Tree Based Predictive Models for Noisy Input Data

Kevin McCoy[1,2] Zachary Wooten[3] Christine Peterson[2]

[1]Department of Statistics, Rice University [2]Department of Biostatistics, UT MD Anderson Cancer Center [3]Department of Biostatistics, St. Jude Children's Research Hospital

Motivation

- Measurement error occurs when the measured value of a quantity of interest is different than the true, unknown value.
- A measurement error model directly accounts for this error in the independent variable, and not just the response variable as in a traditional statistical model.
- Measurement errors are very common across the medical field when exact values cannot be calculated.
- While existing works incorporating measurement error models into regression settings are common, less work has been done incorporating measurement error into more robust and flexible models.
- Bayesian Additive Regression Trees (BART) is an ensemble decision tree model, similar to the random forest or boosted decision trees model, that uses multiple learners to each explain a small portion of the variance of the output data to achieve good predictive performance (Chipman et al., 2010).
- In this work, we develop an extension to the BART model that directly incorporates measurement error in the independent variable(s), which we call meBART.

Background

The standard BART model assumes that data $\{(x_i, y_i) : i = 1, \ldots, n\}$ are IID and generated according to some unknown function f such that:

$$y_i = f(x_i) + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

The goal of BART is to estimate the conditional mean $\mathbb{E}[y_i|x_i]$ by a sum of m decision trees, $f(x_i) = \sum_{h=1}^{m} g(x_i; T_h, M_h)$. An example decision tree is shown below in **Figure 1**.

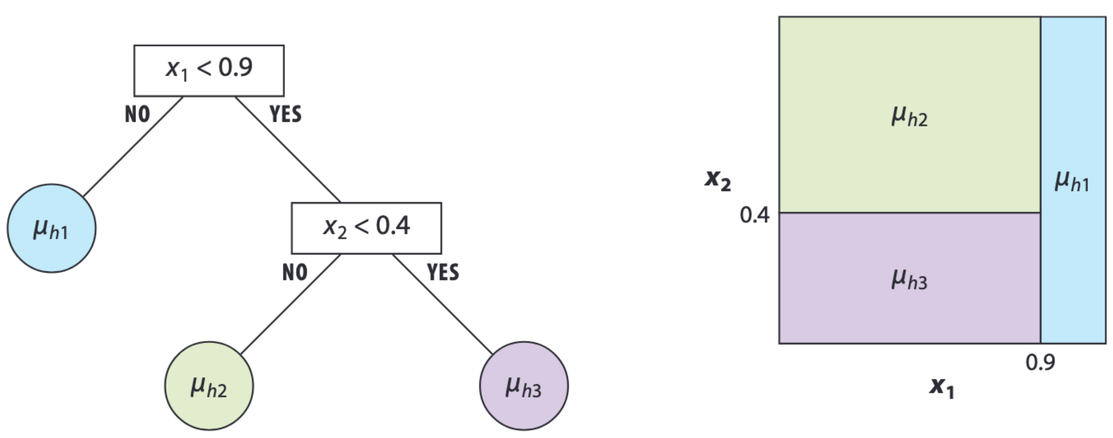

Figure 1. An example of a binary decision tree and its corresponding partition of the data space. Taken from (Hill et al., 2020).

T_h represents the set of decision rules that govern the h'th tree and M_h is the set of leaf node values at the bottom of the h'th tree. This yields the likelihood:

$$p(y_i|T_h, M_h, \sigma) = \mathcal{N}\left(\sum_{h=1}^{m} g(x_i; T_h, M_h), \sigma^2\right) \tag{2}$$

We also assume that the model variance is independent of the trees, that trees are independent from one another, and that the leaf nodes within a tree are independent. This yields the following prior distribution:

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma) = \left[\prod_{h=1}^{m} p(M_h|T_h)p(T_h)\right] p(\sigma) \tag{3}$$

$$p(M_h|T_h) = \prod_{t=1}^{b_h} p(\mu_{ht}|T_h) \tag{4}$$

The full posterior can then be estimated using a Metropolis-Hastings-within-Gibbs sampler, where each tree is updated according to a Bayesian backfitting algorithm.

