictive performance, Hill, 2011 suggests using a BART prior for $\mu(X_i, a)$ directly to flexibly model the outcome and hence obtain treatment effect estimates.

Hahn, Murray, and Carvalho, 2020 note that successful predictive modeling depends largely on careful regularization, and extend the work of Hill, 2011 by noting two shortcomings of regularization in causal inference: (i) strong confounding can lead to highly biased causal estimates (regularization-induced confounding, Hahn, Carvalho, et al., 2018) and (ii) naive priors can lead to high variance estimates. To mitigate (i), they develop a prior for the outcome that depends explicitly on an estimate of the propensity score $\widehat{\pi}$ as a 1-dimensional summary of the covariates. To address (ii), they reparameterize the regression as a sum of two functions. One of which models the prognostic effect of the control variables while the other represents the treatment effect exactly. Specifically, they propose the *Bayesian causal forests*

$$Y_i(a) = \mu(X_i, \widehat{\pi}) + a\tau(X_i) \tag{3}$$

with BART priors on μ and τ. μ is the prognostic function which is dependent on the propensity score $\widehat{\pi}$ and τ is the treatment effect. This model isolates the treatment effect τ in particular, allowing for it to be regularized directly and independently from μ.

Linero and Zhang, 2022 further extend this work by placing this type of model in the context of mediation. Linked to the concept of regularization-induced confounding, they also introduce the concept of prior dogmatism. Prior dogmatism induces regularization-induced confounding by giving a strong prior preference to encourage the model to attribute causal effects on the outcome as being due to the treatment rather than the confounders. To address this issue, they include "clever covariates" $\widehat{m}_{ai}$ $(a = 0, 1)$ into the outcome model. These clever covariates are analagous to the propensity score estimate and are predictions of the mediator M by the predictors X. For treatment value a and mediator value m, they introduce the *Bayesian causal mediation forests* specified by

$$Y_i(a, m) = \mu_y(m, a, X_i) + \epsilon_i \tag{4}$$

$$M_i(a) = \mu_m(a, X_i) + \epsilon_i \tag{5}$$

where μ_y and μ_m are given BART priors and the clever covariates $\widehat{m}_{0i}$ and $\widehat{m}_{1i}$ are included as predictors in the BART model for the outcome $Y_i(a, m)$. This model also stratifies the treatment A into treatment $(A = 1)$ and control $(A = 0)$, specifying a separate model for $\mu_y(m, 0, X_i), \mu_y(m, 1, X_i), \mu_m(0, X_i)$, and $\mu_m(1, X_i)$ rather than treating A as another predictor. Move after potential outcomes section?

The goal of this work is to extend the Bayesian causal mediation forests (Linero and Zhang, 2022) into a form that more closely resembles the Bayesian causal forests (Hahn, Murray, and Carvalho, 2020) so that both the direct and indirect effects can be regularized directly and independently in order to allow researchers to control the amount of heterogeneity of treatment effects.

In Section 2, we discuss the potential outcomes framework for causal mediation analysis and how it can be used to define mediation effects of interest. In Section 3, we state our assumptions which are needed to identify the mediation effects. Section 4 gives an overview of our underlying predictive model, Bayesian Additive Regression Trees (BART). Section 5 outlines our approach to the decomposition and targeted regularization of the direct and indirect effects as well as details on posterior computations. Section 6 contains an analysis of our method on the Medical Expenditures (MEPS) panel survey data. Section 7 conducts a simulation study to show the performance of our model in terms of converage and bias. Lastly, Section 8 concludes with a discussion and possible extensions.

