The Dissertation Committee for Pedro Henrique Filipini dos Santos
certifies that this is the approved version of the following dissertation:

Tree-based models with basis functions

Committee:

Jared S. Murray, Supervisor

Carlos M. Carvalho, Co-supervisor

Antonio R Linero

Magdalena Bennett

Tree-based models with basis functions

by

Pedro Henrique Filipini dos Santos

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin
August 2025

Acknowledgments

I thank my advisor, Jared Murray, for his guidance, patience, and for making sure I would be on track when I needed it most. I am also thankful to Carlos Carvalho, who has been an important source of support, collaboration and research insight.

I am deeply grateful to Hedibert Lopes, my M.S. advisor, who believed in me and encouraged me to pursue a Ph.D., even when I doubted myself. I also owe a great deal to Mauri Oliveira, my undergraduate advisor, who first introduced me to the world of Statistics and inspired my shift in academic direction.

To my friends, thank you for reminding me that there is life beyond research and for standing by me throughout this journey. To my family, especially my parents Selma and Luiz, thank you for the unwavering love and support.

And to my partner Jasmine, for keeping me grounded, focused, and hopeful through every high and low. Your presence and love meant everything to me.

Abstract

Tree-based models with basis functions

Pedro Henrique Filipini dos Santos, PhD
The University of Texas at Austin, 2025

SUPERVISORS: Jared S. Murray, Carlos M. Carvalho

The Bayesian Additive Regression Trees prior is a powerful prediction tool, with a wide range of extensions applied to a variety of problems. Some of these extensions, however, lack computational efficiency due to increased complexity. We present three novel approaches: a scalable BART with targeted smoothness, which uses a reduced-rank Gaussian Process approximation to improve scalability; a locally adaptive linear BART, providing smoother predictions for BART models while maintaining most of the computational efficiency from BART; and a scalable Bayesian causal forest for continuous treatments, which utilizes a reduced-rank Gaussian Process approximation to extend the Bayesian causal forest model. Simulations were conducted to evaluate computational efficiency and predictive performance. The methods were also applied to real-world datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 The advent of tree-based models

A classic and general problem in statistics is to establish the relationships between a response variable, let us say Y, and a set of covariates, which is this case are represented by $\boldsymbol{X}$. There are multiple ways that try to develop solutions to this problem, and most of them are of the form

$$Y = f(\boldsymbol{X}) + \epsilon,$$

where ϵ is the error term that cannot be explained by the set of covariates $\boldsymbol{X}$ and can follow some probability distribution. The problem of establishing the relationship between Y and $\boldsymbol{X}$ becomes the problem of estimating the function $f(.)$. The set of algorithms and tools that are used to estimate the function $f(.)$ is called statistical learning.

Among the statistical learning techniques, the group of models that use binary trees to create partitions of the data is the topic of interest of this dissertation, with models that belong to this group now referred to as tree-based models.

Binary trees are basically a way to represent the decision process of the creation of partitions in a dataset. Let us say, for example, that a person works for a company that estimates the average amount that a costumer with certain characteristics would pay for medical insurance, and this person knows that the insurance companies of their specific city charge a higher price for people that are over 65 years old. In this case, one way that this person might want to separate the people for their insurance price estimation is by creating two groups, one with people that are at least 65 years old, and one for younger people. This binary decision process is basically a binary tree, with the difference that binary trees usually rely on algorithms to determine the cutpoint which will determine to which partition each data point belongs.

As noted by Breiman et al. (1984), the advent of computers allowed the development of the first tree-based models, which were motivated by the need of social scientists to deal with real world problems. The use of these models in regression problems date back to the 1960s with the Automatic Interaction Detection program developed by the University of Michigan, and since then, more sophisticated approaches have been developed to deal with the drawbacks of each model.

The classification and regression trees (CART) model developed by Breiman et al. (1984) in the 1970s helped to build the foundation of the tree-based models in its early days, as well as the development of some of the theoretical background of these models.

Nevertheless, one of the drawbacks of the CART is that due to the way that a binary tree is created in the original algorithm, each tree is deterministic based on the data that the model is trained, and small changes in the dataset could lead to drastically different binary trees. In other words, the model would perform well within the training data, but would have generalization problems and perform poorly when used to predict out-of-sample data, meaning that the model has low bias and high variance. To overcome these issues, Breiman (1996) introduced the bagging predictors, on which an aggregated predictor is created by averaging the responses of a large number of trees, where each one of them had been trained in a bootstrap sample of the original dataset. This idea would later be expanded to create Random Forests (Breiman, 2001), where just a subset of the covariates of each bootstrap sample would be available in each tree, helping to solidify the role of tree ensembles in problems that required satisfactory predictive performance, which, naturally, occur with the trade-off of losing the interpretability of a single binary tree.

1.2 Bayesian tree-based models

Alternatively, on the Bayesian side, Chipman et al. (1998) introduced the Bayesian CART, which now introduced prior probabilities to the depth of the tree, the selected variable to create a rule, cutpoints, leaf node parameters and so on. While the Bayesian CART solves the issue of creating the same tree for a specific dataset due to the nature of posterior sampling in Bayesian models, the model still had issues such as losing some the interpretability of the CART, since the posterior sample of trees could differ reasonably and usually a point estimate such as the posterior mean of a prediction is desirable, but lacks the straightforward clarity of working with a single tree.

Further extending the Bayesian CART, the Bayesian Additive Regression Trees (BART) model was developed by Chipman et al. (2010), basically creating a sum-of-trees model, where each tree is trained on the residuals of the previous trees, naturally allowing the inclusion of additive effects and interactions in the model. The trees are also regularized in a sense that small trees are encouraged by the prior distribution, which means that, in general, each tree is trying to capture only a small contribution of the final prediction based on only a few of the possible interactions allowed in the model. This framework allowed BART to be an powerful tool in prediction-based problems of various natures.

BART models have been applied to a multitude of settings, such as causal inference (e.g. Hill (2011), Hahn et al. (2020), Starling et al. (2020), Yeager et al. (2022)), and survival analysis (e.g. Sparapani et al. (2016), Sparapani et al. (2018), Linero et al. (2022)). More sophisticated problems usually required more sophisticated priors, which leads to the increase of computational cost and lower time efficiency.

In the same way that the bias-variance trade-off motivated the development of ensembles among the tree-based models, limitations in the way that these models are designed motivated the creation of different frameworks that allowed these models to be more successfully applied to a wider range of problems.

Our approaches are focused on proposing scalable alternatives to some of the time efficiency problems found on Starling et al. (2019) and Linero (2018), while also extending these novel approaches for the causal inference setting.

1.3 Outline

This dissertation consists of 5 chapters, with Chapter 1 being an introduction to tree-based models and the motivation behind the development of different techniques for the scalability of tree-based models for distinct kinds of problems.

Chapter 2 focuses on the development of a scalable BART with targeted smoothness, expanding on the ideas of Starling et al. (2019) by using a reduced-rank Gaussian Process approximation (Solin and Särkkä, 2020).

Chapter 3 is based on the development of an alternative to BART focused on smoother function predictions, while being more time efficient than current BART extensions (e.g. Linero and Yang (2018)).

Chapter 4 revisits the reduced-rank Gaussian Process approximation from Chapter 2 and expands the Bayesian causal forest model (a BART extension focused on causal inference with binary treatment) to the general case of continuous treatments.

Chapter 5 provides a discussion on the main results reported in the previous Chapters.

Code availability

The source code used in this dissertation is available at:

`https://github.com/pedrofilipini/UT_dissertation`

Chapter 2: Scalable Bayesian additive regression trees with targeted smoothness

Abstract

Among the methods used to estimate nonlinear functions, Bayesian Additive Regression Trees (BART) priors have shown promising results, especially for prediction, being used in a wide range of different problem settings, including causal inference and variable selection. BART priors represent functions as the sum of many small trees, each parameterizing a step function. However, smoothness is often desirable for parsimony, accurate interpolation, and extrapolation, while keeping the efficiency. We expanded the BART model by replacing locally constant predictions with a set of m basis coefficients that are used in an approximation of the covariance kernel of a Gaussian Process, leading to a scalable alternative of a targeted smooth curve in each leaf. A special case of this more general class of BART models with vectors of basis coefficients attached to the leaves was also developed, where on each leaf we have locally linear predictions, regressing on the variable attached to the parent node.

2.1 Introduction

Having to deal with stressful situations is a burden most people have to live with in the daily basis. Ranging from dealing with immediate danger to meeting a deadline at work, these events arise often in everyday life and can cause some level of discomfort and anxiety, meaning that the ability to cope properly while facing these challenges is desirable. One possible way to cope with stress is by creating awareness of what is happening to the human body during these events and learning to embrace its reaction to the current circumstances, meaning that signals such as sweaty hands and accelerated heart rate can be recognized as ways that the body prepared itself to act in the face of what it considers danger, functioning, in some way, as a defense mechanism.

The way our bodies react to those events is directly related to time, since the stressful situation happens during some period of time, during which the body reacts, and after the event is over the body is expected to return to its baseline behavior, meaning that if some variable that can be considered an indicator of stress is being measured, its behavior is expected to vary over time, especially throughout an stressful event. It can be said that the behavior of this variable is a function $f(t, \boldsymbol{x})$, where t indicates time and $\boldsymbol{x}$ represents the measured variables specific to the subject that is being evaluated.

One way to model this kind of function is by using targeted smoothing Bayesian additive regression trees (tsBART) introduced by Starling et al. (2019), since its main idea is to select a variable, which in this case is time, that the response variable should vary across smoothly. The tsBART is a direct expansion of Bayesian additive regression trees (BART) by Chipman et al. (2010), which a tool that can adapt reliably to a wide variety of problems such as survival analysis (Sparapani et al., 2016), variable selection (Linero, 2018), heteroscedasticity (M. T. Pratola and McCulloch, 2020), and causality (Hahn et al., 2020).

In the specific setting of this paper, the data that is being analyzed is a randomized controlled trial (RCT) where participants went through stressful events in a controlled environment while having some of their physiological variables recorded over time. Our response variable is a variable called total peripheral resistance, which can be seem as an indicator of stress. Two groups were studied, namely treatment and control, where the treatment group was exposed to a 30-minute online training teaching them how stress works and how stressful situations can help a person to improve its abilities, introducing them to the so called synergistic mindsets, which can be seem here a cope mechanism.

Using a modified version of Hahn et al. (2020) Bayesian causal forest (bcf), along with tsBART priors (Starling et al., 2019) it is possible to model the relationship of the response variable smoothly over time at the same time that the Conditional Average Treatment Effect (CATE) can be accounted due to the fact that the bcf prior can naturally capture treatment effects.

One problem of this approach is that using these kind of priors can be computationally expensive, not scaling with the number of unique values for the variable selected for the targeted smoothness. Our contributions to the paper are the introduction of a scalable tsBART using one method of approximation of the covariance function by basis functions, along with the development of a method to properly control the relative error of the approximation.

The article is structured as it follows: Section 2.2 introduces the problem and the motivation to the development of scalable tsBART priors; Section 2.3 defines the notation and reviews BART and some of its extensions; Section 2.4 introduces the novel scalable tsBART alongside a method to properly perform the required basis functions approximations; Section 2.5 reviews similar works on basis functions approximations; on Section 2.6 simulations are performed to quantify the quality of the approximation in comparison to Gaussian processes with an exact covariance function; the analysis of the data is performed on Section 2.7, along with subgroup selection methods; finally, Section 2.8 gives a brief review of the findings of this paper, as well as possible extensions to the methods presented.

2.2 Context of the problem

Dealing constantly with stressful situations may cause some damages to the mental health of individuals , specially adolescents, since those negative interactions may cause a negative feedback loop, such that the person is discouraged to deal with hard situations, which may lead to the adolescent losing many opportunities of growth and the development of required skills in both social and professional life (Yeager et al., 2022).

This study aimed to test the effectiveness of a 30-minute online training on which the person was presented to the so called synergistic mindsets. The idea of synergistic mindsets is to present two types of mindset that can be seen as ways to cope with stress, but that at the same time have a good synergy and can be presented together. These mindsets are the Stress-can-be-enhancing mindset, which basically states that your body is ready to deal with any dangers by increasing the flux of blood to essential organs to be able to deal with the problem that has been presented, and the Growth mindset, which states that challenging situations present a way to the person to hone their skills and grow as an individual due to the learning that will occur as a result of that situation.

One way that the study found to keep track of the stress levels of a participant was to focus on the variable TPR over time, and since the changes in TPR are expected to be smooth over time, as well as the treatment effect, a possible way to express the data is

$$y_{i,t} = \alpha_i + \varphi(\boldsymbol{x}_i, t) + \psi(\boldsymbol{w}_i, t)z_i + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma^2), \tag{2.1}$$

where φ and ψ are independent tsBART priors for the prognostic effect and the treatment effect, respectively, α_i is a random effect variable, and z_i is a binary treatment variable. Details about the data, as well as further information about the experiment can be found on Section 2.7.

2.3 Background and notation

Let $\boldsymbol{x}_i \in \mathcal{X}$, $\boldsymbol{w}_i \in \mathcal{X}$ be vectors of covariates associated with the individual $i \in \{1, ..., N\}$, and let $t \in \mathcal{T}$ be the variable associated with the variable selected for the targeted smoothness. As stated on Section 2.2, the problem that must be solved is of the form

$$y_{i,t} = \alpha_i + \varphi(\boldsymbol{x}_i, t) + \psi(\boldsymbol{w}_i, t)z_i + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma^2), \tag{2.2}$$

where φ (ψ) is a nonlinear function that can capture the heterogeneity of the prognostic (treatment) effect among the $\boldsymbol{x}_i$ $(\boldsymbol{w}_i)$ when $z_i = 0$ $(z_i = 1)$, while also being smooth over the variable t since we believe that y_i changes are smooth over time. Ideally the model should also be scalable to deal with large $|\mathcal{T}|$. We introduce the scalable targeted smoothing Bayesian causal forest (scalable tsbcf) as a possible way of solving this problem. Before getting into the specifics of the novel technique, the original BART (Chipman et al., 2010) is reviewed, followed by the tsBART (Starling et al., 2019), presenting a way to include targeted smoothness into BART. For scalability, a representation of the covariance kernel of Gaussian Processes by using basis functions is presented (Solin and Särkkä, 2020), as well as the approach for selecting the parameters for the approximation. Finally, the scalable tsbcf model is introduced.

2.3.1 Bayesian Additive Regression Trees

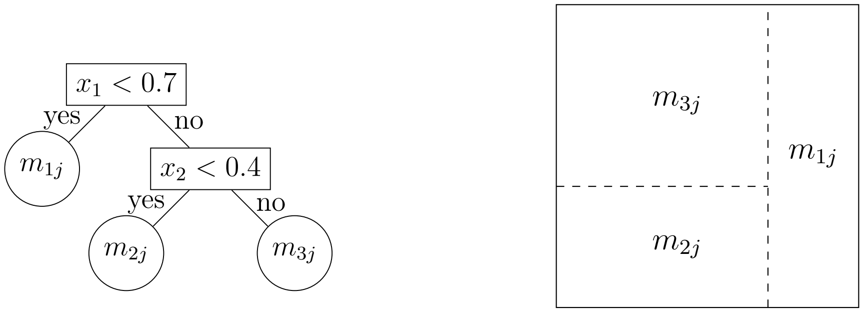
Bayesian Additive Regression Trees (BART), introduced by Chipman et al. (2010), is a nonparametric prior over a regression function $f(\boldsymbol{x}_i)$ constructed as an adaptive basis expansion parameterized by a collection of binary trees $\{T_j : 1 \leq j \leq J\}$ and associated leaf node parameters $\{\Pi_j = (m_{1j}, \ldots, m_{B_j j})' : 1 \leq j \leq J\}$ over the B_j partitions of T_j $\{B_j = |\mathcal{B}_j| : b_{1j}, \ldots, b_{B_j j} \in \mathcal{B}_j, b_{b'j} \cap b_{b''j} = \emptyset, \forall b' \neq b''\}$; see Figure 2.1 for an example. Each tree and its associated parameter vector define a basis function g:

$$g(\boldsymbol{x}_i, T_j, \Pi_j) = \sum_{k=1}^{B_j} \mathbb{I}\left(\boldsymbol{x}_i \in [b_{kj}]\right) m_{b_{kj}j}, \tag{2.3}$$

which returns the parameter associated to the end node to which $\boldsymbol{x}_i$ belongs. A BART prior is constructed by summing many such basis functions:

$$f(\boldsymbol{x}_i) = \sum_{j=1}^{J} g(\boldsymbol{x}_i, T_j, \Pi_j), \tag{2.4}$$

where J is the total number of trees. Regularizing priors put high probability on shallow trees and shrink the end node parameters toward zero; see Appendix **??** for details.



(a) Example of a BART tree.

(b) Partition space of the BART tree.

Figure 2.1: A single BART tree and its associated partition space.

BART priors and their variants (Linero and Yang (2018); Murray (2019); Linero (2018); M. T. Pratola and McCulloch (2020)) have been used in a range of regression models; some examples appear in Hill et al. (2020). For the specific case of causal inference, the first approach by Hill (2011) was to include a binary treatment assignment Z among the covariates to estimate a response surface and conditional average treatment effects, while Hahn et al. (2020) instead utilize distinct BART priors over a nuisance parameter μ and treatment effect function τ to create the Bayesian causal forest (bcf) model:

$$y_i = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{w}_i)z_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2.5}$$

where $\boldsymbol{x}, \boldsymbol{w} \subseteq \mathcal{X}$ are subsets of the covariates posted to be potential effect moderators , μ and τ are independent BART priors, and $z_i \in \{0, 1\}$[1] is a binary treatment variable. Woody et al. (2020) allow z_i to be continuous and introduce appropriate modifications to the prior; see also Deshpande et al. (2024) who develop varying-coefficient models of a similar form.

[1] (Hahn et al., 2020) also introduce a data-adaptive encoding of the binary treatment variable Z; we present the simpler model here for exposition.

2.3.2 BART with Targeted Smoothness (tsBART)

Starling et al. (2019) introduces an extension to BART that imposes smoothness in one dimension for a variable $t \in \mathcal{T}$ excluded from the covariates $\boldsymbol{x}$ used to define splitting rules; see Figure 2.2 for an example. This tree can be represented by using basis functions of the form

$$g(t, \boldsymbol{x}_i, T_j, \Pi_j) = \sum_{k=1}^{B_j} \mathbb{I}\left(\boldsymbol{x}_i \in [b_{kj}]\right) \mu_{b_{kj}j}(t), \tag{2.6}$$

replacing the scalar end-node parameters of Equation (2.3) with functions varying over the t variable $\{\Pi_j = (\mu_{1j}(t), \ldots, \mu_{B_jj}(t))' : 1 \leq j \leq J\}$ which are assigned a univariate Gaussian process (GP) prior:

$$\mu_{b_{kj}j}(\cdot) \sim GP\left(0, \frac{1}{J}C_\theta\right),$$

where J is the total number of trees and C_θ is the covariance kernel of the Gaussian Process with parameters θ.

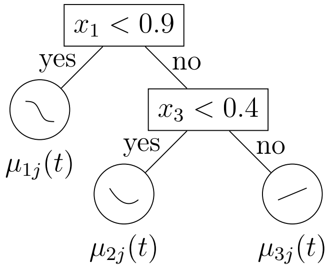


Figure 2.2: An example of a binary tree that is used in tsBART.