Methods

We extend the vanilla BART model by assuming that the independent variable x_i is measured with error such that:

$$x_{i,measured} = x_{i,true} + e_i \qquad e_i \sim \mathcal{N}(0, \sigma_e^2) \tag{5}$$

The proposed model is thus a Bayesian hierarchical model, where the response variable y_i depends on the unobserved, true latent value $x_{i,true}$. This value is an additional parameter in the model, and can be inferred alongside the normal tree model parameters. To do so, we place a normal prior on the measurement error, and assume that the variance of the measurement error is known. Finally, we estimate $x_{i,true}$ via an additional Metropolis-Hastings step within the standard BART MCMC. By directly estimating $x_{i,true}$, we hope to both learn this underlying value and achieve better predictive performance on unseen test data.

Results

Synthetic data were generated according to $X \sim Unif(0, 1)$ and the true underlying step function:

$$f(x) = \mathbb{1}_{[0.5,\infty)}(x) \tag{6}$$

This step function can be thought of as a decision tree with one layer. IID $\mathcal{N}(0, 0.1^2)$ measurement error was added to both X and y variables. Shown below in **Figure 2**, meBART achieves a smoother mean function with 95% credible intervals that fully capture the true underlying function.

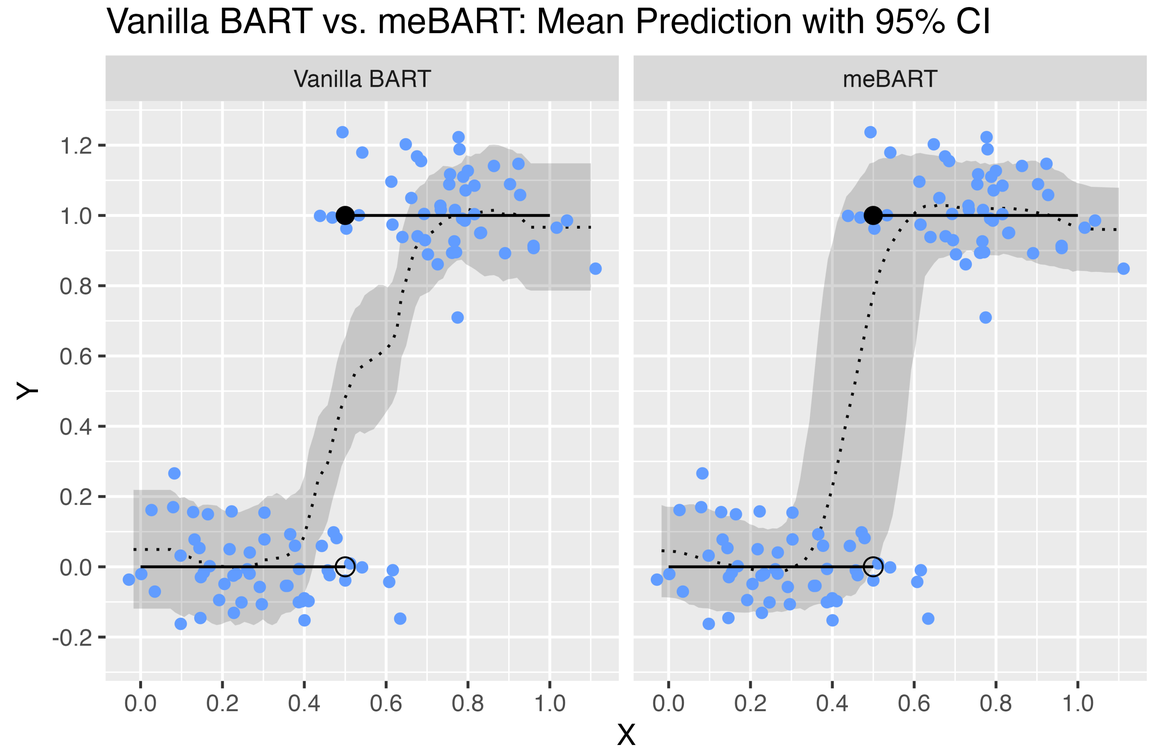

Figure 2. The posterior mean prediction values of vanilla BART and meBART (plotted as black dotted lines) along with their respective 95% credible intervals (the gray shaded regions).