2 Potential Outcomes Framework

We will use the potential outcomes framework for causal mediation analysis (Rubin, 2004; Imai, Keele, and Tingley, 2010). For individuals $i = 1, ..., n$ and treatment $A_i \in \{0, 1\}$, define the potential outcome $M_i(a)$ as the value of the mediator that would have been observed had the individual received treatment a. Note that for each individual, only one of $M_i(0)$ or $M_i(1)$ is actually observed. For treated individuals ($A_i = 1$), $M_i(0)$ can be thought of as the counterfactual. That is, the value of the mediator that would have been observed had the individual been untreated instead. Similarly, the potential outcome $Y_i(a, m)$ is the value of the outcome that would have been observed had the individual received treatment a and had a mediator at level m. For example, $Y_i(0, M_i(1))$ is the value of the outcome that would have been observed if the individual was not treated and had a value of the mediator at the same level they would have had if they were treated. The potential outcomes framework implies that not all mediators and outcomes can be observed. Hence, the causal assumption of *consistency* is made throughout, implying that it is only possible to observe the mediator $M_i = M_i(A_i)$ and the outcome $Y_i = Y(A_i, M_i(A_i))$. Note also that this notation implies that there is no interference between units, known as the *Stable Unit Treatment Value Assumption (SUTVA)*. This is because the potential outcome Y_i for individual i is defined only in terms of a treatment a potentially received by individual i rather than any other individual j.

We can use these potential outcomes to define causal estimates of interest. In causal mediation analysis, we are particularly interested in estimating the natural direct and natural indirect effects (Pearl, 2001; Robins and Greenland, 1992). The *natural direct effect* is defined as

$$\zeta(a) = E[Y_i\{1, M_i(a)\} - Y_i\{0, M_i(a)\}] \tag{6}$$

and the *natural indirect effect* is defined as

$$\delta(a) = E[Y_i\{a, M_i(1)\} - Y_i\{a, M_i(0)\}]. \tag{7}$$

The natural direct effect isolates the effect of the treatment while keeping the potential mediator fixed and can be interpreted as the effect that the treatment A_i has directly on the outcome Y_i. On the other hand, the natural indirect effect isolates the effect of the potential mediator in response to different treatment values while keeping the treatment fixed and can be interpreted as the effect that the treatment A_i has indirectly on the outcome Y_i through the mediator M_i. The total effect of the treatment on the outcome is a sum of the direct and indirect effects. The total effect can be defined as

$$\tau = \zeta(0) + \delta(1) = \zeta(1) + \delta(0) = E[Y_i\{1, M_i(1)\} - Y_i\{0, M_i(0)\}]. \tag{8}$$

3 Assumptions

Following Imai, Keele, and Yamamoto, 2010, we assume the following sequential ignorability assumptions, allowing for the identification and valid inference of the direct and indirect effects defined above.

1. $\{Y_i(a', m), M_i(a)\} \perp\!\!\!\perp A_i \mid X_i = x$ for $a, a' = 0, 1$ and all $x \in \mathcal{X}$.

2. $Y_i(a', m) \perp\!\!\!\perp M_i(a) \mid A_i = a, X_i = x$ for $a, a' = 0, 1$ and all $x \in \mathcal{X}$.

3. $\Pr(A_i = a \mid X_i = x) > 0$ and $p(M_i(a) = m \mid A_i = a, X_i = x)$ for $a = 0, 1$ and all $x \in \mathcal{X}$ and $m \in \mathcal{M}$.

The first assumption states that given the covariates, the treatment assignment is ignorable, that is independent of potential outcomes and potential mediators. This assumption is automatically satisfied when individuals are randomly assigned to treatment and control groups. However, it is not guaranteed to hold in observational studies, in which case researchers often collect as many covariates as possible so that treatment assignment ignorability is plausible after the differences in covariates between treatment groups are accounted for. The second assumption states that given the observed treatment and covariates, the mediator is ignorable, that is independent of potential outcomes. This assumption, however, is not guaranteed to hold even in randomized experiments. In general, it cannot be directly tested from the data and should be assessed through a sensitivity analysis. The third assumption is a positivity assumption for the treatment and mediator, stating that the probability of receiving the treatment and control should be nonzero.

Under the above sequential ignorability assumptions, we can identify the distribution of any counterfactual outcome $Y_i(a, M_i(a))$ nonparametrically. That is, we can identify

$$f(Y_i(a, M_i(a')) \mid X_i = x) = \int_{\mathcal{M}} f(Y_i \mid M_i = m, A_i = a, X_i = x) \, dF_{M_i}(m \mid A_i = a', X_i = x) \tag{9}$$

for any $x \in \mathcal{X}$ and $a, a' = 0, 1$ (Theorem 1, Imai, Keele, and Tingley, 2010). This allows us to make inferences about unobserved counterfactuals (left-hand side) using observed outcomes and mediators (right-hand side). And given that (9) is not dependent on a specific model, we can estimate causal mediation effects in a flexible way. Estimating the potential outcome $Y_i(a, M_i(a))$, and hence the direct effect $\zeta(a)$ and indirect effect $\delta(a)$, therefore only requires specification of $f(y \mid m, a, x)$ and $f(m \mid a, x)$ along with a distribution for the covariates.