Starling et al. (2020) also introduces a version of bcf with targeted smooth trees by applying tsBART priors for both the nuisance parameter and treatment effect function, allowing bcf to estimate smooth functions over a selected variable in the causal inference setting.

2.3.3 Hilbert space Gaussian process approximations

In the studies of Starling et al. (2019) and Starling et al. (2020) the variable $t \in \mathcal{T}$ only takes a few distinct values; as such it was still computationally efficient to conduct full GP inference when updating the leaf parameters. However, when $\{t_i : 1 \leq i \leq n\}$ takes many distinct values – say $\{T : T = |\mathcal{T}|\}$ – MCMC inference involves $O(T^3)$ matrix operations and MCMC algorithm of Starling et al. (2019) rapidly becomes infeasible.

Approximations to accelerate GP inference are well-studied, see e.g. Williams and Seeger (2000), Smola and Bartlett (2000), Quiñonero-Candela and Rasmussen (2005), Lázaro-Gredilla et al. (2010), and Wilson and Nickisch (2015). In our case the GP is over univariate functions, which means that many tractable options are available. Here we utilize a finite-dimensional approximation introduced by Solin and Särkkä (2020). This approximation scheme requires that $\mathcal{T}$ be a compact domain; since we consider only shift-invariant kernels, without loss of generality we assume $\mathcal{T} = [-1, 1]$.

One of the most used kernels for the covariance function of a GP is the square exponential:

$$C_\theta(t, t') = \exp\left(\frac{(t - t')^2}{2\theta}\right), \tag{2.7}$$

and the approximation of Solin and Särkkä (2020) for this kernel takes the form

$$C_\theta(t, t') \approx C^m_{\theta c}(t, t') = \sum_{q=1}^{m} \delta_{\theta c}\,(q)\,\phi_{qc}(t)\phi_{qc}(t'), \tag{2.8}$$

where

$$\phi_{qc}(t) = \sin\left(\frac{\pi q(t + c)}{2c}\right), \qquad \delta_{\theta c}\,(q) = \frac{\theta}{\sqrt{c}} \exp\left(-\frac{\theta^2 \left(\frac{\pi q}{2c}\right)^2}{2}\right), \tag{2.9}$$

and the scalar $c > 1$ defines a window $[-c, c]$ used to control the approximation error; see Appendix A.2 for further detail. We obtain a low-dimensional *fixed* basis, and only the *scale* of the basis functions depends on the parameters of our kernel, which are appealing properties from a computational perspective. The approximation error decays rapidly and is relatively easy to control; see Section 2.4.2. Given Equation (2.8) we can construct approximate draws from $\mu(t) \sim GP(0, C_\theta)$ via the weight-space representation

$$\mu(\cdot) := \sum_{q=1}^{m} \beta_q \sqrt{\delta_{\theta c}\,(q)}\phi_{qc}(\cdot), \quad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{2.10}$$

turning a Gaussian process regression model into a finite-dimensional regression on the basis expansion approximation.

2.4 Methods

In this section we introduce our new methods. We begin by introducing in Section 2.4.1 a generic BART prior with regression within the leaves, which generalizes the original BART prior from Chipman et al. (2010) by making it a special case ($m = 1; \mathcal{T} = 1$), and yields a scalable model for targeted smoothing. Then in Section 2.4.2 we describe how to choose m and c for a given length-scale θ in the approximation described on Equation (2.8) controlling the relative error while maintaining computational efficiency. Finally in Section 2.4.3 we use these insights to introduce an MCMC sampler such that a fully Bayesian inference is performed over the length-scale parameter. Sections 2.4.2 and 2.4.3 will be of independent interest to other models using the Hilbert space approximation for a univariate GP (e.g. Solin and Särkkä (2020), Riutort-Mayol et al. (2020)).

2.4.1 Basis BART: Scaling targeted smoothing

To make tsBART scalable, the procedure introduced on Section 2.3.3 is used. The structure remains almost the same as the presented on Section 2.3.2, except for the leaf nodes of the trees that now are composed of vector $\{\boldsymbol{\beta}_{b_k j} = (\beta_{1 b_k j}, \ldots, \beta_{m b_k j})' : 1 \leq j \leq J\}$; see Figure 2.3 for an example.

In this new framework, the response of a leaf node is given by

$$\mu(t) \approx \sum_{q=1}^{m} \beta_q \omega_q(t), \tag{2.11}$$

and it is defined to have this form because by setting the prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we have

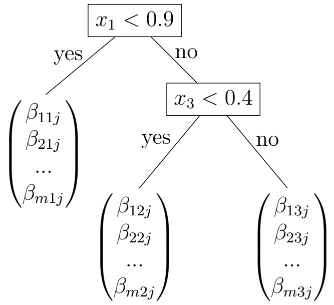

$x_1 < 0.9$
yes
no
$x_3 < 0.4$
yes
no
β_{11j} β_{21j} ... β_{m1j}
β_{12j} β_{22j} ... β_{m2j}
β_{13j} β_{23j} ... β_{m3j}

Figure 2.3: An example of a binary tree that is used in basis BART with scaling targeted smoothing.

that

$$\mathbb{C}ov(\mu(t), \mu(t')) = \sum_{q=1}^{m} \omega_q(t)\omega_q(t'), \tag{2.12}$$

such that

$$C_\theta(t, t') \approx \sum_{q=1}^{m} \omega_q(t)\omega_q(t'), \tag{2.13}$$

which is equivalent to the basis functions presented on Equation 2.8 in Section 2.3.3.

These changes can be easily implemented by using the current tsBART framework, while creating a scalable alternative due to the truncation of the number of basis functions, which means that the number of matrix operations for the MCMC inference is now $\mathcal{O}\left(|\mathcal{T}|m^2\right)$.

Since the scaling targeted smoothing prior is simply an approximation of the tsBART prior, it is possible to readily substitute the tsBART prior in tsbcf (Starling et al., 2020) for the causal inference setting, as well as in the continuous treatment framework (Woody et al., 2020), and even for linear varying coefficient models (Deshpande et al., 2024).

2.4.2 Choosing the dimension m and window c for a fixed length-scale

The quality of the approximation in Equation (2.8) depends on the number of basis functions used (m) and the size of the expanded window $[-c, c]$ in ways that subtly depends on the length-scale θ under consideration. Some discussion of choosing m and c appears in Solin and Särkkä (2020); see also Riutort-Mayol et al. (2020).

As proposed by Riutort-Mayol et al. (2020), one way to measure the quality of the approximation is by using the relative error $e_\theta(m, c)$, that can defined as

$$e_\theta(m, c) = \frac{\int |C_\theta(0, \tau) - C^m_{\theta c}(0, \tau)| d\tau}{\int C_\theta(0, \tau) d\tau}, \tag{2.14}$$

where $C_\theta(t, t') = C_\theta(0, \tau)$, $\tau = t - t'$.

We begin with an example, taking $\theta = 0.1$. Figure 2.4 plots the relative error $e_{0.1}(m, c)$ as a function of c for a range of basis dimensions m. For any fixed c the relative error decreases as the number of basis functions m increases; while the minimum relative error depends on m, the value of c that attains the minimum is nearly the same. There is a relatively compact interval for c over which the relative error is close to the minimum for any given m.

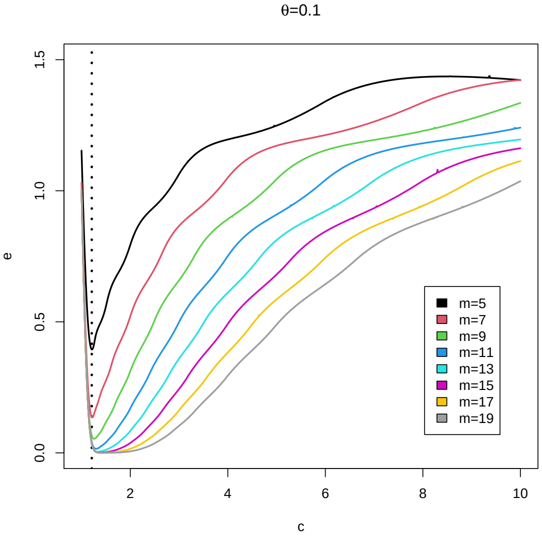


Figure 2.4: The effect of c over the error for different values of m. Length-scale fixed at 0.1. The relative error e is represented in the y-axis. The dashed black line represents the value c that minimizes $e_{0.1}$ for $m = 9$.

Now suppose $\theta = 0.25$ instead, so that typical draws from the prior are smoother functions. Figure 2.5 shows the relationship between c and $e_{0.25}$ for different values of m, while also comparing the optimum value of c when $m = 9$ for in the previous example when $\theta = 0.1$. Since θ is relatively large for the domain $[-1, 1]$ we see that the minimum possible relative error is nearly the same for each m (and close to zero), and is approximately obtained by a wide range of c values. The optimal choice of c for $\theta = 0.25$ and $m = 9$ in this case is given by the dashed red line; note that the optimal c when $\theta = 0.25$ and $m = 9$ is greater than the optimal c when $\theta = 0.1$ and $m = 9$ (the dashed black line).

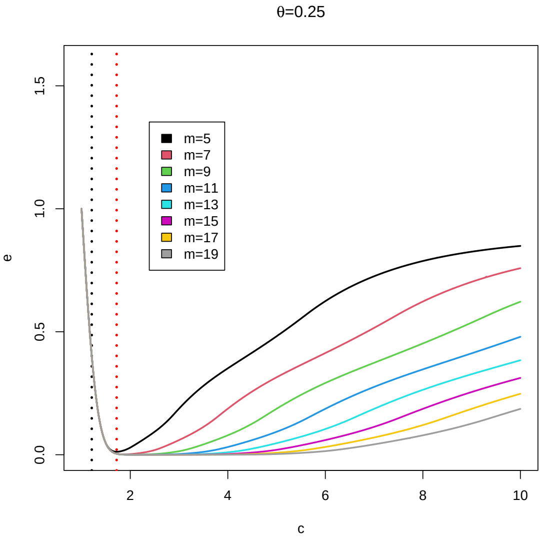


Figure 2.5: The effect of c over the error for different values of m. Length-scale fixed at 0.25. The dashed black line represents the value c that minimizes $e_{0.1}$ for $m = 9$.The dashed red line represents the value c that minimizes $e_{0.25}$ for $m = 9$.

These examples highlight two key considerations in constructing the approximation: The optimal choice of c is highly dependent on the length-scale, as is the optimal value for m when the length-scale is small relative to the domain. For a fixed length-scale we therefore propose the following procedure: Fix a maximum tolerable relative error $\epsilon_{\max}$, and then find the smallest value for m that attains $\epsilon_{\max}$ for some value of c. With m fixed, it is possible to set c to the value that produces the smallest relative error. For example, Figure 2.6 shows that with $\theta = 0.1$, the optimal value

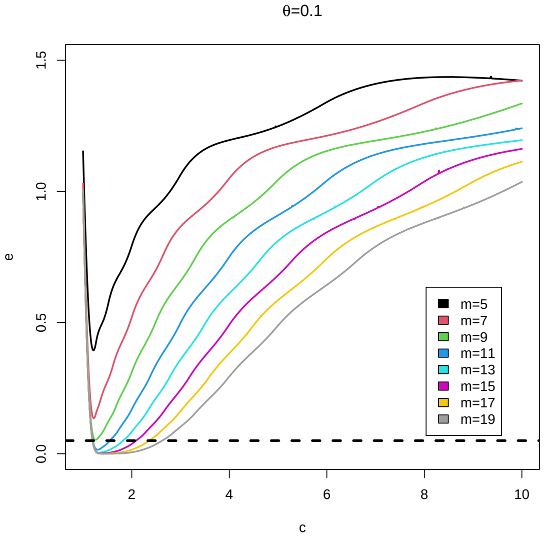


Figure 2.6: Selecting m. Assume $\theta = 0.1$. Dashed line at $\epsilon_{\max} = 0.05$.

for m is 9, as it is the smallest number of basis functions attaining the maximum tolerable error.

2.4.3 Inferring the length-scale

Suppose that that the value of the length-scale parameter is unknown, such that the inclusion of a prior distribution for its value is desirable. Small values of θ usually require a higher number of basis functions to achieve ϵ_{max} for some value c in the approximation due to the fact that the functions with small values of length-scale are wigglier. This behavior can be seen on Figure 2.7, where the black curve represents $\theta = 0.1$, while the red curve represents $\theta = 0.25$. The minimum value of $e_{0.1}(9, c)$ (black curve) is higher than the minimum value of $e_{0.25}(9, c)$ (red curve), therefore, a conservative approach to deal with this problem would be to select a value of θ which can represent the data in a more adequate manner.

One possible approach is to use the same method as Starling et al. (2019), and Kratz (2006), by applying a formula for the expected number of times a random function crosses its mean. If we give an estimate of what is the maximum number of

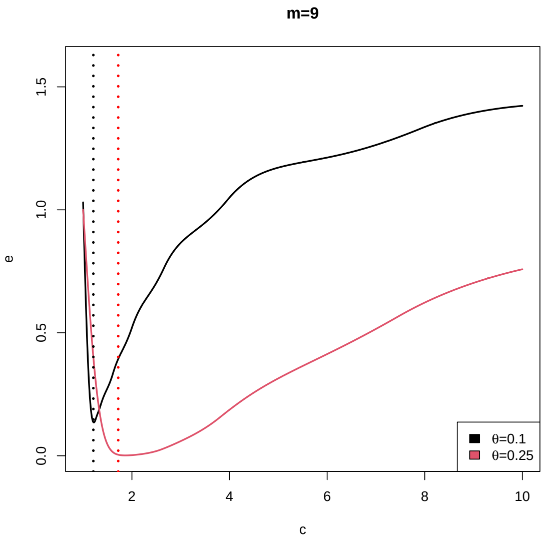


Figure 2.7: The effect of c over the error for different values of θ. m fixed at 9.

times our function would cross its mean, $E_{\max}$, it is possible to adapt the formula to find a $softmin$ of θ, defined as θ_{smin}. The formula is

$$\theta_{smin} = \frac{t_{\max}}{\pi E_{\max}} = \frac{2}{\pi E_{\max}}, \tag{2.15}$$

where $t_{\max}$ is the amplitude of the variable t, but since we are scaling t to be between $[-1, 1]$, then it is possible to substitute its value by 2.

After fixing m (the detailed algorithms to select m and c are detailed in Appendix A.3), we have a sequence of the pairs (c, θ), which can be used to update c every time the value θ is draw from the MCMC. Figure 2.8 show that the relationship between those two parameters is almost linear, so we decided to use interpolation to find the value of c once a new θ has been selected.

Finally, one must select the a prior for θ. Due to the nature of Gaussian processes, greater values of θ could lead to curves that are mostly flat, meaning that the likelihood will be almost the same if the values of the length-scale are draw from the the tails of the prior, so distributions with heavy tails are not desirable. Also, since the number of basis functions has been fixed beforehand, there is a limit for

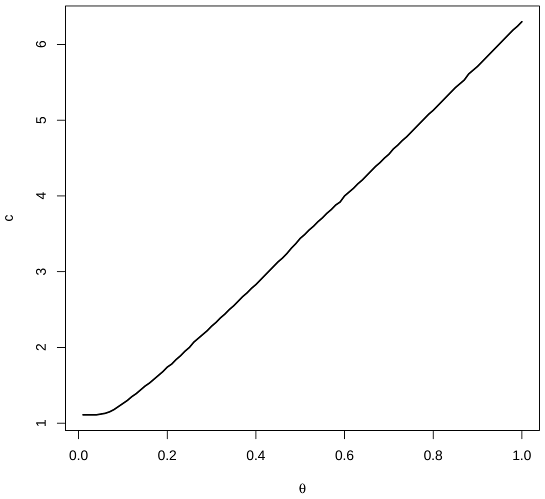


Figure 2.8: c that minimized e for each θ. m fixed at 9.

how much θ can influence how wiggly the Gaussian process is, meaning that even if the density of the prior is high close to zero, it is unlikely that the samples from the posterior distribution will be too close to zero. Having these ideas as base, we decided to use a Half-normal prior, due to the behavior of its tail, as well as how easy it is to find its quantiles. For the selection of the scale parameter of the Half-normal σ_{HNorm}, a procedure similar to the one in Equation (2.15) has been applied, but using a $softmax$ θ_{smax} instead, and the minimum number of times our function would cross its mean, $E_{\min}$. By adapting the formula, we have

$$\theta_{smax} = \frac{2}{\pi E_{\min}}, \tag{2.16}$$

and also

$$\theta \sim \mathcal{HN}(\sigma_{HNorm}), \tag{2.17}$$

so with θ_{smax} we can select the scale of the Half-normal prior given a quantile F, such that

$$\sigma_{HNorm} = \frac{\theta_{smax}}{\sqrt{2} \times erf^{-1}(F)}, \tag{2.18}$$

where erf^{-1} is the inverse error function. In the examples presented in this study $F = 0.9$ is used.

2.5 Related work

Chipman et al. (1998) introduced the Bayesian CART, alongside the ideas for the tree priors, regularizing the depth of a tree, as well as their selected variables and cutpoints. This idea was later expanded by Chipman et al. (2010) into the Bayesian Additive Regression Trees (BART), where contributions from small trees provide a more flexible model for predictions.

Hill (2011) applies BART in a causal inference setting, estimating average treatment effects for binary outcomes. Hahn et al. (2020) expands BART causal inference application into the Bayesian causal forest (bcf), a model that uses two BART priors, one to estimate the prognostic effect and one for the treatment effect.

Starling et al. (2019) introduces a BART prior with targeted smoothing (tsBART), maintaining the structure of BART trees, using a Gaussian Process prior on the leaf nodes to create a smooth function of a selected variable. This method is later expanded by Starling et al. (2020) using the bcf framework, creating the tsbcf, a targeted smoothing approach focused on causal inference. One drawback of these approaches is that using a GP prior is computationally costly, meaning that this framework is not scalable when the targeted smoothing variable has a large number of unique values.

One approach to improve the computational efficiency of a Gaussian Process is the reduced-rank method developed by Solin and Särkkä (2020), where the covariance function of a GP is approximated using a set of basis functions, providing a fundamental piece for the development of our scalable tsBART prior, discussed in Section 2.4. Riutort-Mayol et al. (2020) presents recommendations for the selection of parameters of the reduced-rank method, focusing on their relationships and providing applied examples.

2.6 Simulation studies

2.6.1 Scalability gains

Since an approximation is used, the practitioner must think carefully if it makes sense to use it. In the case of our approximation of the covariance function, the main reason for the use of these kinds of models is regarding the time spent training the model due to the fact that fitting a Gaussian process requires $\mathcal{O}(|\mathcal{T}|^3)$ matrix operations for MCMC inference. For small values of $|\mathcal{T}|$, the practitioner might want to use the exact covariance function instead since the difference in time is low. It must be noted, however, that since the error can be controlled, there is little drawback from using the approximation. Solin and Särkkä (2020) and Riutort-Mayol et al. (2020) have a few comparisons of the general behavior of the approximation for different numbers of basis functions.

To provide guidance to the time savings that the approximation achieves, therefore expanding the tools available to help the practitioners to make the decision of what model should be used, a simulated example from Starling et al. (2019) has been adapted.

