Bayesian Trees and Causality.

– A semiparametric modeling approach using Bayesian additive regression trees with an application to evaluate heterogeneous treatment effects.

Bret Zeldow[1], Vincent Lo Re III[2], Jason Roy[3]

1. Department of Health Care Policy, Harvard Medical School, 2. Department of Medicine, University of Pennsylvania, 3. Department of Biostatistics and Epidemiology, Rutgers School of Public Health

1. Introduction and Causality

a) Overview:

- Challenges in prescription of the correct drug for people with HIV persist, specifically for those with co-morbidities.
- In the US 25% of those with HIV also have Hepatitis-C (HCV) which can lead to liver failure
- Improving HIV-related symptoms has a negative overall effect when it is accompanied by a fatal decline in liver function.
- In previous work, it was discovered that increased cumulative exposure to mtNRTIs (mitochondrial toxic nucleoside reverse transcriptase inhubator) imposes higher risk of decompensation and death [].
- In this article, the results are extended to a potential modifier of the effect FIB-4 (fibrinogen-4).

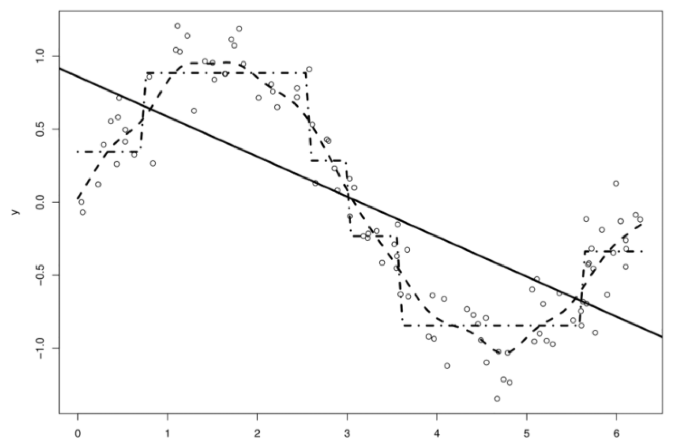

Fig. 1.
Illustration of a BART fit with a univariate predictor space $x \in [0, 2\pi]$ and mean response $y = \sin(x) + \epsilon$. The solid line is the fit using linear regression, the dashed line is the fit of BART, and the dashed–dotted line is the fit of a single tree.

b) The question: Does the effect of mtNRTIs on the risk of death (within two years) change for individuals with varying FIB-4 levels?

c) Model basic idea: A new model is proposed to solve the above question that is highly interpretable, due to its parametric nature, while also maintaining the flexibility of nonlinearity on the remaining confounders. The model is called *semi-BART* and its potential can be seen in Fig 1, where it manages to capture the non-linearity of a sinusoid much better than linear regression and a usual tree [1].

2. Model

The modification of BART proposed in the paper, called semi-BART, partitions the covariates space into $L = L_1 \cup L_2$. One for the covariates that are **directly** influencing the relevant question (L_2), and **all the other covariates** (L_1), that can increase the model accuracy. The treatment and effect-modifiers are modeled in **linear terms** for interpretability, while the second partition with **BART** for flexibility.

Mathematical Formulation

- BART: uses sum-of-trees to predict a binary or real-valued target, given some predictors. $Y = \omega(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, and $\omega()$ is the unknown function, that relates X to Y.
Usually, $\omega(x) = \sum_{j=1}^{m} \omega_j(x; T_j, M_j)$ where each $\omega_j(x)$ is a tree with T_j parameter that represents the tree structure, and M_j the one for nodes.
MCMC is used, in order to estimate the fixed but unknown σ^2 hyperparameter of the ϵ term.

- semi-BART: Usually in research problems, only a few covariates are of scientific interest.
So we write $Y_i = \omega(L_1) + h(L_2, \psi) + \epsilon_i$, where $h()$ is a parametric function of its covariates in ψ, estimated using linear regression and $\omega()$ is of unspecified form estimated using BART.

3. Simulation

A comparison of semi-BART was performed on simulated data for $n = 500$ and $n = 5000$, between semi-BART, BART, GAM (Generalized additive models) and linear / logistic regression. Since the data were generated from a known distribution, the bias, 95% CI, and empirical standard deviation could be measured. Various other comparisons have been conducted in the paper, but the results were similar across all of them.