4 A Review of Bayesian Additive Regression Trees

We will use the Bayesian Additive Regression Trees (BART) proposed by Chipman, George, and McCulloch, 2010. Consider an unknown function f that predicts an output Y_i using a vector of inputs $\boldsymbol{x_i}$

$$Y_i = f(\boldsymbol{x_i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{10}$$

We can model $f(\boldsymbol{x_i})$ by a sum of m regression trees $f(\boldsymbol{x_i}) \approx r(\boldsymbol{x_i}) = \sum_{j=1}^{m} g(\boldsymbol{x_i}; T_j, M_j)$. T_j is a binary decision tree consisting of a set of interior node decision rules as well as a set of terminal nodes. The decision rules are binary splits of the predictor space. $M_j = \{\mu_{j1}, \mu_{j2}, ...\mu_{jb_j}\}$ is a set of parameter values associated with each of the b_j terminal nodes of tree T_j. Each x_i is associated with a single terminal node of T_j and is then assigned the μ_{jk} value associated with that terminal node. Hence, for a given tree T_j and terminal node parameters M_j, $g(\boldsymbol{x_i}, T_j, M_j)$ denotes the function that assigns a $\mu_{jk} \in M_j$ to $\boldsymbol{x_i}$. Thus (10) can be approximated by the sum-of-trees model

$$Y = r(\boldsymbol{x_i}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{11}$$

Under (11), $E(Y_i|\boldsymbol{x_i})$ equals the sum of all the terminal node u_{jk}'s assigned to $\boldsymbol{x_i}$ by the $g(\boldsymbol{x_i}; T_j M_j)$'s.

It is also necessary to specify a prior over all parameters of the sum-of-trees model, i.e., (T_j, M_j) for all j. This prior should regularize the fit by keeping individual tree effects from being disproportionately influential. The prior consists of two components: a prior for each tree T_j and a prior on the terminal nodes $M_j|T_j$ where $p(T_j, M_j) = p(M_j|T_j) \times p(T_j)$. We can write $r \sim \text{BART}(p(T_j), p(M_j|T_j), m)$. That is, r has a BART prior with m trees, a tree prior $p(T_j)$, and a terminal node prior $p(M_j|T_j)$. The posterior distribution is computed by Markov chain Monte Carlo.

The prior $p(T_j)$ is specified by three aspects: the probability that a node is interior, the distribution of the splitting variable assignments at each interior node, and the distribution of the splitting rule assignment in each interior node conditional on the splitting variable. The probability that a node at depth d is interior is

$$\alpha(1+d)^{-\beta}, \quad \alpha \in (0,1), \beta \in [0, \infty) \tag{12}$$

with $\alpha = 0.95$ and $\beta = 2$ being the default that favors small trees. The distribution of the splitting variable assignments at each interior node and the distribution of the splitting rule assignment in each interior node conditional on the splitting variable are both given a uniform prior.

For the prior on the terminal nodes, note that $p(M_j|T_j) = \prod_{k=1}^{b_j}[\mu_{jk}|T_j]$. First, shift and rescale Y so that the observed transformed y values range from $y_{min} = -0.5$ to $y_{max} = 0.5$. Then, we use the prior

$$\mu_{jk}|T_j \sim N(0, \sigma_\mu^2) \quad \text{where } \sigma_\mu = \frac{0.5}{k\sqrt{m}} \tag{13}$$

for a suitable value of k with default $k = 2$. Note that this prior shrinks the terminal node values μ_{jk} towards zero and applies greater shrinkage as the number of trees m is increased, ensuring that each tree is a weak learner in the ensemble of trees.

5 BART for Mediation

We will use BART in modeling the mediator given the treatment and covariates as well as the outcome given the treatment, covariates, and mediator. That is, we model components of $[M_i|A_i, \boldsymbol{X_i}]$ and $[Y_i|M_i, A_i, \boldsymbol{X_i}]$ using BART. These models are discussed in the context of binary, ordinal, and continuous mediators.