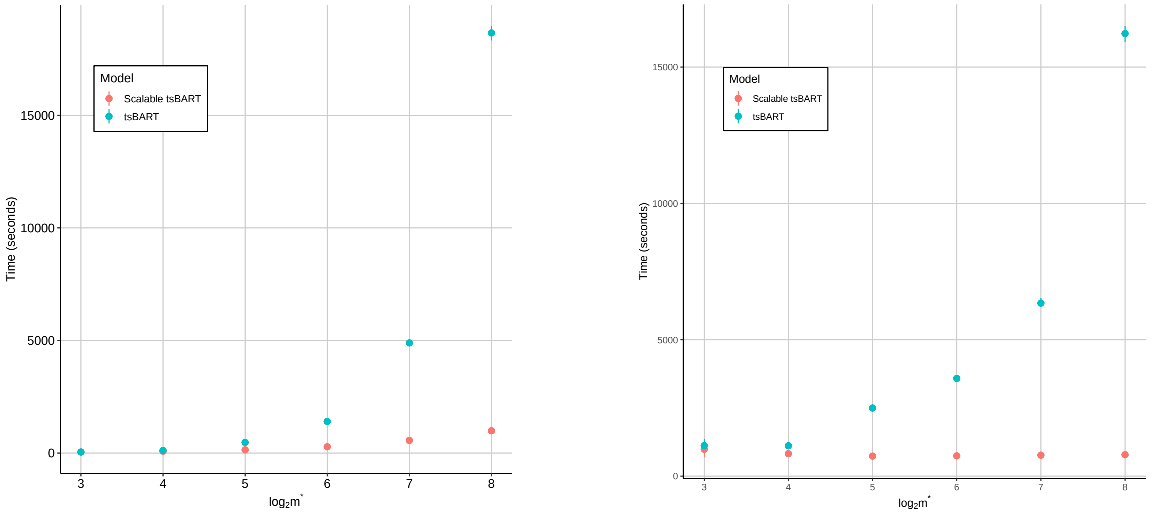
In the original example, $\mathcal{T} = \{1, ..., 8\}$, but since the idea of this example is to expand the size of $|\mathcal{T}|$, real numbers between 1 and 8 are inserted equally spaced among the possible values of $\mathcal{T}$. For simplicity, let us define $|\mathcal{T}| = m^*$. Since the idea is to increase the size of m^* until the estimation of the the original tsBART becomes unfeasible, the size of m^* is increased in log_2, which means its value is doubled for each set of simulations.

The covariates are sampled such that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \begin{pmatrix} X_3 \\ X_4 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right),$$

for $n = 100$, where each observation is associated with m^* response values. The response is generated as

$$y_i(t, \boldsymbol{x}_i) = (x_{i1} + x_{i2}) \times \cos\left(t + 2\pi(x_{i3} + x_{i4})\right) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, 1).$$



(a) Variable sample size: As the number of basis functions increase, the number of subjects remain the same, increasing the sample size.

(b) Fixed sample size: As the number of basis functions increase, the number of subjects decrease, keeping the sample size fixed.

Figure 2.9: Time comparison between tsBART and scalable tsBART. The x-axis represents $\log_2 m^*$, while the y-axis represent the time in seconds. The vertical bars represent the lowest and highest values of the 10 simulations for a specific model at a specific $\log_2 m^*$ value, however some of the values are so similar that the bars are covered by the points, which represent the average of those 10 simulations.

10 parallel simulations with 1000 burn-in samples and 1000 saved posterior samples were performed for each size of m^*, therefore simulating a short chain for demonstrative purposes. The hyperparameters of the model were set to default, including the number of trees, which was set to 200.

The simulations were carried out up to $m^* = 256$, since the laptop used to run those chains did not have enough memory to run the 10 parallel chains for $m^* = 512$. The simulation was performed in a notebook with a processor Intel Core i9-13980HX, 32GB DDR5 5600 MT/s RAM, Microsoft Windows 11 Home OS, and R version 4.3.1.

The implementations of both tsBART and the scalable tsBART were performed using C/C++ functions through the Rcpp package in R. For tsBART, the original implementation of Starling et al. (2019) has been used.

The results shown on Figure 2.9a clearly illustrate the difference in scalability between the tsBART and scalable tsBART. Even for $m^* = 32$ there is a considerable difference in times, that will be accumulated for long MCMC chains. Our advice is that for any scenarios where the number of basis functions m is lower than $|\mathcal{T}|$, the approximation is preferred due to the accumulated time savings.

2.6.2 Comparing fixed and variable length-scale parameters

One of the contributions of this work is the sampling of a length-scale parameter using MCMC. The drawbacks of doing such an update, based on a few examples that were tested, can be considered negligible, since the increase in MCMC time was of about 10%, which is significantly lower than the scalability gains in Section 2.6.1.

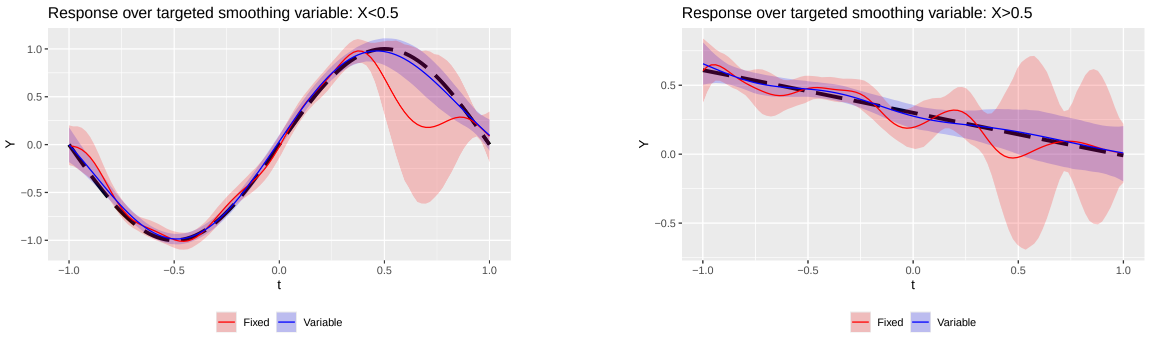
The gains of such an update, however, are noticeable. It is not uncommon that a practitioner does not know the length-scale parameter that should be applied to the dataset that is being analyzed, and that may lead to scenarios where there is a lot of uncertainty, especially since defining a reasonable length-scale is one of the main concerns when fitting any kind of Gaussian Process regression.

Small length-scale values lead to correlation values that decline quickly, meaning that the GP can model wiggly functions due to the high flexibility of the function. On the other hand, large values of length-scale lead to a slow declining correlation, creating functions that vary slower.

Let us analyze a simple toy-example scenario. Suppose that the true data generating process is given as

$$y_i(t, \boldsymbol{x}_i) = \sin(\pi t)\, \mathbb{I}(x_i \leq 0.5) + (\sin(-0.1\pi t) + 0.3)\mathbb{I}(x_i > 0.5) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, 0.1^2).$$

where $x_i \in \{0.25, 0.75\}$ with uniform probability, and t is the scaled (between -1 and 1) version of a variable sampled from the beta distribution $\mathcal{B}(2, 5)$. This example has two different curves that depend on its x values and a large number of unique values for t, which is the kind of structure that the scalable ts-BART prior is designed to capture.



(a) $\sin(\pi t)$

(b) $\sin(-0.1\pi t)$

Figure 2.10: Comparison between scalable tsBART with and without a fixed values for the length-scale parameter. Fixed length-scale is represented in red, while variable length-scale is represented in blue. The x-axis represents the targeted smoothing variable t, while the y-axis represents the response. The dashed black line represents the true data generating function. Solid lines represent the posterior mean over different values of t, with the shaded area representing 95% credible intervals.

Assume two different scenarios. In the first scenario consider fixed length-scale set to be 0.13, which is relatively low, meaning the points have a rapid-declining correlation, with points with a distance of 0.5 (considering the range to be between -1 and 1) having basically a correlation of zero. Let us compare this scenario with the scenario that the length-scale of 0.13 is set as θ_{smin} and θ is updated in each MCMC step using a Half-normal prior.

1000 burn-in samples and 3000 saved posterior samples were performed, therefore simulating a short chain for demonstrative purposes. The hyperparameters of the model were set to default, except the number of trees, which was set to 100.

Figures 2.10a and 2.10b exemplify the difference in behaviors between fixing the length-scale parameter and letting it be updated in each MCMC step. In figure 2.10a the true function requires a length-scale that is lower than the true function of 2.10b, however in both scenarios the fixed length-scale is not ideal for being too low.

On Figure 2.10a the function is reasonably estimated by both scenarios up to $t \approx 0.3$, however due to the lack of high values for t, since those were sampled from

$\mathcal{B}(2,5)$, the curves estimated using the fixed length-scale were way more wiggly in that region, leading to wide credible intervals and a posterior mean reasonably different from the true function, while the curves estimated using length-scales updated at each step could more easily adapt to the changes in the data, leading to a more reasonable fit with tighter credible intervals. Figure 2.10a presents a similar behavior, but now the true function is almost flat, causing the curves estimated using a fixed length-scale to overfit in some level, leading to a fit not as close to true function as using the updated length-scale for each MCMC step. Also, the same wide credible intervals can be observed for $t > 0.3$.

In cases where the length-scale is properly defined, both scenarios yield very similar curves with virtually the same results, but that's rarely the case. Having an instrument that facilitates the work of the practitioner and uses the data to properly estimate the length-scale definitely has value in the estimation of GP curves, which means that, given the small difference in time for the estimation of the curves in both methods analyzed, our recommendation is that the update of the length-scale parameter should be used whenever the real length-scale is unknown.

2.7 Application

2.7.1 Background

In this study, subjects were exposed to stress-inducing events while their total peripheral resistance (TPR) was measured (Yeager et al., 2022). TPR can be seen as an indicator of threat-type responses by measuring the vasoconstriction of the limbs. The idea is that in the event of stressful situations your body is prepared to react to danger, prioritizing blood to more important organs, such as the brain, and this can be measured by TPR. By being able to understand what is happening with the body along the idea that a stressful situation can be viewed as an opportunity to improve your skills, the current stressful situation may not be viewed as a threat.

2.7.2 Dataset

2.7.2.1 Study design

This study aimed to analyze the impact that the introduction of a synergistic mindset treatment would have in people who are undergoing stressful situations. The synergistic mindset is a composition of two kinds of mindsets that are called Growth mindset and Stress-can-be-enhancing mindset. The general idea behind the Growth mindset is that to develop skills, people need to be subjected to difficult situations where those skills will have the opportunity to be honed, meaning that challenges are not supposed to be viewed as negative experiences, but instead as opportunities for people to grow and develop different kinds of skill sets. The Stress-can-be-enhancing mindset argues that stress can be beneficial to the person, since signals of stress, such as an increased heart-rate or some feeling of uneasiness, can be perceived as your body preparing itself to face challenges, for example by increasing the amount of oxygenated blood to the brain.

After removing any data that presented possible technical issues, the dataset consisted of 153 students from a university social science subject pool, each one receiving a 20 dollar compensation and 2 hours of course credit. 111 (42) of the students declared themselves as female (male), and their age ranged from 18 to 26. 70 subjects from the dataset were part of the treatment group, being exposed to the synergistic mindsets as a possible way to cope with stressful situations through a 30-minute online training program.

The study was based on a Trier Social Stress Test (Kirschbaum et al., 1993). Self-reported variables were recorded before the beginning of the test to set the baseline stress physiology levels of each subject. The participants were subjected to a 21-minute long test. After the first 5 minutes used to record their baseline TPR levels, participants had 3 minutes to prepare a speech about their personal strengths and weaknesses. Once that time had passed, they had to deliver the speech to two evaluators, who provide negative nonverbal feedback. The third part of the test was a 5-minute long surprise math section, where the participants were asked to count backwards starting at 996 in steps of 7, while the evaluators could correct any mistakes that the participants made throughout the test. Finally, the participants had a 3-minute rest period.

2.7.2.2 Description of the variables

The variables used in the model are described on Table 2.1. The first column defines the variable name, the second column contains a brief description of the variable, while the third and fourth columns state if the variable was added among controls and/or moderators.

Variable	**Description**	**Control**	**Moderator**
Sex	Self-reported sex	Yes	Yes
Fixed mindset	Self-reported fixed mindset levels	Yes	Yes
Stress mindset	Self-reported stress mindset levels	Yes	Yes
Both mindsets	Scaled multiplicative mindset interaction	Yes	Yes
Self-esteem	Self-reported self-esteem levels	Yes	No

Table 2.1: Description of the covariates included in the model

2.7.3 Proposed model

TPR is the variable of interest and considered here as the main measurement of stress-type responses. The variable went through a simple transformation by subtracting the average TPR from the baseline epoch of each subject from the recorded TPR of the remaining epochs. The changes in TPR are expected to be smooth in time, such that the variable Time has been selected to be the variable where the targeted smoothness is applied. The Time variable has also been transformed to be scaled between -1 and 1 in the model. The vector $\boldsymbol{x}_i$ is composed of the self-reported mindsets variables, the prior self-esteem, and sex, while the vector $\boldsymbol{w}_i$ vector is composed of the self-reported mindsets and sex for the ith individual of the dataset. The prognostic and the treatment effects, are represented by the scalable tsBART priors (developed on Section 2.4) φ and ψ respectively. Finally, a random effect α_i is added to capture the effects specific from each subject. Therefore, our problem can be written as

$$y_{i,t} = \alpha_i + \varphi(\boldsymbol{x}_i, t) + \psi(\boldsymbol{w}_i, t) z_i + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma^2). \tag{2.19}$$

2.7.3.1 Calibrating the approximation

The proposed model is only worth using if there is an advantage when comparing $|\mathcal{T}|$ and m. In this case, $|\mathcal{T}| = 16$, since there are 21 minutes for which TPR was recorded for every subject of the study with the first 5 minutes used for baseline, which means that a matrix of 16 × 16 needs to be inverted to solve the Gaussian process regression for the MCMC inference on each leaf of the model. To analyze if it is advantageous to use the scalable tsBART prior, we must first select m. The advantage of using the proposed model is noted for values of m lower than 16.

First, it is necessary to select a reasonable number for the maximum expected number of crossings to estimate θ_{smin}. The selection must be done for both tsBART priors φ and ψ. For this, the expected behavior of the outcomes can be taken into account. It is expected that the levels of TPR will first rise until a peak at the Speech phase, followed by a decrease as this phase ends. Once the unexpected Math phase begins, the same behavior is expected, then returning to the original base level at the recovery period. This means that, in general, the outcome is expected to cross the average around four times. This behavior should not be so different for any of the priors, so values above four can be seem as reasonable values for the maximum expected number of crossings, therefore we can use

$$\theta_{smin} = \frac{2}{\pi \times 4} \approx 0.1591549.$$

The remaining parameters of the approximation are selected to control the error and can be selected to be as low as possible (or as high as possible in the case of $c_{\max}$). In this study we selected $\epsilon_{\max} = 0.01$, $\textit{step} = 0.05$, $\textit{step}_\theta = 0.05$, $\theta_{\min} = 0.01$, and $c_{\max} = 10$.

Finally, Algorithm 2 resulted on $m = 13$ for both scalable tsBART priors, meaning that there there is a time advantage on using the developed methods for this study, since the number of selected basis functions is lower than the number of unique values that the selected variable of the targeted smoothing can assume.

2.7.3.2 Priors and hyperparameters

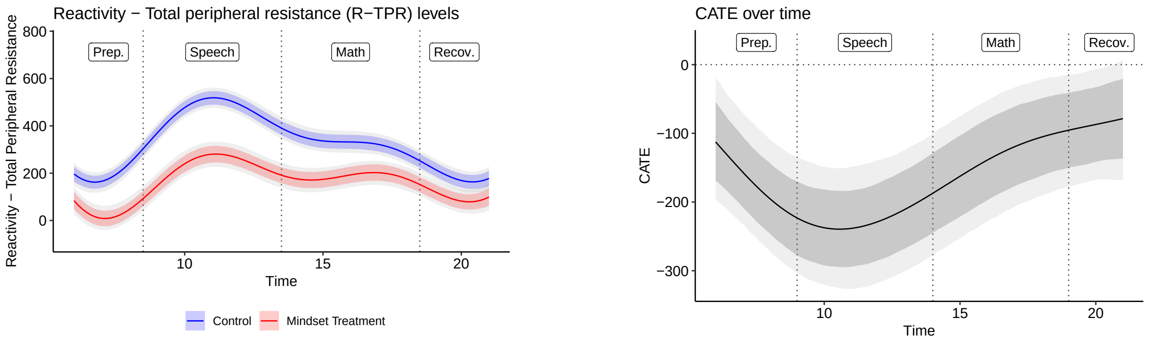
Regarding the selection of the hyperparameters of the model, a possible first step is to set θ_{smin} and θ_{smax} to select the hyperparameter related to the prior of the length-scale θ. By using the proposed method introduced on Equation (2.16) it is possible to select extreme scenarios for the expected number of crossings, A reasonable "extreme" case scenario could be seem as when the expected number of crossings is 1, since 0 would mean that we are modeling flat curves, meaning that $\theta_{smax} \to \infty$. In this study, setting $\theta_{smax} = 2/\pi$ and $F = 0.9$ leads to $\sigma_{HNorm} \approx 0.3870373$.

The tsBART priors can have their hyperparameters and hyperpriors set separately due to their independence from each other, so all priors related to trees can be adapted as needed. For the prognostic tsBART prior, we set 250 trees, $\alpha = 0.95$ and $\beta = 2$, following the default BART prior. For the tsBART prior related to treatment, following Hahn et al. (2020) advice, the tree depth is penalized to shrink towards additive treatment effects, and this was done by setting 50 trees, $\alpha = 0.25$, and $\beta = 3$.

The Gaussian process prior remains almost the same as the tsBART prior, being

$$\mu_{b_k j}(t) \sim GP\left(0, \frac{\zeta}{J} C_{\theta c}^{m}\right),$$

where $C_{\theta c}^{m}$ is the basis function approximation of the squared exponential kernel where each entry is specified on Equations (2.8) and (2.9). The scale parameter ζ is assigned a half-Cauchy prior such as in Starling et al. (2019), Hahn et al. (2020), and Gelman (2006). The σ prior also follows a χ^2 distribution, following Chipman et al. (2010) approach. Random effects were also added to account for individual-level heterogeneity



(a) TPR - Estimated curves for the average treatment and control over time.

(b) Conditional Average Treatment Effect (CATE) curves over time (in minutes).

Figure 2.11: Estimated TPR and CATE curves over time.

in TPR.

2.7.4 General results

Using the settings proposed on Section 2.7.3.2, the model with scalable tsBART priors has been fitted, setting the first 150000 posterior draws treated as burn-in, 10 as thinning, and 3000 as the final number of posterior draws.

Figure 2.11a shows the estimated resulting curves for TPR reactivity over time, where the blue (red) curve represents the average control (treatment) estimated responses over time, with the light blue (red) bands representing the 80% and 95% credible intervals. As expected, peaks are observed in the estimated TPR curves at both speech and math sections. The curves of treatment and control have different rates of change for TPR, which indicates that the treatment might have been heterogeneous, alongside the fact that since the curves do not overlap, the treatment must have been effective in comparison to the control group.

Since the treatment effect seems to be heterogeneous over time, there is an indication that the treatment effect might be stronger on more stressful situations. Figure 2.11b shows that the CATE was indeed stronger in the speech section, returning close to the base level in the recovery period. This is an evidence that the

30-minute online treatment might have presented benefits to the participants that were exposed to synergistic mindsets, also indicating that knowing how your body works under stress might lead to better ways to cope with stressful situations when these events arise.