- Continuous outcome with binary treatment and no effect modification.
Table 1 shows that for small n, there is some bias, but all the algorithms are biased by the same amount and in the same direction.

Table 1

Results from simulation study (scenario 1) with no effect modifiers. Bias: mean absolute bias across 500 datasets. Cov: Confidence/credible interval covarage (percent of simulations where the true value falls within the 95% interval). ESD: Empirical standard deviation defined as the standard deviation of the 500 estimates

Method	Parameter	Bias	Cov.	ESD
			$n = 500$	
Semi-BART	ψ_1	−0.02	0.96	0.153
GAM	ψ_1	−0.02	0.94	0.371
BART	ψ_1	−0.02	0.94	0.153
Regression	ψ_1	−0.02	0.95	0.390
			$n = 5000$	
Semi-BART	ψ_1	0.00	0.95	0.036
GAM	ψ_1	0.00	0.94	0.111
BART	ψ_1	0.00	0.92	0.037
Regression	ψ_1	0.01	0.94	0.119

- Misspecified linear term.
Table 4 clearly shows that all results are quite similar, but semi-BART has slightly lower ESD.

Table 4

Results from simulation study (scenario 3) for binary outcomes. Bias: mean absolute bias across 500 datasets. Cov: Confidence/credible interval covarage (percent of simulations where the true value falls within the 95% interval). ESD: Empirical standard deviation defined as the standard deviation of the 500 estimates. The true parameter values are $\psi_1 = 0.3$, $\psi_2 = -0.1$, and $\psi_3 = 0.1$

Method	Parameter	Bias	Cov.	ESD
			$n = 500$	
Semi-BART	ψ_1	0.03	0.92	0.144
	ψ_2	0.00	0.94	0.140
	ψ_3	0.00	0.93	0.106
Regression	ψ_1	−0.01	0.93	0.131
	ψ_2	0.01	0.94	0.127
	ψ_3	−0.01	0.94	0.101
			$n = 5000$	
Semi-BART	ψ_1	0.00	0.94	0.039
	ψ_2	0.00	0.95	0.039
	ψ_3	0.00	0.94	0.029
Regression	ψ_1	−0.03	0.84	0.038
	ψ_2	0.01	0.93	0.036
	ψ_3	−0.01	0.93	0.029

4. Medical data application

Data are gathered from Veterans Aging Cohort Study (VACS) 2002-2009. The sample consists of patients with HIV/Hepatitis C coinfections. The data contain variables, such as demographics, time of initiation of treatment, HIV characteristics, other laboratory measures etc. The **outcome** of this analysis is a binary variable, that indicated survival of the patient within the two-year period.

- The analysis consisted of $m = 50$ trees with 20,000 iterations.
- L is partitioned into L_2, variables of mtNRTI and the FIB-4 index and L_1, all the other variables.

The three models are

1. Without continuous effect modifier.
2. With continuous effect modifier.
3. With binary effect modifier.

The effect modifier is the FIB-4 index and, the result is:
When FIB-4 is present, the effect of mtNRTI is magnified.

Table 7

Comparison of point estimates 95% confidence/credible intervals from our data analysis using semi-BART and probit regression. The outcome is a binary indicator of death. ψ_1 is the parameter for the treatment (mtNRTI use) effect and ψ_2 is the parameter for the interaction between mtNRTI use and FIB-4 (binary or continuous)

Analysis	Parameter	Semi-BART	Probit regression
No effect modifier	$\hat{\psi}_1$	0.15 (−0.02,0.33)	0.18 (0.01, 0.35)
Continuous effect modifier	$\hat{\psi}_1$	0.18 (0.00,0.36)	0.20 (0.03, 0.37)
	$\hat{\psi}_2$	0.07 (0.02,0.12)	0.06 (0.01, 0.11)
Binary effect modifier	$\hat{\psi}_1$	0.07 (−0.12,0.26)	0.10 (−0.08, 0.29)
	$\hat{\psi}_2$	0.38 (0.07, 0.69)	0.34 (0.04, 0.64)

5. References

[1] Bret Zeldow, Vincent Lo Re III, and Jason Roy. 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, 2019.