For continuous and ordinal mediators, the outcome and mediator models are specified as follows.

$$Y_i(a, m) = \mu(X_i) + a\zeta(X_i) + md(X_i) + \epsilon_i \tag{14}$$

$$M_i(a) = \mu_m(X_i) + a\tau_m(X_i) + \varepsilon_i \tag{15}$$

where we use BART priors for μ, ζ, d, μ_m, and τ_m. τ_m is dependent on an estimate of the propensity score $\widehat{\pi} = P(A_i = 1|X_i = x)$ while μ, z, and d are dependent on the clever covariates $\widehat{m}_{0i}$ and $\widehat{m}_{0i}$ where $\widehat{m}_{ai}$ $(a = 0, 1)$ estimates $E(M_i \mid A_i = a, X_i = x)$. The natural direct effect is exactly $\zeta(x)$ and the natural indirect effect is computed as $\delta(x) = d(x)\tau_m(x)$ respectively.

Binary mediators can also be easily specified with a slight modification of the mediator model.

$$Y_i(a, m) = \mu(X_i) + a\zeta(X_i) + md(X_i) + \epsilon_i \tag{16}$$

$$M_i(a) = \Phi(\mu_m(X_i) + a\tau_m(X_i)) + \varepsilon_i \tag{17}$$

We again use BART priors for μ, ζ, d, μ_m, and τ_m. The model for M_i is a probit regresssion of M_i, where we use truncated normal latent variables z_i (J. H. Albert and Chib, 1993) such that

$$z_i \mid M_i, \mu_m, \tau, A, X_i \sim \begin{cases} N\{\mu_m(X_i) + a\tau_m(X_i), 1\}\ \mathbb{I}(-\infty, 0) & \text{if } M_i(a) = 0 \\ N\{\mu_m(X_i) + a\tau_m(X_i), 1\}\ \mathbb{I}(0, \infty) & \text{if } M_i(a) = 1 \end{cases}$$

so that we can use the continuous outcome BART model of Section 4. Under this model, the natural direct effect is again $\zeta(x)$ and the natural indirect effect is computed as $\delta(x) = d(x)[\Phi(\mu_m(x)+\tau_m(x)) - \Phi(\mu_m(x))]$.

This parameterization allows us to isolate the components μ, z, d, μ_m, and τ_m and apply differing amounts of regularization to them. Hence, we can focus exactly on the natural direct or indirect effects in order to exert direct control on their estimation process.

5.1 Posterior Computation

To identify heterogeneous treatment effects, individual natural direct and indirect effects can simply be estimated by

$$\widehat{\zeta}_i(a) = \zeta^*(X_i) \qquad \text{and} \tag{18}$$

$$\widehat{\delta}_i(a) = d^*(X_i)\tau_m^*(X_i) \tag{19}$$

where z^*, d^*, and τ_m^* are sampled from their respective BART posteriors.

For average effects, note that the marginal distribution of $Y_i(a, M_i(a'))$ is given by

$$f(Y_i(a, M_i(a')) = \int f(Y_i(a, M_i(a')) \,|\, X_i = x)\, dF_X. \tag{20}$$

Hence, it is necessary to specify a model for the distribution of the covariates F_X. Often, when this distribution is not modeled explicitly, the empirical distribution is used as an estimate. The empirical distribution is modeled by a multinomial distribution with a fixed weight of $\frac{1}{n}$ for each covariate. The empirical distribution of the confounders can be written as $f(x) = \sum_{i=1}^n \omega_i \delta_{x_i}$ where δ_{x_i} is a point mass at x_i and $\boldsymbol{\omega}_i = \frac{1}{n}$, where $\omega_i \geq 0$ and $\sum_{i=1}^n \omega_i = 1$.