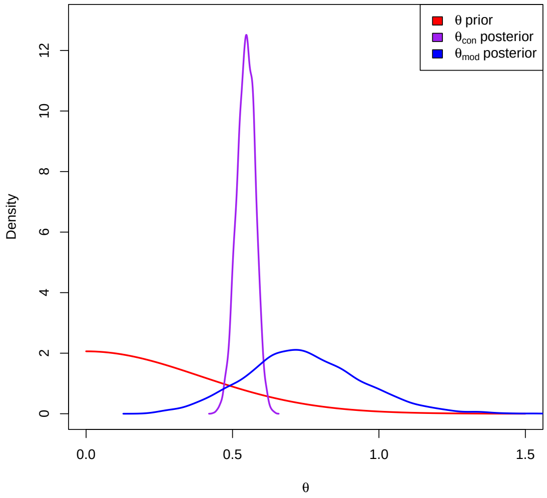


Figure 2.12: Density comparison between the prior and posterior distributions of θ parameters.

Regarding the length-scales, as seem on Figure 2.12, the posterior for the control curve (θ_{con}) is mostly concentrated and generally lower than the treatment effect curve (θ_{mod}). This means that the control curve is expected to be more wiggly, while the treatment effect curve is expected to be more flat. As seem on Figures 2.11a and 2.11b, this is the general behavior observed.

2.7.5 Heterogeneity analysis

2.7.5.1 Subgroup analysis

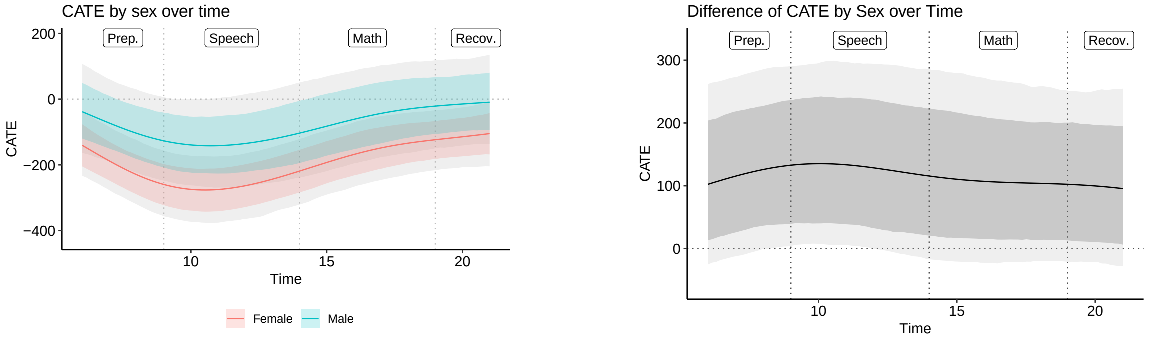
The analysis of possible heterogeneous treatment effects may uncover some underlying story that may gives us more information about the expected effectiveness of the treatment. In our case, three different approaches were utilized to select these groups.

The first approach was trying to find simple subgroups that may have some kind of difference in the treatment effects. Analyzing the variable *sex* is straightforward, but still relevant, since it is a single binary variable, making it easier to evaluate the results while still providing information about of a relevant part of the population.

On Figure 2.13a the CATE curves for Male (in blue) and Female (in red) can be observed. The solid line represents the posterior mean of the estimated CATE over time. The color-shaded areas represent the respective 80% credible interval for each curve, while the shaded-grey area for each curve represents the 95% credible interval. As the curve approaches the Speech epoch, the absolute estimated CATE becomes larger, and the effect is dissipated as the curves approach the Recovery epoch. The credible interval bands of the CATE curve of the Female subgroup remained below zero over all four epochs, while the for the Male subgroup, zero was observed to be within the 80% credible intervals for all epochs, except for Speech, where zero mostly remained between the 80% and 95% bands. This result suggests that the treatment is more effective on females than in males, but it is necessary to evaluate if the CATE curves are actually distinct from each other enough to corroborate the hypothesis that the conditional average treatment effect differs from each subgroup.

Since the credible interval bands of both curves overlap, calculating the difference of CATE curves is a simple way of evaluating if the subgroups are indeed different. Figure 2.13b shows that despite the fact that zero is included in the 95% credible interval for most of the time, it is outside the 80% credible interval for the whole period. Furthermore, it is possible to notice that the difference is slightly larger as the curve gets closer to the beginning of the Speech epoch.

Since the estimated difference in CATE is continuously changing over time, a possible approach is to the the average of estimated differences of each epoch for each posterior draw and evaluate the distribution. Figure 2.14 uses violin plots to



(a) Conditional Average Treatment Effect (CATE) curves over time (in minutes) for each sex.

(b) Difference in Conditional Average Treatment Effect (CATE) curves over time (in minutes) for sex subgroups.

Figure 2.13: CATE and difference in CATE curves by sex.

estimate the posterior probability that the difference in CATE for each epoch is larger than zero. For all epochs, this probability is over 90%, with the largest probability being recorded by the Speech epoch (98%). This means that the practitioner can be quite confident that there is a non-zero difference in the conditional average treatment effect between males and females, where the average female seems to benefit more of the treatment than the average male.

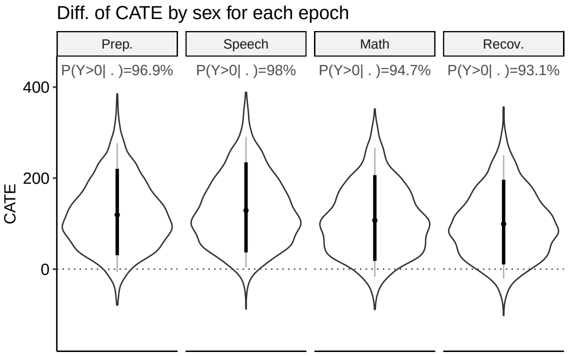


Figure 2.14: Difference in Conditional Average Treatment Effect (CATE) for sex subgroups by epoch average.

A second idea of subgroups was explored based on the assumption that, since the treatment of this study is a change in the mindset of students, students whose mindsets were more pessimistic would have a stronger treatment effect than the students who had more optimistic mindsets. The mindsets were measured with the variables *Fixed mindset* and *Stress mindset*. People who scored lower on these two variable are the people who are in the group defined as having *negative prior mindsets*, while people who scored higher in both variables are in the *positive prior mindsets* group.

One important detail is that a thresholds should be delimited to define which observations belong to each group. Figure A.1 shows the general idea of how the groups would be selected. The left and the center image showcase examples of valid groups, while the right image shows an example of invalid thresholds, they would create subgroups with overlapping observations (the observations in the central quadrant belong to both groups), which is not desired. The black dotted lines are the observations selected for each threshold, while the solid red represent the boundary of the selected subgroups, where the bottom left boundary represents the *negative prior mindsets* subgroup, and the upper right boundary represents the *positive prior mindsets* subgroup.

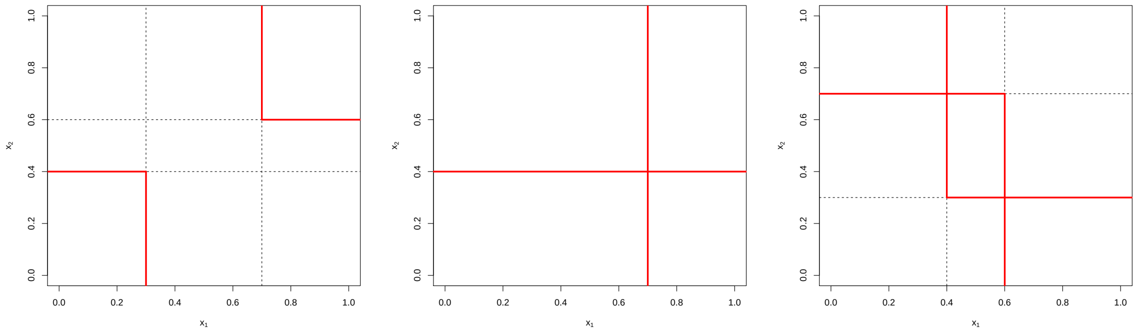


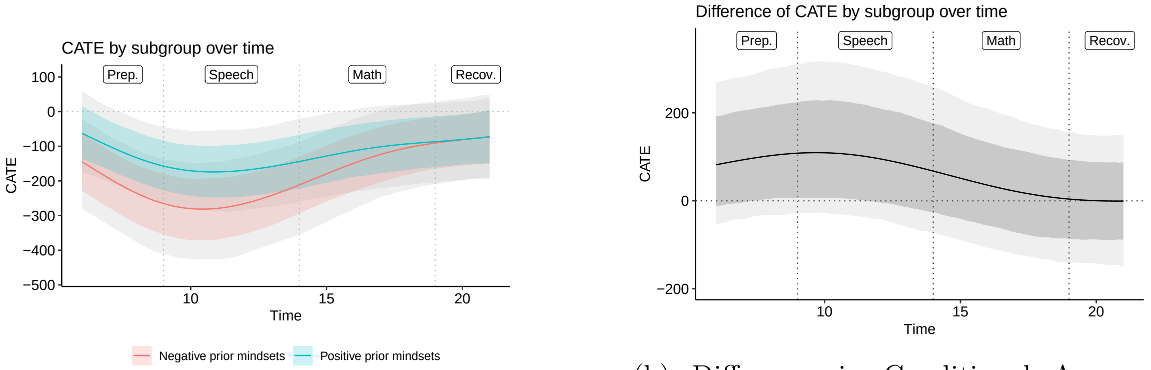
Figure 2.15: Types of subgroups: (left) values of x_1 ans x_2 selected for each subgroup are distinct, and selected such that no observation is in both groups at the same time; (middle) values of x_1 ans x_2 selected for each subgroup are equal, and selected such that no observation is in both groups at the same time; (right) values of x_1 ans x_2 selected for each subgroup are distinct, but some observations are in both groups at the same time, therefore, these subgroups are invalid.

Performing a search over all valid partitions of the data is a simple but effective way of going through all possible scenarios to select the thresholds for the groups. Group with less than 10% of the sample size have also been considered invalid, to avoid groups that were composed of only a few observations. For the response variable of the search, the posterior mean of the estimated prognostic effect has been used. Also, since only 153 subjects are present in the dataset, the epochs have been aggregated by taking the average of the prognostic effect over time, and the Speech epoch has been since it is the epoch where the treatment effects seems to be stronger. The selected set of subgroups was the one where the groups had largest absolute difference in the average response. The *positive prior mindsets* selected threshold was 3.25 for the Fixed mindset variable and 2.1875 for the Stress mindset variable, while the Fixed and Stress mindsets threshold for *negative prior mindsets* were 3.25 and 3.0625, respectively.

Figure 2.16a shows the CATE curves over time for each of the subgroups found. As expected, the *negative prior mindsets* groups shows a stronger treatment effect in comparison to *positive prior mindsets*, specially in Speech. Over time, the difference in subgroups starts to wane, to the point where the curves basically converge in the Recovery epoch. It is not completely clear whether or not there is a difference in the subgroups since the credible interval bands overlap, so looking at the difference in CATE curves is suggested.

The curve of the difference in CATE curves can be seen on Figure 2.16b. As previously discussed, the difference is stronger during the Speech epoch and wanes as the curve approaches the Recovery epoch, so aggregating each epoch by taking the average over each epoch and looking at the probability that the difference bigger than zero gives us a better view of the scenario.

On Figure 2.17 the estimated probability that the difference is bigger than zero is 90.2% and 90.6% in the Prep. epoch and Speech epoch respectively, which might seem like a reasonably larger number and it seems to corroborate our expectations of



(a) Conditional Average Treatment Effect (CATE) curves over time (in minutes) for the *positive prior mindsets* and *negative prior mindsets* subgroups.

(b) Difference in Conditional Average Treatment Effect (CATE) curves over time (in minutes) for the *negative prior mindsets* and *positive prior mindsets* subgroups.

Figure 2.16: CATE and difference in CATE curves by prior mindsets.

what would happen in these subgroups. The probability then decreases up to a point where it is basically a coin toss on the Recovery epoch. This result implies that it is likely that people with mindsets that are mostly pessimistic have the most to gain from the proposed treatment in comparison with people that already had optimistic mindsets.

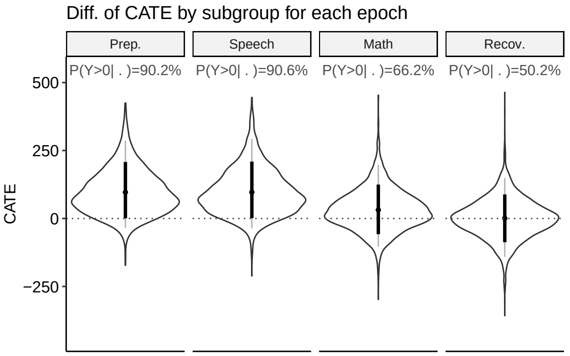


Figure 2.17: Difference in Conditional Average Treatment Effect (CATE) for the *negative prior mindsets* and *positive prior mindsets* subgroups by epoch average

2.7.5.2 Reported results

In general, the results observed in the analysis were promising, with the estimated CATE curves showing the behavior that would be expected from a successful treatment. One important aspect that was observed is that the treatment is specially effective in the most stressful part of the experiment, meaning that more stressful situations, where an individual may feel overwhelmed, can be more easily dealt with using a simple and effective treatment.

Another interesting point is that groups that are considered more prone to stress, such as people in the *negative prior mindsets* subgroup presented a stronger treatment effect, meaning that people that are in a more vulnerable position regarding their mindsets a more meaningful impact in comparison to people with more solid previous mindsets. A stronger treatment effect was also observed in females in comparison to males, meaning that sex also plays a role in how effective the treatment is.

These results are important to better understand the population that would benefit the most from this treatment, being specially beneficial to people with mindsets that are more negative and therefore are in a more vulnerable position, and females. The simplicity of the treatment and its low cost help advocate for more studies regarding this question and even its implementation in a large scale setting, as in schools, for example.

2.8 Discussion

The targeted smoothing BART prior is a useful tool for a wide variety of applications. In our analysis, the model performed as expected, corroborating the results originally found by Yeager et al. (2022) while expanding the analysis of subgroups with promising heterogeneous treatment effects.

In addition, a method to approximate the covariance kernel of Gaussian Processes was studied and developed to adapt to the ts-BART, solving the scalability problem that existed in the original model. This method can now be adopted to expand the use of ts-BART and ts-bcf models to different experiments, with a targeted smoothing variable with a large number of unique values due to the newly developed scalable ts-BART prior. Since this prior can now be further expanded to deal with continuous values of t, meaning that a continuous treatment version of bcf is now feasible due to the techniques developed in this study.

Furthermore, the selection of a length-scale parameter has been facilitated with the advent of a Bayesian method for sampling the parameter using $softmin$ and $softmax$ values. This approach helps practitioners to deal with the problems of using a fixed length-scale to GP regression methods when the parameter is unknown.

Chapter 3: LineART: Locally adaptive linear Bayesian additive regression trees

Abstract

Bayesian Additive Regression Trees (BART) priors have shown promising results, especially for prediction, being used in a wide range of different problem settings. BART priors represent functions as the sum of many small trees, each parameterizing a step function. However, smoothness is often desirable for parsimony, accurate interpolation, and extrapolation, while keeping the efficiency. We expanded the BART model by replacing locally constant predictions with locally linear predictions, regressing on the variable attached to the parent node. This is a special case of a general class of BART models with vectors of basis coefficients attached to the leaves.

3.1 Introduction

The advent of Bayesian additive regression trees (BART) by Chipman et al. (2010) introduced a notable tool that provided better predictive performance than some of the most used machine learning models, such as Random Forests (Breiman, 2001), Boosting Friedman (2001), and LASSO (Tibshirani, 1996).

Many BART extensions have been proposed to account for more complex relationships among the covariates. One of these extensions is the SoftBART prior (Linero and Yang, 2018), which, despite its superior predictive power due to its more complex structure, has issues of scalability when dealing with large sample sizes.

Aiming for a more flexible prior, this study extended BART by performing a change on the leaf nodes to allow a simple linear regression as a response using a locally adaptive linear BART (LineART) prior. The new prior allows more flexibility while maintaining the scalability found in BART. This prior also provides a natural extension of the Bayesian causal forest (bcf) model (Hahn et al., 2020).

Section 3.2 introduces the novel LineART prior, providing the mathematical details of its implementation; Section 3.5 provides a performance comparison with BART and SoftBART, as well as results regarding the predictive performance and timing trade-off; Section 3.6 extends LineART to the causal inference setting with lbcf, revisiting some of the studies performed by Yeager et al. (2022); lastly, Section 3.7 provides a discussion about the general results and the usefulness of the LineART prior.

3.2 Methods

Let $\boldsymbol{x}_i \in \mathcal{X}$ be the vector of covariates related to the ith individual $\{i \in 1, 2, \ldots, n\}$ of a dataset. Let $y_i \in \mathcal{Y}$ be the scalar response associated with the ith individual of the dataset. The problem then is of the form

$$y_i = \phi\left(\boldsymbol{x}_i\right) + \epsilon_i,$$

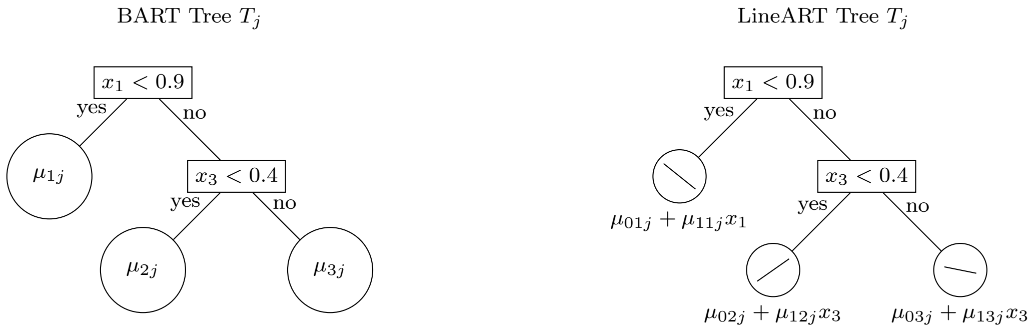
where $\phi(.)$ is a function that can capture the relationship between the covariates and the scalar response, and ϵ_i is the error term.

3.2.1 Locally adaptive linear BART

By using the framework proposed by Chipman et al. (2010),

$$\phi\left(\boldsymbol{x}_i\right) = \sum_{j=1}^{J} g\left(\boldsymbol{x}_i, T_j, \Pi_j\right),$$

where each $g(\boldsymbol{x}_i, T_j, \Pi_j)$ is a Bayesian CART as defined in Chipman et al. (1998), meaning that the framework is composed of a collection of binary trees $\{T_j : 1 \leq j \leq$



(a) Representation of a regular BART Tree.

(b) Representation of a LineART Tree.

Figure 3.1: Comparison between different regression trees.

$J\}$ and associated leaf node parameters $\{\Pi_j = \mu_{1j}, \dots, \mu_{B_j j}\}$ over the B_j partitions of T_j $\{B_j = |\mathcal{B}_j| : b_{1j}, \dots, b_{B_j j} \in \mathcal{B}_j, b_{b'j} \cap b_{b''j} = \emptyset, \forall b' \neq b''\}$. This sum-of-trees approach is known as Bayesian additive regression trees (BART).