In traditional BART, the mixing of σ is taken as a proxy for the quality of mixing for the other parameters and overall quality of the fitted model. Shown below in **Figure 3**, meBART achieves a posterior distribution much closer to the true underlying value than that of vanilla BART.

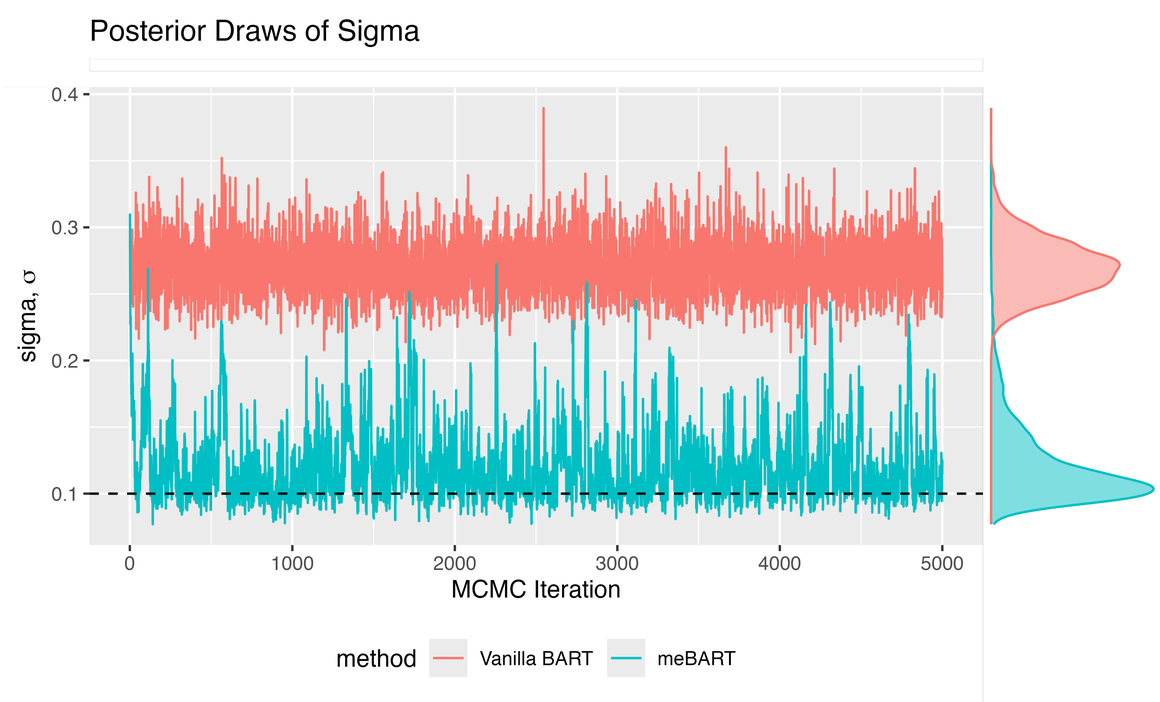

Figure 3. Posterior trace plots of σ. The true underlying value is shown as a black dashed line.

Discussion

We show that in the presence of measurement error, our model allows for much better recovery of the true underlying function and more accurate estimation of model parameters. Going forward, we are actively investigating multivariate and more complex functions with measurement error. We also are exploring various real life datasets that would benefit from having measurement error modeled.

References

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), Mar. 2010. ISSN 1932-6157. doi:10.1214/09-AOAS285. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-1/BART-Bayesian-additive-regression-trees/10.1214/09-AOAS285.full.

J. Hill, A. Linero, and J. Murray. Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, Mar. 2020. ISSN 2326-8298, 2326-831X. doi:10.1146/annurev-statistics-031219-041110. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-031219-041110.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842494, as well as The Ken Kennedy Institute Computational Science and Engineering Recruiting Fellowship, funded by the Energy HPC Conference.