An alternative to the empirical distribution is the Bayesian bootstrap (Rubin, 1981). The Bayesian bootstrap is similar to the empirical distribution, but rather than using weights $\boldsymbol{\omega} = (\frac{1}{n}, ..., \frac{1}{n})$, the weights $\boldsymbol{\omega} = (\omega_1, ..., \omega_n) \sim \text{Dirichlet}(1, ..., 1)$ are used instead. Under the Bayesian bootstrap, the counterfactual means are identified as

$$E[Y_i(a, M_i(a')] = \sum_i \omega_i \int y\, f(Y_i \,|\, M_i = m, A_i = a, X_i = x) f(M_i \,|\, A_i = a', X_i = x)\, dm. \tag{21}$$

Hence, under sequential ignorability, we can estimate the average natural direct and indirect effects by

$$\widehat{\zeta}(a) = \sum_i \omega_i \, \zeta^*(X_i) \qquad \text{and} \tag{22}$$

$$\widehat{\delta}(a) = \sum_i \omega_i \, d^*(X_i) \tau_m^*(X_i). \tag{23}$$

6 Medical Expenditures Panel Survey Data

As an illustration of our model, we consider the medical expenditures panel survey dataset. The goal of this survey was to examine the effect of smoking on medical expenditures. For our purposes, we want to examine both the direct effect of smoking on medical expenditures as well as the indirect effect of smoking on medical expenditures through the mediator of overall perceived health. Therefore, the outcome Y is the natural logarithm of medical expenditures, the treatment A is whether or not an individual smokes (0: non-smoke, 1: smoke), and M is an ordinal measure of overall perceived health (1: excellent, 2: very good, 3: good, 4: fair, 5: poor). Our model also includes the following patient attributes as confounders:

- `age:` age in years at the time of the survey
- `bmi:` bmi at the time of the survey
- `education_level:`
- `income:`
- `poverty_level:`
- `region:`
- `sex:` male or female
- `marital_status:`
- `race:`
- `seatbelt:`

We fit our model using a propensity score $\widehat{\pi}$ and clever covariates $\widehat{m}_0$ and $\widehat{m}_1$. The posterior distribution of the average direct and indirect effect is shown in Figure 1. We see that there is evidence of an average mediation effect in this dataset.

To further explore heterogeneity in the indirect effect and identify covariate importance and interactions, we use posterior summarization to better interpret our model. Posterior summarization was introduced by

Woody, Carvalho, and Murray, 2021 as a way to understand how nonparametric models make predictions by creating parsimonious, interpretable summaries of complex models. A post-hoc investigation is conducted on the fitted model using lower-dimensional surrogates as summaries in order to answer relevant inferential questions. For a generic regression model given by $E[y|x] = f(x)$ with observations y and covariates x, posterior summarization allows us to explore the posterior for f with interpretable explanations of model behavior. Consider the objective function

$$\mathcal{L}(f, \gamma, \tilde{X}) = d(f, \gamma, \tilde{X}) + q_\lambda(\gamma) \tag{24}$$

where $d(\cdot, \cdot, \tilde{X})$ measures the discrepancy in prediction between the original model f and the parsimonious summary γ over some $\tilde{n}$ specified covariate locations of interest $\tilde{X}$, and q_λ is an optional penalty function governed by parameters λ. Then, the summary is the function minimizing this objective, i.e.,

$$\gamma(x) = \arg\min_{\gamma' \in \Gamma} \mathcal{L}(f, \gamma', \tilde{X}) \tag{25}$$

where we use the posterior samples for f. The optimal point estimate for the summary minimizes the posterior expected loss, i.e.,

$$\hat{\gamma}(x) := \arg\min_{\gamma' \in \Gamma} E[\mathcal{L}(f, \gamma', \tilde{X})|Y, X]. \tag{26}$$

We take $d(\cdot, \cdot, \tilde{X})$ to be the squared-error and let the posterior mean $\widehat{f}$ take place of f.

We project the indirect effect of our model onto a single regression tree as well as a generalized additive model (GAM). The results under a regression tree are shown in Figure 2. The tree indicates that for the indirect effect, the most predictive covariates in our model are race, age, and sex with higher order interactions present between them. Figure 3 shows the average indirect effect within the five subgroups in the terminal nodes of Figure 2. In general, those who are white, over the age of 32, and male have the greatest indirect effect through the mediator of overall perceived health. The results under a GAM are shown in Figure 4. While not able to identify higher order interactions as in the regression tree, we can see that the indirect effect is significantly higher for individuals over the age of approximately 35.

To measure the adequacy of the summary function approximation to the actual regression function, we show the R^2 for the regression tree and GAM in Figure 5 as a measure of the predictive variance in the original model explained by the summarization. For the posterior average of the indirect effect, both a regression tree and GAM explain approximately 95% of the variance in the original model.