One of the characteristics of the BART prior is that every single observation that ends in the same leaf node will have exactly the same response for that specific tree. By the nature of the model, it may be difficult to approximate some functions, as for example, a linear function with a slope different from zero. In this study, $\phi(.)$ is a locally adaptive linear BART (LineART) prior of the form

$$g\left(\boldsymbol{x}_i, T_j, \Pi_j\right) = \sum_{k=1}^{B_j} \mathbb{I}\left(\boldsymbol{x}_i \in [b_k]\right) \boldsymbol{x}_{ib_k j} \boldsymbol{\mu}_{b_k j}$$

where $\{\Pi_j = \boldsymbol{\mu}_{1j}, \dots, \boldsymbol{\mu}_{B_j j}\}$ is the set of parameters for each leaf node b_k, and $\boldsymbol{x}_{ib_k j}$ is a 2×1 vector with 1 as the first element, and the second is the element of $\boldsymbol{x}_i$ that is used in the locally adaptive linear regression.

3.3 Prior details and edge cases

3.3.1 Proposing a prior for the leaf parameters

Let $y_{b_{kj} i}$ denote the ith observation in the terminal node $b_{kj}, k \in \{1, 2, ..., B_j\}$ in a tree. In this example the tree index j and the partition index k are omitted for simplicity. It is assumed that for a terminal node $b \in \{1, \ldots, B\}$,

$$\boldsymbol{y}_b {\sim} \mathcal{N}\left(\Omega_b \boldsymbol{\mu}_b, \sigma^2 \boldsymbol{I}\right), \qquad \boldsymbol{\mu}_b {\sim} \mathcal{N}_2\left(\mathbf{0}, \Sigma_b\right),$$

such that $n_1 + n_2 + ... + n_B = n$, $\boldsymbol{y}_b = \left(y_{b1}, y_{b2}, ..., y_{bn_b}\right)^T$, $\boldsymbol{\mu}_b = \left(\mu_{b0}, \mu_{b1}\right)^T$, and Ω_b is a matrix $n_b \times 2$, where the first column is composed of 1, and the second column is composed of one of the numeric covariates in the dataset.

In this framework, the variable can be chosen in few different scenarios. The first scenario occurs if the tree is a stump (a tree without any splits, composed of only one leaf node). For this case, a variable is selected by using a discrete uniform distribution among the available covariates. The second scenario occurs if the tree has at least one split, on which case the variable used on the split of the parent of the leaf node will be used as a covariate.

Let us assume only numeric covariates and that the leaf node b has a parent node with split given by $X_d < c$, where $d \in \{1, ..., p\}$, then in this case Ω_b will be given by

$$\Omega_b = \begin{bmatrix} 1 & X_{d1} \\ 1 & X_{d2} \\ 1 & X_{d3} \\ ... & ... \\ 1 & X_{dn_b} \end{bmatrix}.$$

The general idea for the prior $\mathbb{P}(\boldsymbol{\mu}_b)$ is to control the total contribution of each tree to the prediction, so instead of regularizing each parameter separately, let us focus on the contribution of a specific tree j to the fit of observation i.

For partition b let us introduce a prior that uses the design matrix Ω_b as a way to add information to the prior similar to a g-prior (Zellner, 1986), giving us

$$\boldsymbol{\mu}_b \sim \mathcal{N}\left(\mathbf{0}, \frac{k}{m}\left(\Omega_b^T \Omega\right)^{-1}\right).$$

It is well-known that for $\mathbb{P}(\boldsymbol{\mu}_b) \sim \mathcal{N}\left(\mathbf{0}, \Lambda\right)$ the posterior distribution of the parameters would be given by

$$\mathbb{P}(\boldsymbol{\mu}_b | \boldsymbol{y}_b) \sim \mathcal{N}\left(\left(\frac{\Omega_b^T \Omega_b}{\sigma^2} + \Lambda^{-1}\right)^{-1} \frac{\Omega_b^T \Omega_b}{\sigma^2} \hat{\boldsymbol{\mu}}_b, \left(\frac{\Omega_b^T \Omega_b}{\sigma^2} + \Lambda^{-1}\right)^{-1}\right),$$

where $\hat{\boldsymbol{\mu}}_b = \left(\Omega_b^T\Omega_b\right)^{-1}\Omega_b^T\boldsymbol{y}_b$, therefore by using our proposed prior, we have that the posterior of $\boldsymbol{\mu}_b$ is given by

$$\mathbb{P}(\boldsymbol{\mu}_b|\boldsymbol{y}_b) \sim \mathcal{N}\left(\frac{k}{m\sigma^2+k}\hat{\boldsymbol{\mu}}_b, \frac{k}{m+k/\sigma^2}\left(\Omega_b^T\Omega_b\right)^{-1}\right).$$

3.3.2 Comparison with another prior

Another way of looking at this prior is by analyzing the prior on the prediction of an observation. Let us take the ith observation from partition b as an example (only a single tree is considered). We have that for a given x_i

$$\begin{pmatrix}1 & x_i\end{pmatrix}\begin{pmatrix}\mu_{b0}\\ \mu_{b1}\end{pmatrix} = \phi_b\left(x_i\right) \sim \mathcal{N}\left(0, \frac{k}{m}\left(\frac{1}{n_b} + \frac{\left(x_i - \bar{x}_b\right)^2}{\sum_r\left(x_r - \bar{x}_b\right)^2}\right)\right),$$

and since the term

$$\frac{\left(x_i - \bar{x}_b\right)^2}{\sum_r\left(x_r - \bar{x}_b\right)^2}$$

is on average $1/n_b$, then on average the variance of $\phi_b(x_i)$ is $(k/m)(2/n_b)$. Since the response variable is scaled, we are setting $k = \tau^2 n_b/2$, where τ is a scale parameter, meaning that, on average, the variance is τ^2/m, with a lower bound of $0.5\tau^2/m$ (when $x_i = \bar{x}_b$).

By comparing our proposed prior with a "vanilla" prior such as

$$\boldsymbol{\mu}_b \sim \mathcal{N}\left(\mathbf{0}, \frac{k}{m}\begin{pmatrix}v_0^2 & 0\\ 0 & v_1^2\end{pmatrix}\right).$$

would lead to

$$\phi_b(x_i) \sim \mathcal{N}\left(0, \frac{k}{m}\left(v_0^2 + v_1^2 x_i^2\right)\right),$$

which scales with x_i.

Furthermore, the mixing of our proposed prior has been superior to the "vanilla" prior in some settings. Let us consider four simulated examples, the first one inspired by Friedman and Silverman (1989) is a mix of linear and exponential functions given by

$$y_i = 0.1\exp(4x_{1i}) + \frac{4}{1+\exp(20(x_{2i}-0.5))} + 3x_{3i} + 2x_{4i} + x_{5i} + \epsilon_i.$$

The second one is inspired by Friedman (1991), being a mix of non-linear and linear additive functions

$$y_i = 10\sin(\pi x_{1i}x_{2i}) + 20(x_{3i}-0.5)^2 + 10x_{4i} + 5x_{5i} + \epsilon_i.$$

The third case is a sum of smooth functions (sine, cosine, and power functions), given by

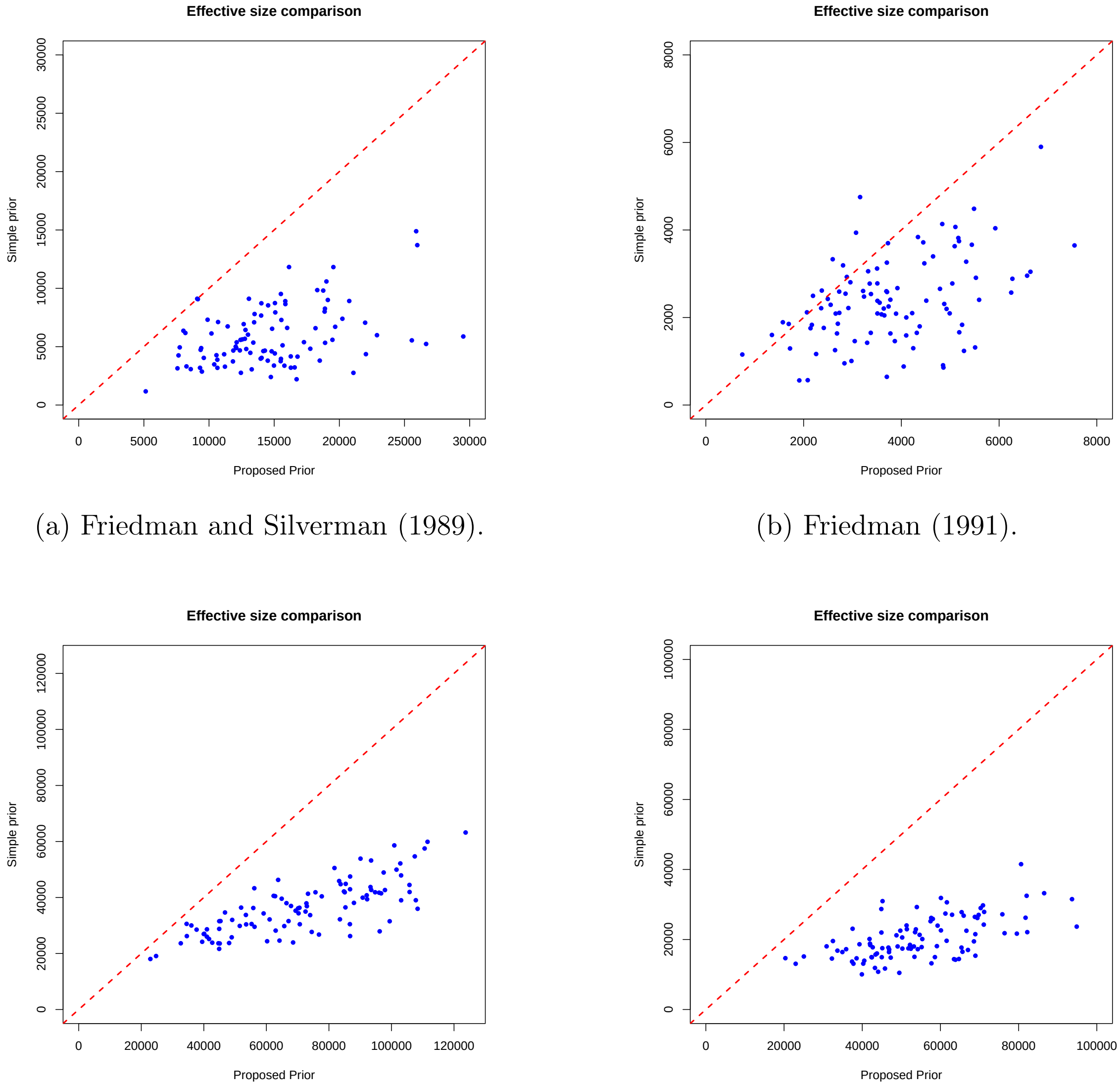
$$y_i = \sin(x_{1i}) + \cos(2x_{2i}) + 0.1x_{3i}^3 - 0.05x_{4i}^4 + 0.2x_{5i}^2 + \sin(x_{1i}x_{2i}) + \cos(x_{3i}+x_{4i}) + \epsilon_i.$$

Finally, the fourth case is an additive linear functions with interactions, given by

$$y_i = 3x_{1i} - 2x_{2i} + 0.5x_{3i} + x_{4i} - x_{5i} + 1.5x_{1i}x_{2i} - 2x_{3i}x_{4i} + 0.8x_{2i}x_{5i} + \epsilon_i.$$

In all cases the 5 covariates ($n = 100$) have been sampled from a standard uniform distribution and $\epsilon \sim \mathcal{N}(0,1)$. 20 chains were sampled for each example, with 50000 draws of each chain discarded as burn-in, and 5000 posterior draws kept as sample for each chain of each model. These samples were, then, aggregated and the effective sample size of the estimated response function was calculated using *effectiveSize()* from the R package *coda* for the estimated samples of models using each prior.

Figure 3.2 contains the plots for each example of the compared effective sample sizes between the "vanilla" prior and the "g-prior". In all examples, our proposed prior showed a better performance than the "vanilla" prior in terms of mixing, since the effective sample size of our proposed prior was, in general, larger.



(a) Friedman and Silverman (1989).

(b) Friedman (1991).

(c) Additive smooth functions.

(d) Additive linear functions.

Figure 3.2: Comparison of effective sample sizes between a more simple "vanilla" prior (y-axis) and the proposed "g-prior" (x-axis). Observations below the red dashed line mean that the effective sample size of the proposed "g-prior" is bigger than the competitor.

3.3.3 Considering edge cases

The previous results assume that the x_i in question is being treated as a numeric variable and that the matrix $\Omega_b^T \Omega_b$ is rank 2. Edge-cases for dummy variables or when $\Omega_b^T \Omega_b$ is rank 1 are treated such that $\boldsymbol{\mu}_b = (\mu_{b0}, 0)^T$. Since $\mu_{b1} = 0$ for that case, it is natural to consider that the prior mean is also zero, and since the value is a constant, it is also reasonable to consider that the correlation between μ_{b0} and μ_{b1} in those cases is also equal to zero, therefore, for those cases we consider the prior

$$\mu_{b0} \sim \mathcal{N}\left(0, \sigma^2_{\mu_{b0}}\right),$$

which is basically the original BART prior (Chipman et al., 2010).

Another possible edge-case scenario is that our linear regression wants to estimate a covariate whose value is way higher than all values in the dataset. In these cases the predicted value of y_i may have a large magnitude that is not aligned with the values that have been observed in the training dataset. To avoid extreme values for the predictions, the linear function is truncated in the lowest and the highest values of every single covariate observed in training, such that

$$\phi_b(x_i) = \begin{cases} \phi_b(x^{(1)}) & \text{if } x_i < x^{(1)}; \\ \phi_b\left(x_i\right) \sim \mathcal{N}\left(0, \frac{k}{m}\left(\frac{1}{n_b} + \frac{(x_i - \bar{x}_b)^2}{\sum_r (x_r - \bar{x}_b)^2}\right)\right) & \text{if } x^{(1)} \leq x_i \leq x^{(n)}; \\ \phi_b(x^{(n)}) & \text{if } x_i > x^{(n)}. \end{cases}$$

Further explaining, there may be scenarios where it is desirable to make predictions out of the range of the training data. In these cases, the predictions will likely show extreme values due to the way that the posterior distribution of $\boldsymbol{\mu}_b$ is calculated, since the variance of the posterior distribution depends on the the inverse of the design matrix, and therefore, whenever the determinant of the design matrix is close to zero, the inverse of the design matrix will be populated with large values, leading to higher levels of uncertainty.

3.3.4 The scale parameter

Regarding the scale parameter τ we can consider that we have

$$y_i = \tau\phi\left(\boldsymbol{x}_i\right) + \epsilon_i$$

and by setting the half normal prior

$$\tau \sim \mathcal{HN}\left(0, \gamma^2\right)$$

we have that

$$\begin{aligned}
\mathbb{P}(\tau|\boldsymbol{y},\gamma) &\propto l\left(\boldsymbol{y}|.\right)\mathbb{P}(\tau|\gamma)\\
\mathbb{P}(\tau|\boldsymbol{y},\gamma) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \tau\phi\left(\boldsymbol{x}_i\right)\right)\right\}\exp\left\{-\frac{1}{2\gamma^2}\tau^2\right\}\\
\mathbb{P}(\tau|\boldsymbol{y},\gamma) &\propto \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n} y_i^2 - 2\tau\sum_{i=1}^{n} y_i\phi\left(\boldsymbol{x}_i\right) + \tau^2\sum_{i=1}^{n}\phi\left(\boldsymbol{x}_i\right)^2}{\sigma^2} + \frac{\tau^2}{\gamma^2}\right]\right\}\\
\mathbb{P}(\tau|\boldsymbol{y},\gamma) &\propto \exp\left\{-\frac{1}{2}\left[-2\tau\frac{\sum_{i=1}^{n} y_i\phi\left(\boldsymbol{x}_i\right)}{\sigma^2} + \tau^2\left(\frac{\sum_{i=1}^{n}\phi\left(\boldsymbol{x}_i\right)^2}{\sigma^2} + \frac{1}{\gamma^2}\right)\right]\right\}
\end{aligned}$$

and therefore

$$\mathbb{P}(\tau|\boldsymbol{y},\gamma) \sim \mathcal{HN}\left(\left(\frac{1}{\frac{\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)^2}{\sigma^2} + \frac{1}{\gamma^2}}\right)\left(\frac{\sum_{i=1}^{n} y_i\phi\left(\boldsymbol{x}_i\right)}{\sigma^2}\right), \left(\frac{1}{\frac{\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)^2}{\sigma^2} + \frac{1}{\gamma^2}}\right)\right).$$

It is also possible to set γ with an inverse-gamma prior to get a half Cauchy prior instead of a half normal. In our case, we set $\gamma = 1$, leading to a half normal distribution.

3.4 Related work

Chipman et al. (2010) introduced the Bayesian Additive Regression Trees (BART) prior, creating a highly adaptive sum-of-trees framework where each regression tree is trained using the residuals of all other trees, allowing the model to capture a wide range of relationships between the covariates and the response.

Since their advent, BART models have been expanded to solve a multitude of problems. Chipman et al. (2010) utilized the Albert and Chib (1993) data augmentation to create a probit BART for classification. To deal with variable selection in high-dimensional settings, Linero (2018) replaced the Uniform prior used to sample variables in the trees with a Dirichlet prior, which updates the probabilities of selecting each variable. Murray (2019) developed a log-linear BART to deal with unordered categorical and count responses using latent variables. Sparapani et al. (2016) adapted BART to survival analysis by also utilizing data augmentation.

The original BART framework, however, has the drawback of utilizing only a scalar at each leaf node to make predictions, meaning that the predictions are non-smooth step-functions. Linero (2022) noted that the non-smooth nature of BART leads to suboptimal function estimates when making predictions for smooth functions. To address this issue, Linero and Yang (2018) introduced the SoftBART prior, which uses soft decision trees, meaning that the partition space is probabilistic, and not deterministic. According to benchmark tests reported in Linero and Yang (2018), the default SoftBART prior exhibited better predictive performance than BART in most settings. The drawback of the SoftBART prior is the high computational cost, leading to a trade-off between predictive power and time efficiency.

Taking time efficiency into consideration, using linear regressions in the leaf nodes is a more efficient method that might lead to results smoother than BART, and therefore, better predictions. This motivated the development of the methods described in Section 3.2.

Other authors also explored similar approaches. Prado et al. (2025) utilizes a semi-parametric BART that is modeled as a sum of linear regression and a BART, while Prado et al. (2021) uses a BART with model trees, fitting a multiple linear regression at each leaf node. This approach, however, has a high computational cost on cases with too many covariates, leading to a considerable increase in the number of parameters that need to be sampled, different from the LineART approach that only selects a single covariate for each leaf node, which is a scalable alternative.

3.5 Performance assessment

One of the possible advantages of the LineART prior is that the use of the locally adaptive linear model inside leaf nodes can more easily capture effects than the original BART prior due to the fact that BART uses a scalar for each leaf node. To evaluate its predictive performance, a default setting for the LineART prior is selected, then compared with BART (Chipman et al., 2010) and SoftBART (Linero, 2022). Naturally, SoftBART is expected to show the best predictive performance overall, but due to the large number of operations performed, the time to run the MCMC chains greatly increases. The speed of the LineART prior is also one of its major advantages due to the low number of operations needed for each MCMC draw (since inverting a 2 × 2 matrix is fast).