7 Simulation

We conduct a simulation study to demonstrate the reliability of our model. We use a data generating process where covariate and treatment assignment are directly sampled from the MEPS dataset and the mediator and outcomes are simulated based on applying the model in Section 5 with known true values of μ_m, ζ, d, μ_m, and τ_m. These true values are taken from an initial fit of the model to the entire dataset. Hence, the mediator and outcome models are defined as functions of known variables and the true direct and indirect effects are known as well. Each simulation setting is replicated 200 times for an $N = 8215$ training set and $N = 8215$ out of sample test set. We fit our model to each simulated dataset and measure the mean point estimate, the quantiles and width of a 95% credible interval, and whether or not the interval captures the true parameter for each replication. Using the 200 replications, we also measure the root mean square error, absolute bias, the average width of the intervals, and the coverage probability.

8 Discussion

This paper introduces a model to separately identify and regularize the natural direct and indirect effects of Bayesian causal mediation forests. The mediation effects can be simply identified as products of coefficients in a similar way to LSEMs. This allows researchers to have a general modeling framework for causal mediation that does not rely on an underlying parametric model to define mediation effects. Under the sequential ignorability assumptions, we demonstrate our model on the MEPS dataset and conduct a simulation study to show the reliable performance of our model.

8.1 Extensions

- ordinal mediator?
- treatment/mediator interaction?

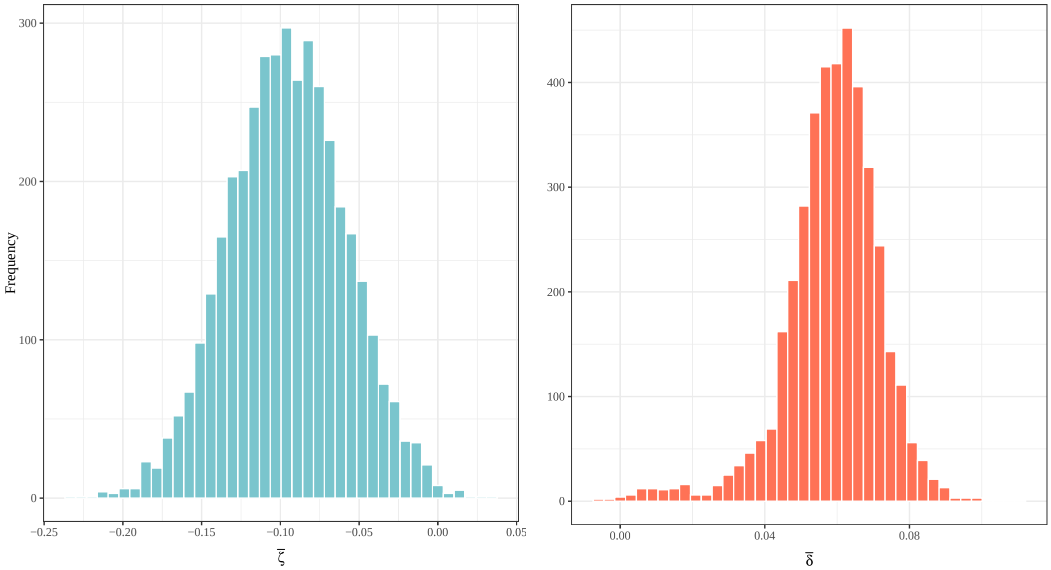

Figure 1: Posterior distribution for the average direct effect $\overline{\zeta}$ and average indirect effect $\overline{\delta}$.

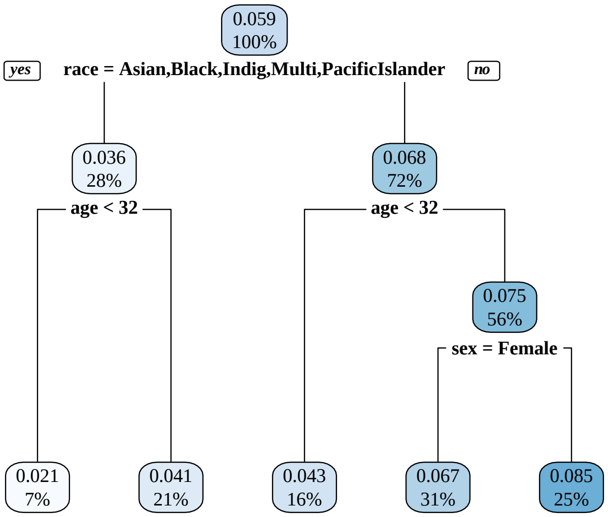

Figure 2: Posterior summarization of the indirect effect using a single regression tree.