3.5.1 Predictive performance

α	β	$ntree$	# of times better than BART
0.50	1	50	448
0.50	1	100	430
0.80	2	100	429
0.50	2	50	424
0.50	2	100	422
0.80	1	50	421
0.50	3	200	420
0.50	2	150	415
0.50	3	100	412
0.50	1	200	406

Table 3.1: Settings of the LineART prior and the number of times that specific setting had a performance better than BART over 820 simulations (41 datasets with 20 simulations each).

First, to evaluate the general predictive performance of the prior, a default set of hyperparameters for the prior is desirable. To select a default that works in a wide variety of settings, the datasets used in the simulations of this section are the same used by Chipman et al. (2010) for their BART evaluation. The following values of parameters were considered:

$$\alpha = \{0.5, 0.8, 0.95\},$$

$$\beta = \{1, 2, 3\},$$

$$ntree = \{10, 50, 100, 150, 200\}.$$

For each dataset, 5/6 of the data has been used for training and 1/6 used for testing, with the data selected for train and test being randomized over 20 replications, with each simulation having the first 10000 MCMC draws treated as burn-in, and 5000 saved MCMC samples with thinning of 3. To compare the results, the RMSE calculated using the posterior mean of the predictions has been compared between each LineART combination of settings and the default setting of BART. Out of the available choices, 9 combinations of settings have shown a better out-of-sample predictive performance than BART, with the top performing set of hyperparameters being chosen as default for LineART. Therefore, the selected default is $\alpha = 0.5$, $\beta = 1$, $ntree = 50$. This combination of hyperparameters was also the best LineART setting in comparison to SoftBART (showed predictive performance superior to SoftBART in 359 out of the 820 simulations). The top 10 hyperparameter combinations are presented on Table 3.1. As a measure of comparison, SoftBART presented lower RMSE than BART in 493 out of the 820 simulations.

The number of trees selected as default is lower than BART, and the hyperparameters selected to define the tree splitting probability makes the initial probability of splitting small trees and stumps (a tree that consist of a single leaf node) lower, which makes sense since the idea of using the LineART prior is to allow the trees to more easily capture the underlying complexity of the data, therefore, not as many trees and splits are necessary to have a performance similar to BART.

One measure to aggregate the predictive performance among all datasets is the relative RMSE (RRMSE) used in Chipman et al. (2010), which is a way to standardize

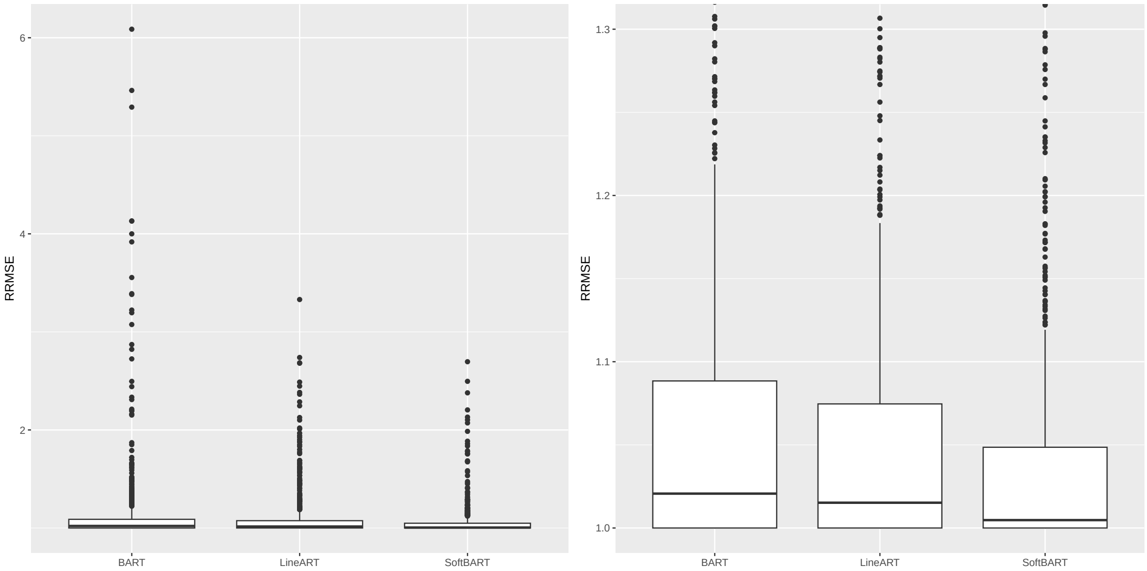


Figure 3.3: Boxplots of out-of-sample relative root mean square error. 820 simulations (41 datasets with 20 simulations each). Left graph includes all extreme points from all competitors. Right graph is a the same graph with y-axis limited between 1 and 1.3. Percentage of RRMSE values over 1.3: BART 10.9%; LineART-default 8.4%; SofBART 4.1%.

the out-of-sample RMSE measures where, for each simulation, the calculated RMSE for all competitors is divided by the lowest RMSE among them. This way, all the 820 simulations can be aggregated. In this case, the competitors are the default settings of BART, LineART and SoftBART. As seen on Figure 3.3, BART and LineART had similar results in general, with BART showing slightly higher values for Q2, and Q3. Besides that, BART had a higher percentage of values over 1.3 (10.9%) in comparison to LineART (8.4%). SoftBART, as expected, presented the best performance out of the competitors.

Analyzing the results by dataset on Figure 3.4 also shows that, in general, the predictive performance of LineART-default and LineART-CV (hyperparameters selected using 5-fold cross validation) are similar, meaning that a practitioner that uses the default hyperparameters should have similar results in comparison to a practitioner that decides to use cross validation. Table B.1 complements Figure 3.4 by

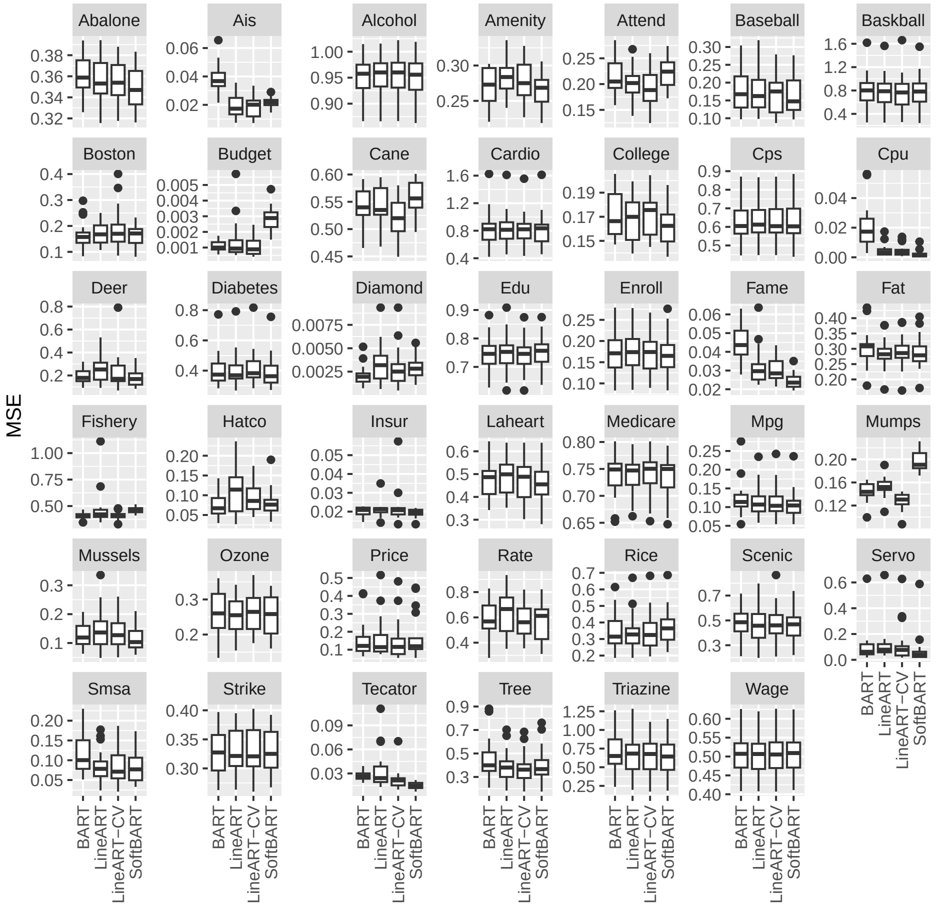


Figure 3.4: Boxplots of out-of-sample mean square error by dataset (20 simulations each) and competitor.

presenting the average out-of-sample MSE for each competitor, alongside their respective standard deviation.

Figure 3.5 shows the average values of the ratio $\log\left(RMSE_{comp}/RMSE_{BART}\right)$, meaning that 0 represent no change between the predictive performance of BART and the competitor, while positive (negative) values represent that the BART (competitor) had a better predictive performance. To assess if the values represent a meaning difference, a two-sided t-test has been performed to test if the average is equal to zero. For LineART, all datasets outside the region covered by the 10% dashed red

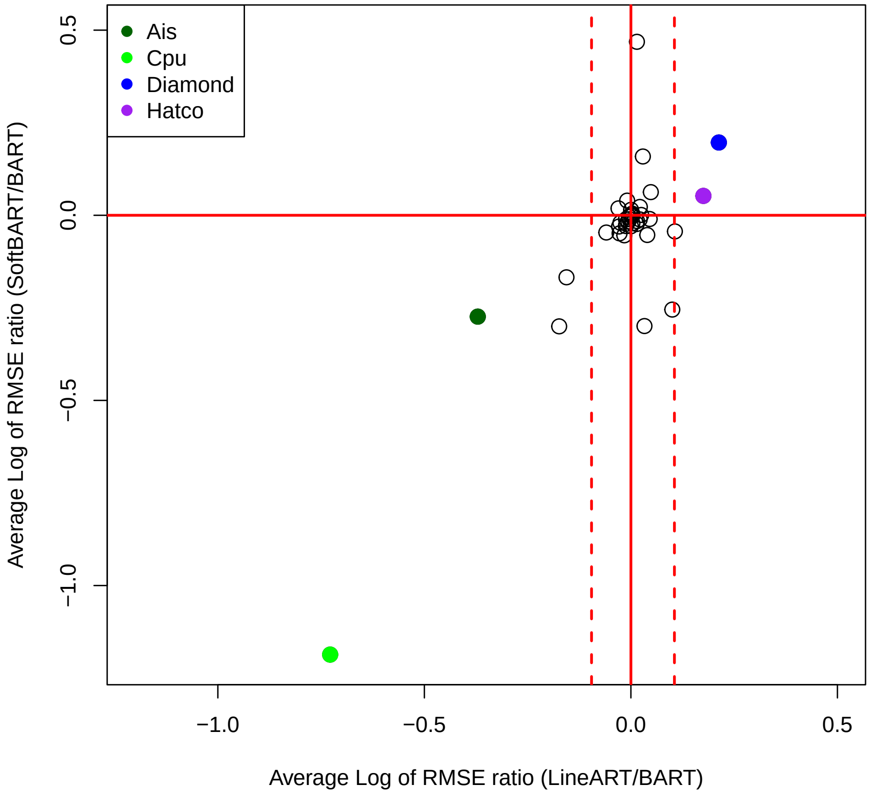


Figure 3.5: Average log of RMSE ratio between competitor and BART for each dataset. LineART is represented in the x-axis, while SoftBART is represented in the y-axis. Dashed red lines represent 10% increase (decrease).

lines presented a p-value lower than 5%. The list of all p-values for each dataset is available in Table B.1.

In most datasets that SoftBART presented a relatively large improvement on predictive performance (Ais and Cpu, for example), LineART was also able to present a decent improvement in comparison to BART, while in datasets that the performance of SoftBART was worse than BART (as in datasets Budget and Mumps), LineART presented a relatively good performance. Our take is that in datasets that BART strive to capture the underlying relationships that exist in the data, LineART is generally able to provide better results than BART, usually achieving a performance closer to what SoftBART would have. For all other datasets, the results are not expected to deviate much from the results expected from BART.

Cpu (Feldmesser, 1987) is a dataset on which CPU performance is estimated using characteristics of the hardware. Only one out of the seven independent variables is categorical. This is the kind of example that LineART is expected to show improvements over BART. Even with the possible interactions, there are not many categorical/dummy variables that could create specific partitions that require deeper trees. Also, the relationship between hardware and performance naturally have some level of linearity (for example, more memory or cache should lead to a better performance), but a smooth function would probably be more appropriate. Since BART has restrictions using step-functions, LineART and SoftBART both show large relative improvements over BART.

The Ais dataset (Cook and Weisberg, 1994) is populated with data from athletes of the Australian Institute of Sport. In our case, the response variable is body fat. The dataset contains only two categorical/dummy variables (sport category and sex) out of 12 independent variables. Like in the Cpu example, body fat is expected to be closely related to physiological characteristics of the athletes, meaning that models with linear and smooth functions should present a better performance when generalizing over a model that uses step-functions. Again, LineART and SoftBART presented considerable improvements regarding out-of-sample predictive performance in comparison to BART.

Diamond (Chu, 2001) is the dataset on which LineART presents its worst performance in comparison to BART. There are some possible reasons that might lead to this behavior. First, out of its four independent variables (carat, clarity, color and cut), only one is numeric (carat), so there is only one variable that may be impacted by the proposed prior to predict the response (scaled ln of price), since selecting one of the categorical variables (that are converted into dummy variables) would lead to the original BART prior. Second, the R^2 of a linear regression model for the Diamond dataset is the highest among all datasets (97.1%) without accounting for interaction terms (the model achieves an R^2 of 98.7% with interaction terms) meaning that the data is mostly linear and even with the proposed prior creating locally adaptive linear regressions, there is a limitation on the improvement that a more flexible model could achieve. And third, the dataset presents some repeated observations, which can lead the model to be overconfident in some partitions, leading to overfit. SoftBART also had issues with this dataset presenting a predictive out-of-sample performance inferior to BART.

Hatco is another dataset that LineART did not present a good general performance. This dataset has originally been introduced by Hair et al. (1998) as an example for teaching multivariate techniques. The response variable is the overall costumer satisfaction, and is defined to be between 0 and 10, and 8 out of the 14 independent variables in the dataset are numeric. The interactions among the variables can be more easily captured in a model with a larger number of trees or with deeper trees. The default of the LineART prior utilizes only 50 trees and has $\alpha = 0.5$, penalizing deep trees. Increasing α or increasing the number of trees lead to results closer to BART.

In general, SoftBART is a more complex model that is expected to outperform BART and LineART, while LineART is a middle-ground alternative, using a prior that can capture data patterns that are hard for BART to capture while maintaining almost the same speed provided by the prior introduced by Chipman et al. (2010).

3.5.2 Timing

One of the main positive features of LineART is the low number of operations required for likelihood calculation and posterior sampling. In general, the number of operations is similar to the BART prior, with the exception of a 2×2 matrix that has to be inverted for each leaf node. Figure 3.6 present the ratio of time spent by

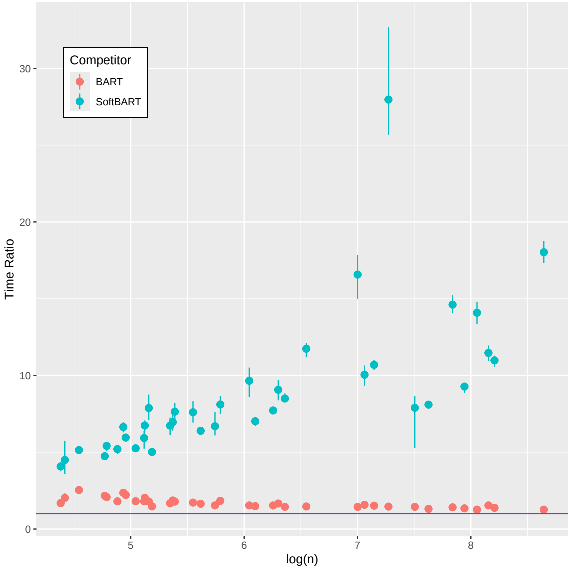


Figure 3.6: Time ratio of competitor versus LineART, both with default settings. 41 datasets with 20 simulations each. Solid purple line is equal to 1 and indicates when the time spent is the same as LineART. Vertical lines extend from the minumum to the maximum time observed among simulations for the same dataset.

the competitors (BART and SoftBART) versus the time spent by LineART for 20 replications of each dataset presented on Section 3.5.1.

Due to the lower number of trees in its default configuration, LineART MCMC chains were sampled at a faster rate than BART MCMC chains. However, as the sample size of the datasets increases, this difference starts to wane. SoftBART, on the other hand, presented ratios that increased substantially as the sample size increased. It must also be noted that the number of covariates for each dataset also played a role on how much time the MCMC chains using the SoftBART prior took to be sampled, but there is a clear positive trend in the calculated ratio.

Therefore, besides having a predictive power lower than SoftBART, LineART is presented as a middle-ground alternative by being more flexible than BART and more scalable than SoftBART.

3.6 Application

One of the most direct uses of the LineART prior is by applying it to bcf (Hahn et al., 2020), since there are basically no drastic structural changes in the model aside from the prior on the leaf nodes, which means that a practitioner can choose to use the LineART prior instead of the BART prior in bcf.

The original bcf is given by

$$\phi(\boldsymbol{x}_i, \boldsymbol{w}_i, z_i) = \mu\left(\boldsymbol{x}_i\right) + z_i \tau\left(\boldsymbol{w}_i\right), \tag{3.1}$$

where $\mu(.)$ and $\tau(.)$ are BART priors as defined by (Chipman et al., 2010), $\boldsymbol{x}_i$ is the subset of covariates related to control, $\boldsymbol{w}$ is the subset of covariates selected as moderators, and z is a binary treatment variable. The LineART prior introduces more flexibility by allowing locally adaptive linear regressions in the leaf nodes, and the practitioner might want to use the new prior in either $\mu(.)$, $\tau(.)$, or both. This possibility is readily available since LineART is a natural extension of the BART prior in the sense that most of the fundamentals of the original prior are kept the same.

For simplicity, a bcf with both BART priors substituted by LineART priors from now on is defined as lbcf, while a bcf with only one of the BART priors substituted is now defined as partial lbcf.

As a way to analyze any possible differences in results, a comparison between bcf and (partial) lbcf has been conducted in an study by (Yeager et al., 2022). In that study, adolescents from an university were separated in treatment and control groups. The treatment group was presented with the synergistic mindset intervention, a short (about 30-minute) online course that aimed to present the growth mindset (the idea that stressful situations are opportunities for the personal growth of a person) and the stress-can-be-enhancing mindset (the notion that the responses that one may notice in their body during stress situations, from anxiety to having a fast heart rate, can viewed as beneficial, since their body is basically preparing itself to have an enhanced performance, so instead of being worried about it, they can embrace these responses) in an integrated way.

The main variables in the study are the self-recorded prior mindset levels, named fixed mindset (the idea that intellectual ability cannot change, basically the opposite of the growth mindset) and stress mindset (or stress-is-debilitating mindset, which is the idea that stress is negative and harmful).

One of the ideas that the original study wanted to test was that people with negative prior mindsets (high levels of both fixed and stress mindsets) could benefit more, on average, of the synergistic mindsets intervention in comparison to people with positive prior mindsets (low levels of both fixed and stress mindsets). To separate these subgroups and analyze the average treatment effect on each subgroup, the same strategy employed by Yeager et al. (2022) has been applied, where all possible rectangular disjoint partitions using the values of stress and fixed mindsets are found, and the disjoint partitions (with the restriction of having at least 15% of the subjects are in each partition) with the highest difference in treatment effects (as opposed to Yeager et al. (2022) that uses the difference in prognostic effects) are selected.

In the presented dataset, each model was selected after a 5-fold cross validation, using lowest average RMSE from the test set ($RMSE_{test}$) across folds. The $RMSE_{test}$ is defined as

$$RMSE_{test} = \sqrt{\sum_{i=1}^{n_{test}} \frac{(y_i - \hat{y}_i)^2}{n_{test}}},$$

where $\hat{y}$, in this case, is the posterior mean of the prediction of y for the ith individual of the test set; y is the recorded response from the ith individual of the test set; and n_{test} is the number of elements in the test set.

The available parameters used in the cross-validation are available on Table 3.2. During this phase, each model sampled 1000 burn-in samples, and 1000 posterior samples were recorded, no thinning was used. The low number of samples is due to the high number of models that were being tested.

Hyperparameter	**Value**
Number of trees for control	{50, 100, 200}
Number of trees for treatment	{25, 50, 100}
α for control	{0.5, 0.8, 0.95}
α for treatment	{0.25, 0.5, 0.8}
β for control	{1, 2, 3}
β for treatment	{1, 2, 3}
Scale parameter for control	{yes, no}
Scale parameter for treatment	{yes, no}
Linear prior for leaves	{none, control, treatment, both}

Table 3.2: Hyperparameters available for the models during the 5-fold cross validation. In total, 11664 models were tested.

After the selection of the hyperparameters of the model, 10 chains with different seeds were created, each with 15000 samples of burn-in, and 5000 posterior samples recorded, with thinning of 3. The posterior samples of all 10 chains were then aggregated.

3.6.1 Anxiety

Study 6 from Yeager et al. (2022) tried to identify if the synergistic mindsets treatment impacted the perceived anxiety of adolescents during times of negative stress. In this case, the environmental stressor was the academic pressure alongside the social isolation during the early stages of the COVID-19 pandemic.

A 5-fold cross validation was performed in this dataset. The default settings of bcf presented a better result than the default settings of lbcf, and the best performance was achieved by a partial lbcf with a linear prior over the treatment effects. For simplicity, the partial lbcf-CV will be referred as lbcf-CV in this section.

The selected model for lbcf-CV uses $ntree_{con} = 50$, $ntree_{mod} = 25$, $\alpha_{con} = 0.5$, $\alpha_{mod} = 0.25$, $\beta_{con} = 3$, $\beta_{mod} = 2$, with a half-normal scale for $\mu(.)$, and a linear prior only for $\tau(.)$. On the other hand, bcf-CV uses $ntree_{con} = 50$, $ntree_{mod} = 50$, $\alpha_{con} = 0.5$, $\alpha_{mod} = 0.5$, $\beta_{con} = 2$, $\beta_{mod} = 3$, without half-normal scales.

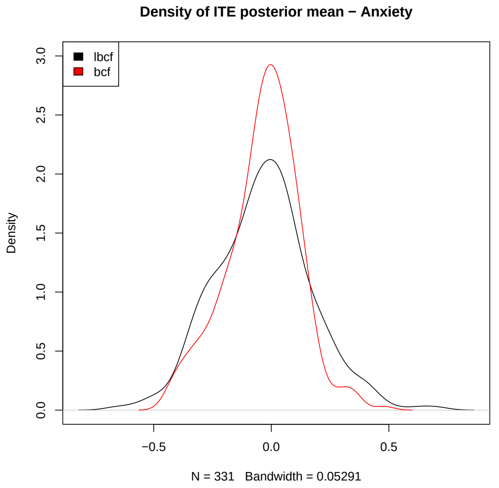


Figure 3.7: Density of the posterior mean of the estimated individual treatment effects on s.d. levels of Anxiety for bcf-CV and lbcf-CV.

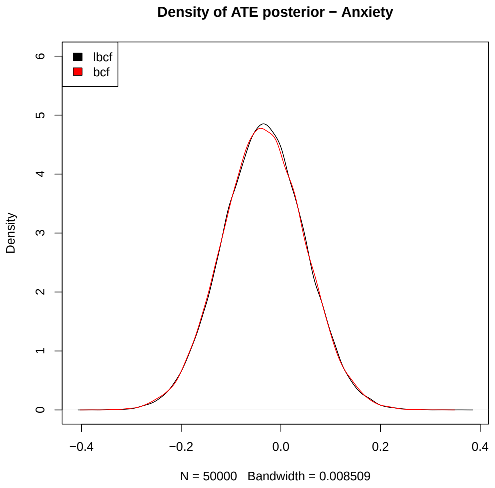


Figure 3.8: Density of 50000 posterior samples of the estimated average treatment effects on s.d. levels of Anxiety for bcf-CV and lbcf-CV.

For this study, the treatment effects have been divided by the standard deviation of the response variable, such that the treatment effects are now given in standard deviations of the levels of Anxiety. This change was made to allow a direct comparison with the original study from Yeager et al. (2022).

Figures 3.7 and 3.8 represent the posterior mean of the estimated individual treatment effects, and the posterior of the estimated average treatment effects. Figure 3.7 shows that the posterior mean of the ITEs is less concentrated around zero for lbcf-CV in comparison to bcf-CV. The posterior of ATE between the two models is basically the same, as seem on Figure 3.8.

GAMs showed upward trend in pss and downward trand at fixed mindsets. High values of stress mindsets also started to show a downward trend. High values of perceived social stress for internalizing symptoms is not great here, but that is also perceived stress, so people may be more overwhelmed than anything else.

The partial effects estimated by using GAMs to perform a lower-dimensional

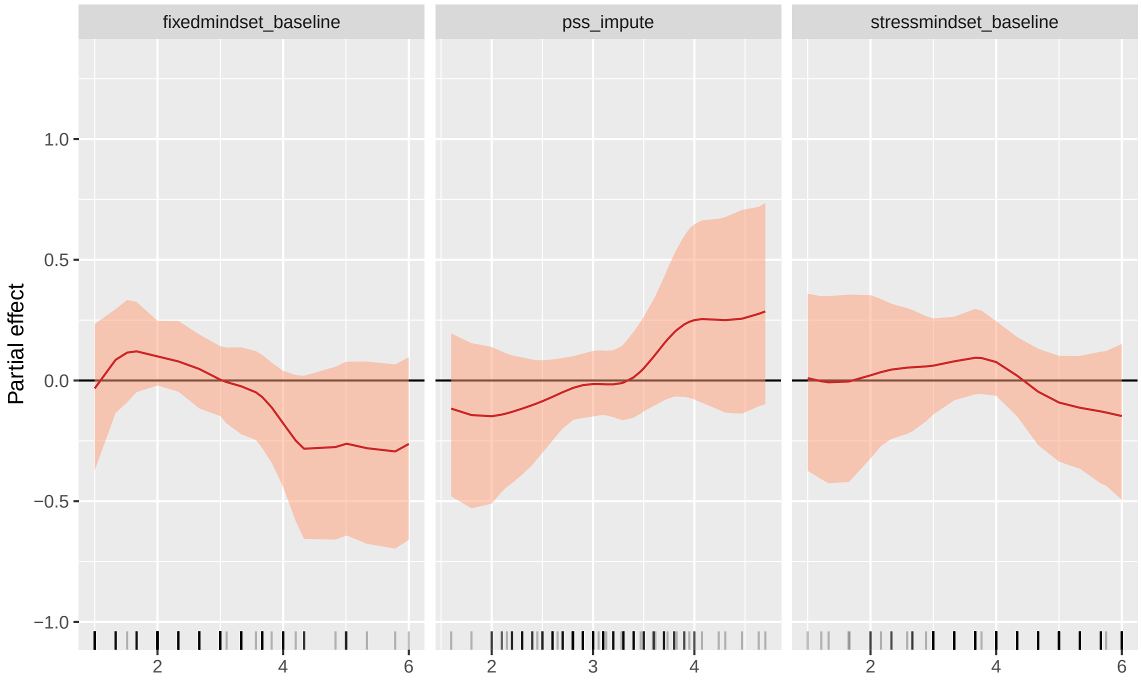


Figure 3.9: Partial effect of GAMs by separated by moderators for bcf-CV on treatment effects for s.d. levels of Anxiety.

summarization are shown on Figures 3.9 and 3.10 for bcf-CV and lbcf-CV, respectively. The general behavior of the the curves is maintained when comparing the partial effects of both models, however, the curves for perceived social stress (pss) and fixed mindsets are more steep in the summarization of the treatment effects from lbcf-CV.

The partial effect of fixed mindsets decreases as the level of fixed mindsets goes up, which is aligned with the idea that people with higher levels of this negative prior mindset would benefit more of synergistic mindsets treatment. The partial effect also shows a little bit of a tendency of decline for high values of the stress prior mindset. This provides a hint that people with negative prior mindsets would benefit more of the treatment than people with positive prior mindsets (low values for fixed/stress mindset). On the other hand, the partial effect for pss increases as the levels of pss increase, this could be due to fact that pss is a baseline measure

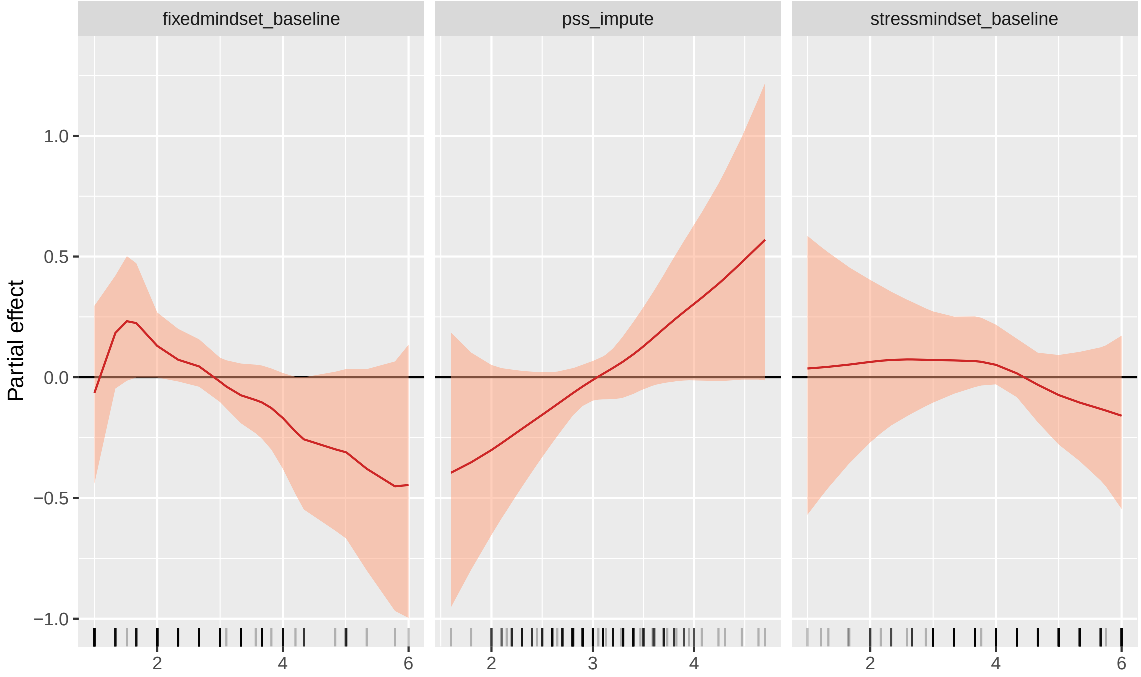


Figure 3.10: Partial effect of GAMs by separated by moderators for lbcf-CV on treatment effects for s.d. levels of Anxiety.

of internalizing symptoms, so the effect on people with high levels of internalizing symptoms is being mitigated, different from the effect of people with low levels of pss, that have a negative partial effect, and therefore, lower levels of anxiety.

Figure 3.11 presents the R^2 summary for both lbcf-CV and bcf-CV. Again, this is a measure of how much of variation of the estimated treatment effect by each model is being explained by the lower-dimensional summarization method, that in this case are the GAMs. The distribution of the R^2 summary for lbcf-CV is more concentrated around 1 than bcf-CV, but both GAMs seem to reasonably estimating the variation of the posterior of the estimated treatment effects.

As for subgroups, there are a few subgroups of interest that can be explored. First, grouping the participants by sex and verifying if there is a meaningful difference between the treatment effects of male and female participants, and second, aggregat-

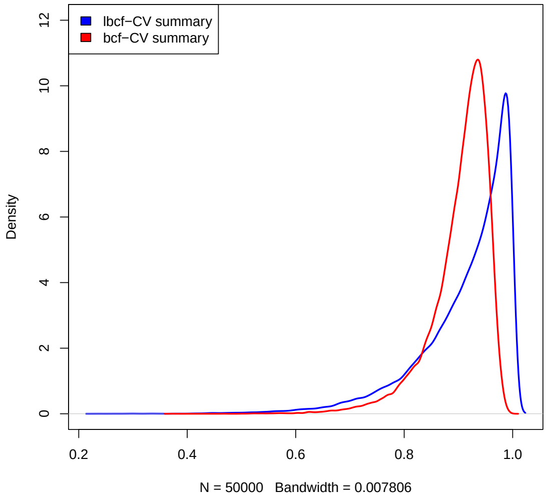


Figure 3.11: Summary R^2 for the GAMs of bcf-CV and lbcf-CV on levels of Anxiety.

ing people by prior mindsets, especifically negative and positive prior mindsets, since, as it can be seem on Figures 3.9 and 3.10, high levels of prior mindsets are linked with negative levels of partial effects.

By analyzing Figures 3.9 and 3.10, 3.17 is a natural point to separate fixed mindsets in two partition (also verified by using a CART model), and by calculating the subgroup difference for both bcf-CV and lbcf-CV, the probability that the difference is bigger than zero is 95.9% for bcf-CV, and 96.2% for lbcf-CV, which by itself is already a meaningful difference.

Figure 3.12 illustrate the difference between the treatment effects divided by sex. In general, the densities of the difference of subgroups for both bcf-CV and lbcf-CV appears to be centered around zero, with bcf-CV presenting a 64.8% probability of a difference larger than zero, against a 61.1% probability of a difference larger than zero from lbcf-CV. This means that there might some indication that sex plays a role on intensity of the treatment effects, but even with this difference being higher than by chance (50.0%), this effect is not as meaningful as the sole effect of the difference of high and low values of fixed mindsets.

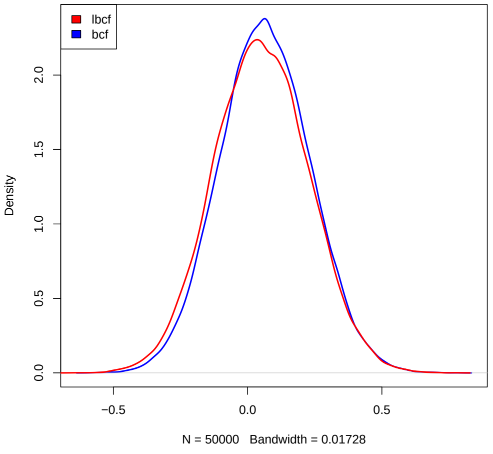


Figure 3.12: Density of difference of the estimated average treatment effects of sex subgroups on levels of Anxiety for bcf-CV and lbcf-CV.

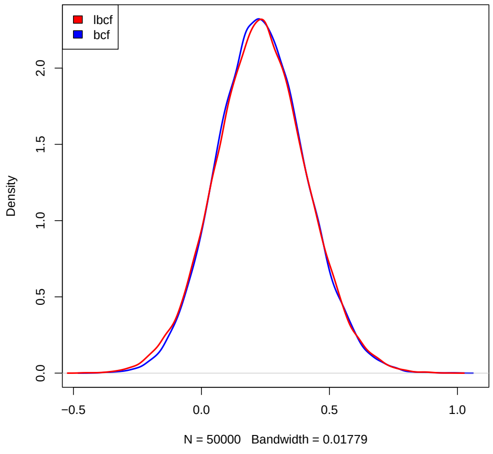


Figure 3.13: Density of difference of the estimated ATEs between negative and positive prior mindsets subgroups on levels of Anxiety for bcf-CV and lbcf-CV.

Figure 3.13 shows the difference between subgroups with positive and negative prior mindsets, with the negative prior mindsets being set as values higher than 3.17 (3.83) for fixed (stress) mindset, while positive prior mindsets are set as values lower than 2.50 (3.83) for fixed (stress) mindset. The density for both lbcf-CV and bcf-CV are mostly located over zero, with the probability that the difference is higher than zero being estimated as 91.8% and 90.7% for bcf-CV and lbcf-CV, respectively. Therefore, both by looking at the difference of high/low values of fixed mindsets by itself or at the difference of negative/positive prior mindsets, the treatment effect seems to be more effective on the subgroups that with higher levels of prior mindsets.

3.7 Discussion

The novel LineART prior is a clear alternative to the original BART prior due to the increase flexibility proportioned by the use of a locally adaptive simple linear regression instead of a scalar on the leaf nodes. The default set of hyperparameters chosen for LineART seemed to accommodate most scenarios tested on Section 3.5, presenting results closely related to hyperparameters selected via cross validation, making LineART an easily accessible and ready to use prior.

In comparison to SoftBART, which is a more complex alternative, LineART seemed to be able to reasonably capture the underlying relationships in most of the datasets that SoftBART presented predictive capabilities superior to BART, and since in some cases the time spent to run a model using the SoftBART prior could lead to times up to 30 times the time spent by models with the LineART prior (depending on the sample size and number of covariates). Our novel prior presents a trade-off between scalability and predictive performance, since in some cases the use of the SoftBART prior becomes unfeasible, and another flexible approach is demanded.

The application of the prior to causal inference settings makes it a natural extension of bcf, introducing lbcf and adding a simple and new tool for practitioners. Based on the data analysis of Section 3.6, the results were mostly in line with bcf, but when estimating partial effects with GAMs, the estimated credible intervals were generally tighter, and the produced curves presented sections that more closely resemble linear regressions, leading to higher values of summary R^2.

Therefore, LineART is a practical substitute of BART in a wide variety of settings, and it is expected that the novel prior will outperform BART in most scenarios. Also, due to its similarities with BART, some extensions are readily available for implementation, like the inclusion of a variable selection Dirichlet prior (Linero, 2018).

Chapter 4: Scalable Bayesian causal forest for continuous treatments

Abstract

The Bayesian additive regression trees (BART) prior and its extensions are considered reliable options for prediction in a wide variety of settings. One of its many use cases is at the causal inference field, however limitations regarding either the treatment or the association between the response and the levels of the treatment have to be made due to the lack of scalability in those models. We introduce a scalable Bayesian causal forest prior for continuous treatments by approximating the covariance kernel of a Gaussian Process prior.

4.1 Introduction

The estimation of causal effects is a widely studied problem in Statistics, and one of the most used frameworks is the Neyman-Rubin causal model, which apparently was first proposed by Neyman (1923) (as mentioned by Rubin (2005)) and later extended by Rubin (1974) in the potential outcomes framework.

In general, when dealing with the estimation of causal effects, one does not have the ability to observe what would have happened to the response of a certain observation since, by definition, one can only observe the response given the received treatment, which means that this is a missing data problem (Ding and Li, 2018). The response if the assigned treatment was different than the observed treatment is called counterfactual, and it is one of the main studied elements in causal inference settings.