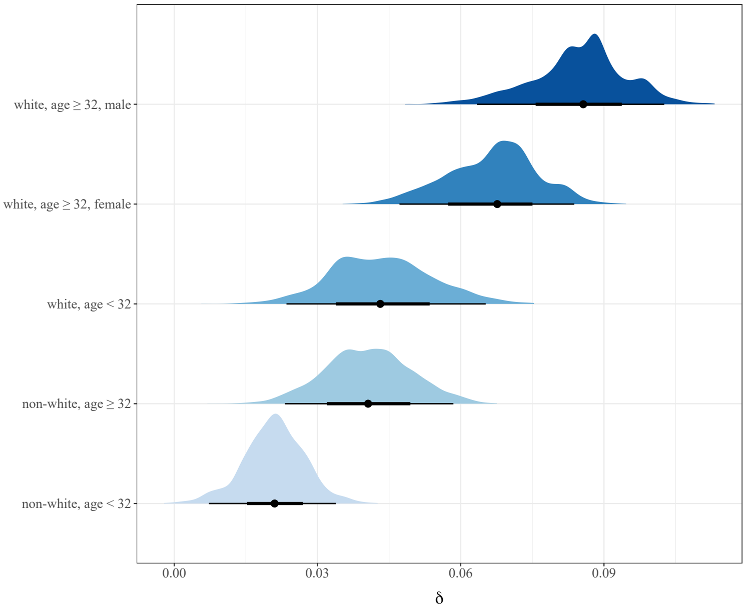

Figure 3: Posterior density for the average indirect effect within subgroups from Figure 2.

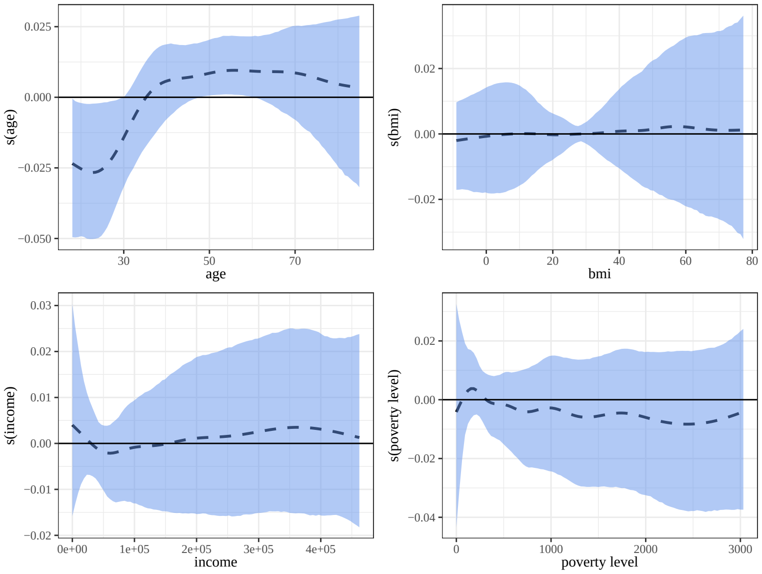

Figure 4: Posterior summarization of the indirect effect using a GAM. The respective function for each covariate is shown. The dashed line gives the posterior mean and the shaded area gives a posterior 95% credible interval.

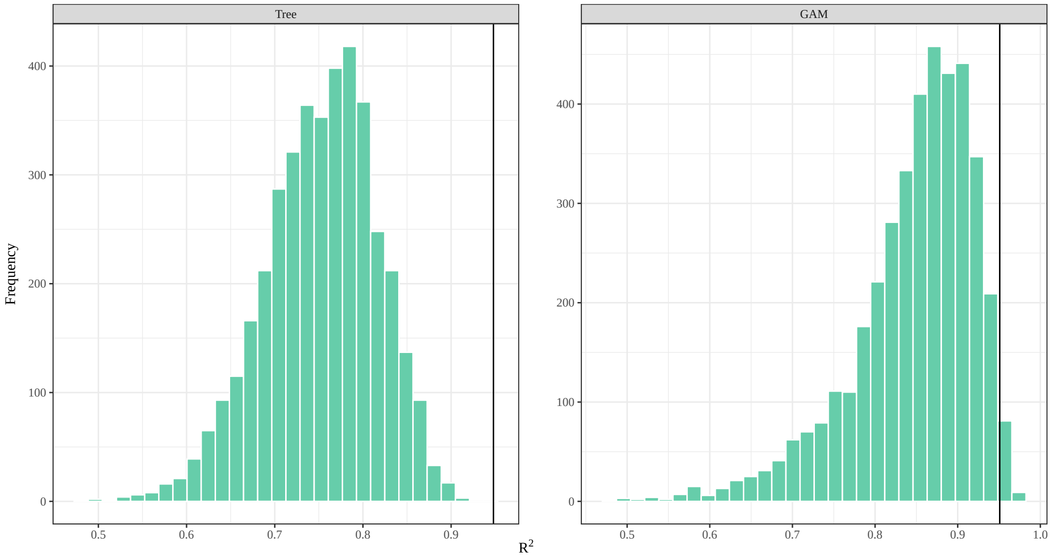

Figure 5: R^2 to quantify the sufficiency of the summarization with a regression tree and a GAM. The black line indicates the R^2 for the posterior average of the indirect effect.

References

Albert, James H and Siddhartha Chib (1993). "Bayesian analysis of binary and polychotomous response data". In: *Journal of the American Statistical Association* 88.422, pp. 669–679.

Albert, Jeffrey M (2008). "Mediation analysis via potential outcomes models". In: *Statistics in Medicine* 27.8, pp. 1282–1304.

Baron, Reuben M and David A Kenny (1986). "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." In: *Journal of Personality and Social Psychology* 51.6, p. 1173.

Chipman, Hugh A, Edward I George, and Robert E McCulloch (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Dorie, Vincent et al. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition". In: *Statistical Science* 34.1, pp. 43–68.

Hahn, P Richard, Carlos M Carvalho, et al. (2018). "Regularization and confounding in linear regression for treatment effect estimation". In: *Bayesian Analysis* 13.1, pp. 163–182.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho (2020). "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)". In: *Bayesian Analysis* 15.3, pp. 965–1056.

Hill, Jennifer L (2011). "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Imai, Kosuke, Luke Keele, and Dustin Tingley (2010). "A general approach to causal mediation analysis". In: *Psychological Methods* 15.4, p. 309.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010). "Identification, inference and sensitivity analysis for causal mediation effects". In: *Statistical Science* 25.1, pp. 51–71.

Linero, Antonio R and Qian Zhang (2022). "Mediation analysis using Bayesian tree ensembles." In: *Psychological Methods.*

MacKinnon, David P (2008). *Introduction to statistical mediation analysis.*

MacKinnon, David P and James H Dwyer (1993). "Estimating mediated effects in prevention studies". In: *Evaluation Review* 17.2, pp. 144–158.

Pearl, Judea (2001). "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., pp. 411–420.

Robins, James M and Sander Greenland (1992). "Identifiability and exchangeability for direct and indirect effects". In: *Epidemiology*, pp. 143–155.

Rubin, Donald B (1981). “The Bayesian bootstrap”. In: *The Annals of Statistics*, pp. 130–134.

— (2004). “Direct and indirect causal effects via potential outcomes”. In: *Scandinavian Journal of Statistics* 31.2, pp. 161–170.

VanderWeele, Tyler J (2016). “Mediation analysis: A practitioner’s guide”. In: *Annual Review of Public Health* 37, pp. 17–32.

Wendling, Thierry et al. (2018). “Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases”. In: *Statistics in Medicine* 37.23, pp. 3309–3324.

Woody, Spencer, Carlos M Carvalho, and Jared S Murray (2021). “Model interpretation through lower-dimensional posterior summarization”. In: *Journal of Computational and Graphical Statistics* 30.1, pp. 144–161.

Wright, Sewall (1921). “Correlation and causation”. In: *Journal of Agricultural Research.*