The Neyman-Rubin causal model focus mostly on the case of dealing with binary treatments. The broader case of continuous treatments, has some specific challenges. In the case of Bayesian tree-based models the practitioner might want to assume that given a set of covariates, the treatment effect is a smooth function of the continuous treatment Z. This case is basically a Bayesian causal forest (bcf) (Hahn et al., 2020) where the treatment effect term $\psi(.)$ has a tsBART prior (Starling et al., 2019). As mentioned in Chapter 2, the issue with this kind of approach is that with Z as the targeted smoothing variable, the computational time would grow rapidly and using such a model would be infeasible. However, with the methods developed on Chapter 2, this model is now feasible. With the Bayesian causal forest prior for continuous treatments (bcf+) model, it is possible to revisit the analysis of Imai and Van Dyk (2004) of the NMES smoking data and analyze the impact of smoking on the expected annual medical expenditure.

4.2 Methods

Using Section 2.3 as a reference, let $\boldsymbol{x}_i \in \mathcal{X}$, $\boldsymbol{w}_i \in \mathcal{X}$ be vectors of covariates associated with the individual $i \in \{1, ..., N\}$, and let $z \in \mathcal{Z}$ be the continuous treatment variable.

Our identifying assumptions are the same as those observed in Imai and Van Dyk (2004):

(i) *Stable Unit Treatment Value Assumption*:
The distribution of potential outcomes for one unit is assumed to be independent of potential treatment status of another unit given the observed covariates;

(ii) *Strong Ignorability of Treatment Assignment*:
$Y(z) \perp Z|\boldsymbol{X} = \boldsymbol{x} \quad \forall \;\; z \in \mathcal{Z},\; \boldsymbol{x} \in \mathcal{X}$;

(iii) *Positivity*:
$p(z|\boldsymbol{X} = \boldsymbol{x}) > 0 \quad \forall \;\; z \in \mathcal{Z},\; \boldsymbol{x} \in \mathcal{X}$.

SUTVA establishes that other observations receiving a treatment do not affect the outcome of an observation (no interference), and that there are no alternate versions of a given treatment. Strong ignorability assumes that given the covariates, the treatment is independent of all potential outcomes. Positivity guarantees that given a set of observed covariates, any treatment has a positive density, such that there is a non-zero probability of sampling a treatment in any interval around z.

For any two treatments z and z', given our assumptions, it is possible to compare any two levels potential outcomes, such that an estimand that can now be computed is the conditional average treatment effect between two treatments, namely

$$CATE_{(z,z')}\left(\boldsymbol{x}\right) = \mathbb{E}\left(Y(z) - Y(z')|\boldsymbol{x}\right), \tag{4.1}$$

mostly used to calculate the average treatment effects given a set of covariates, which can be used for subgroup comparison.

It is also possible to estimate the average treatment effect between two treatments z and z', defined as

$$ATE_{(z,z')} = \mathbb{E}\left(Y(z) - Y(z')\right), \tag{4.2}$$

being an estimand that is used to estimate the overall treatment effect between any two treatments.

Finally, the conditional dose-response function is defined as

$$\xi_z(\boldsymbol{x}) = \mathbb{E}\left(Y(z)|\boldsymbol{x}\right), \tag{4.3}$$

and can be used to represent the expected response differences for a given set of covariates while varying the treatments.

4.2.1 Model statement

The problem is of the form

$$y_i = \varphi(\boldsymbol{x}_i) + \psi(\boldsymbol{w}_i, z_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{4.4}$$

Like bcf, the Bayesian causal forests for continuous treatments (bcf+) is separated in two terms. The term φ is a BART prior (as defined by Chipman et al. (2010)) associated with the variables that estimate the contribution to the response regardless of a given treatment. ψ is a scalable tsBART prior as defined in Section 2.4.1, where the covariance kernel of the Gaussian Processes is approximated using basis functions (Solin and Särkkä, 2020).

4.2.2 Regularization Induced Confounding

Hahn et al. (2018) notes that even in scenarios where strong ignorability and positivity hold, the regularization that is present in Bayesian models can introduce bias in the estimated treatment effects of linear models, due to regularization induced confounding (RIC).

Hahn et al. (2020) notes that this bias can be diminished if no correlation is observed between Z and $\boldsymbol{X}$, and also that despite having no closed-form in nonparametric regression models with heterogeneous treatments, the RIC phenomenon can be recreated in nonlinear settings.

One way of looking at the RIC phenomenon is that when the treatment Z and the covariates $\boldsymbol{X}$ are correlated, regularization might attribute changes that are related to treatment effects to the prognostic effect instead. Using the example of targeted selection, which are settings that, given a set of covariates, the treatment assignment is partially based on the prognostic effect, Hahn et al. (2020) argues that the use of the propensity score as a covariate helps to mitigate the RIC phenomenon.

Let us assume that the true propensity score is known and given by $e(\boldsymbol{x}) = \mathbb{P}(Z = 1|\boldsymbol{X} = \boldsymbol{x})$, and that, for simplicity, the same variables are used as controls and moderators ($\boldsymbol{x} = \boldsymbol{w}$). In the original bcf framework, with binary treatments, it is possible to use the residual treatment $\tilde{z}_i = z_i - e(\boldsymbol{x}_i)$ instead of z, such that

$$y_i = \varphi(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)(z_i - e(\boldsymbol{x}_i)) + \epsilon_i, \tag{4.5}$$

which makes $(Z_i - e(\boldsymbol{x}_i))$ and $\boldsymbol{X}$ orthogonal, and therefore, diminishes the effect of the RIC phenomenon. However, the estimation of treatment effects using z provides a level of interpretability that could be lost by the use of $\tilde{z}$.

It is possible to rewrite the bcf model such that

$$y_i = \varphi(\boldsymbol{x}_i) - \psi(\boldsymbol{x}_i)e(\boldsymbol{x}_i) + \psi(\boldsymbol{x}_i)z_i + \epsilon_i,$$

and assuming that adding $e(\boldsymbol{x}_i)$ among the covariates of $\varphi(.)$ would lead to

$$\varphi(\boldsymbol{x}_i, e(\boldsymbol{x}_i)) \approx \varphi(\boldsymbol{x}_i) - \psi(\boldsymbol{x}_i)e(\boldsymbol{x}_i), \tag{4.6}$$

due to the high flexibility of the BART prior, then having the propensity score as a covariate leads to a model that can be considered equivalent to the model represented in Equation 4.5.

Therefore, the use of the propensity score as a covariate as expressed by (Hahn et al., 2020) works in the case of bcf and in the case of Woody et al. (2020), where z is continuous but assumes that the effect of the exposure on the outcome is linear.

In our approach, for a specific leaf node,

$$\mu(z) \approx \sum_{q=1}^{m} \beta_q \omega_q(z),$$

as in Equation 2.11. Equation 2.10 shows that z is used as the input of a sin function, such that using $\tilde{z}$ does not lead to a result similar to Equation 4.6. However, due to the approximation, using

$$\mu'(z) \approx \sum_{q=1}^{m} \beta_q \left(\omega_q(z) - \mathbb{E}\left(\omega_q(z)|\boldsymbol{x}\right)\right),$$

provides a similar effect. Now, since

$$\mathbb{E}\left(\omega_q(z)|\boldsymbol{x}\right) = c_{1q}\mathbb{E}\left(\sin\left(\frac{\pi q}{2c}z + \frac{\pi q c}{2c}\right)|\boldsymbol{x}\right) = c_{1q}\mathbb{E}\left(\sin\left(c_{2q}z + c_{3q}\right)|\boldsymbol{x}\right),$$

analyzing this term is the same as analyzing

$$\mathbb{E}\left(\sin(az + b)|\boldsymbol{x}\right),$$

where a and b are constants.

Using the first terms of the Taylor expansion about the point z' leads to

$$\sin(az+b) = \sin(az'+b) + a\cos(az'+b)(z-z') + ...,$$

and applying the conditional expectation,

$$\mathbb{E}\left(\sin(az+b)\,|\boldsymbol{x}\right) = \sin(az'+b) + a\cos(az'+b)(\mathbb{E}\left(z|\boldsymbol{x}\right) - z') + ...,$$

we can define $z' = \mathbb{E}(z|\boldsymbol{x})$,

$$\mathbb{E}\left(\sin(az+b)\,|\boldsymbol{x}\right) = \sin(a\mathbb{E}(z|\boldsymbol{x})+b) + a\cos(a\mathbb{E}(z|\boldsymbol{x})+b)(\mathbb{E}\left(z|\boldsymbol{x}\right) - \mathbb{E}(z|\boldsymbol{x})) + ...,$$

which leads to the approximation

$$\mathbb{E}\left(\sin(az+b)\,|\boldsymbol{x}\right) \approx \sin(a\mathbb{E}(z|\boldsymbol{x})+b).$$

Therefore,

$$\mathbb{E}\left(\omega_q(z)|\boldsymbol{x}\right) = c_{1q}\mathbb{E}\left(\sin\left(c_{2q}z + c_{3q}\right)|\boldsymbol{x}\right) \approx c_{1q}\sin\left(c_{2q}\mathbb{E}\left(z|\boldsymbol{x}\right) + c_{3q}\right).$$

Now, expanding the term

$$\begin{aligned}
\mu'(z) &\approx \sum_{q=1}^{m}\beta_q\omega_q(z) - \sum_{q=1}^{m}\beta_q\mathbb{E}\left(\omega_q(z)|\boldsymbol{x}\right) \\
&\approx \sum_{q=1}^{m}\beta_q\omega_q(z) - \sum_{q=1}^{m}\beta_q c_{1q}\sin\left(c_{2q}\mathbb{E}\left(z|\boldsymbol{x}\right) + c_{3q}\right) \\
&\approx \mu(z) + \mu\left(\mathbb{E}(z|\boldsymbol{x})\right),
\end{aligned}$$

it is possible to assume that the function $\mu\left(\mathbb{E}(z|\boldsymbol{x})\right)$ can be approximated by a BART prior, such that the inclusion of $\mathbb{E}(z|\boldsymbol{x})$ among the covariates of $\varphi(.)$ should lead to the mitigation of the effects of the RIC phenomenon.

Lastly, it must be noted that the function targeted smoothing function in terms of z is, by definition, smooth in each node. Subtracting $E(z|\boldsymbol{x})$ directly from z could lead to inconsistent curves, motivating the approach discussed in this section.

4.2.3 Priors

The priors follow the same guidelines as discussed in Hahn et al. (2020) and Chapter 2. $\varphi(.)$ is a BART prior with default parameters and a half-normal prior on a scale parameter.

$\psi(.)$ is a scalable tsBART prior as discussed in Chapter 2, with the default parameters suggested by Hahn et al. (2020) for bcf. The length-scale parameter uses a half-normal prior, as suggested in Chapter 2, and the default expected number of crossings is set to 3.

4.2.4 Related work

Hill (2011) applied BART to a causal inference setting with binary treatments by adding the treatment variable Z among the covariates of the model.

Hahn et al. (2020) introduced the bcf model by separating the prognostic effect and the treatment effect into two BART priors, and recognized that including the propensity score among the covariates of the model reduced the bias provenient from the RIC phenomenon on targeted selection settings.

Still on the binary treatment case, Starling et al. (2020) expanded bcf to settings with a targeted smoothing variable by using a Gaussian Process prior on leaf nodes. This idea motivated the developments in Chapter 2, where the reduced-rank Gaussian Process approximation developed by Solin and Särkkä (2020) was applied to tsBART and tsbcf.

On the continuous treatment case, Woody et al. (2020) developed a tsbcf with continuous treatments with the assumption that the treatment effect is linear, conditional on the value of moderator covariates. Hirano and Imbens (2004), on the other hand, discusses the use of the generalized propensity score on settings with continuous treatments.

4.3 Simulation study

This is an experimental toy example with the objective of estimating the ATE for different pairs of treatment values.

Consider

$$X_1 \sim \mathcal{U}(0,1),$$

$$X_2 \sim \mathcal{N}(0,1),$$

$$Z \sim \mathcal{U}(-1,1).$$

Let the response be defined such that

$$y_i = 2x_{2i} + (1 + x_{1i})z_i + \sin\left(\pi z_i(1 + 2x_{1i})\right) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,1).$$

In this case, we have that our treatment variable is continuous in the interval $[-1,1]$ and our treatment effect is dependent on X_1. The treatment has a linear relationship with the covariate X_1, since both values are multiplied, and a non-linear relationship due to the sine function.

For this simulation two models are compared, the first one is original BART (Chipman et al., 2010) with default parameters, and the treatment variable Z included among the covariates of the model. The second model is the novel bcf+ with default parameters. It must be noted that on bcf+, X_1 was added as a moderator and X_2 as a control, such that the model is correctly specified. For each model, the first 50000 MCMC samples are treated as burn-in, and 1000 posterior samples are recorded, with a thinning of 5.

Figure 4.1 presents the densities of posterior samples of the estimated $ATE_{(z,z')}$ for $(z,z') \in \{(0,0.5),(-0.5,0),(-0.5,0.5)\}$. For both models, the true value was within the range of sampled values, however, BART presented higher levels of uncertainty, even including zero within the 95% credible interval for $ATE_{(0,0.5)}$ in the left graph. The mode of the densities estimated by using bcf+ are also closer to the true

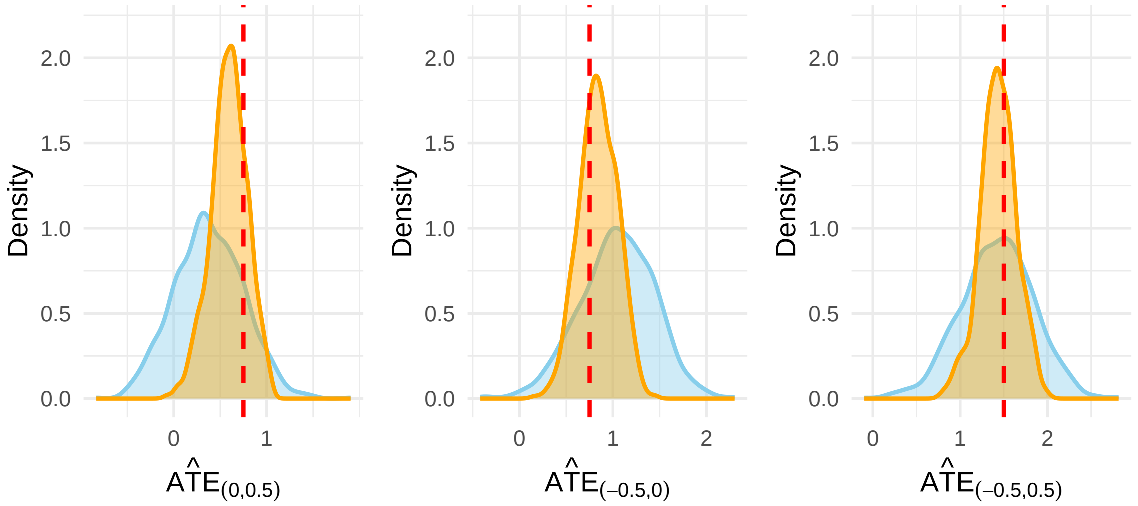


Figure 4.1: $ATE_{(z,z')}$ estimation for different treatments z and z'. In orange, the density of the $ATE_{(z,z')}$ as estimated by bcf+. In blue the density of the $ATE_{(z,z')}$ as estimated by BART. The red dashed line is the true sample average treatment effect (0.75 for the left and middle graphs, 1.5 for the graph on the right).

value of $ATE_{(z,z')}$ in all scenarios in comparison to BART. In general, bcf+ seems superior, which is expected due to its higher levels of flexibility.

Figure 4.2 presents the $CATE_{(z,z')}(\boldsymbol{x})$ curves for different values of X_1. For these curves, the value of X_2 was kept constant at 0. The BART estimates present curves that kept the true function within their 95% credible intervals for all cases, however, in all cases the posterior mean drifted away from the true function. Nevertheless, on the bcf+ case the posterior mean followed the general trend of the true functions, with some degree of variation. Therefore, bcf+ captured moderation more accurately in comparison to BART.

By fixing X_1 to a set of predetermined values ($X_1 \in \{0.2, 0.5, 0.8\}$) and fixing $X_2 = 0$ to remove the prognostic effect from the functions, it is possible to analyse the conditional dose-response curves, as shown on Figure 4.3. BART, as expected, due to the use of scalars in the leaf nodes of its trees, has curves that lack smoothness since those curves are created using step-functions. On the other hand, bcf+ pre-

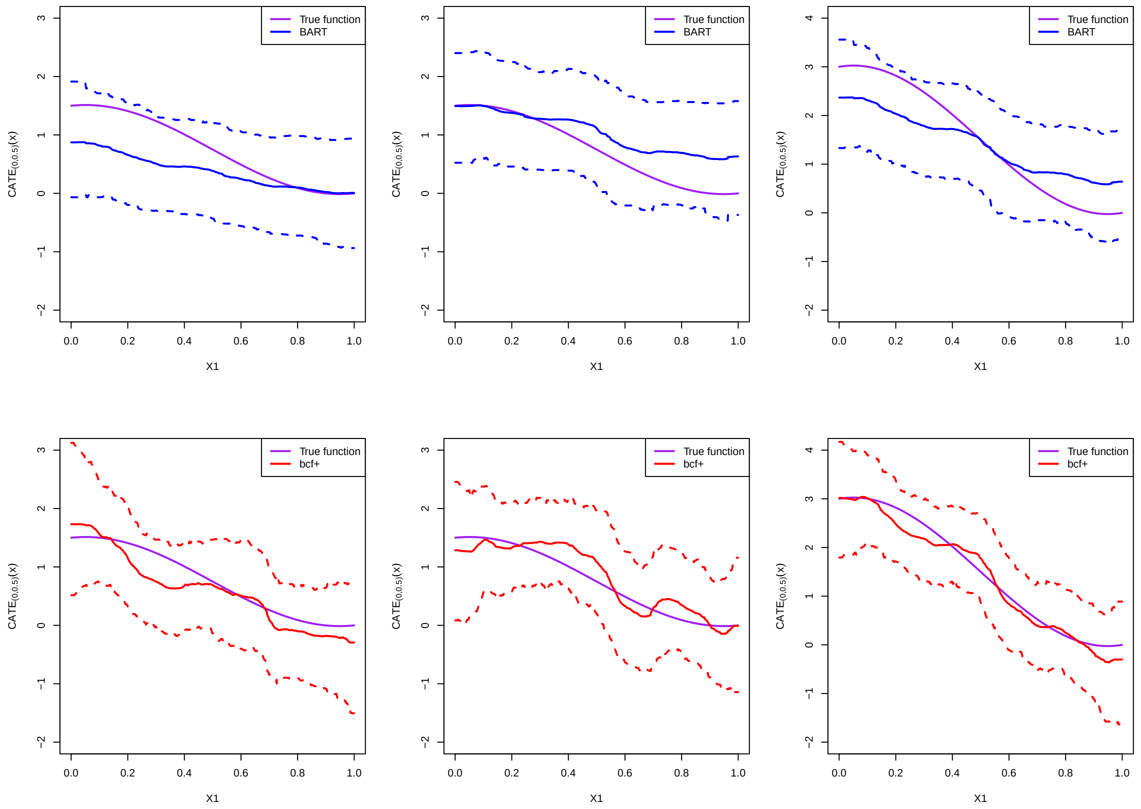


Figure 4.2: $CATE_{(z,z')}(\boldsymbol{x})$ curves over X_1. The control covariate X_2 was set to 0. The purple solid lines represent the true curves. On the top (bottom) graphs, the blue (red) solid lines represent BART (bcf+) posterior means, and the dashed lines represent their respective 95% credible intervals. On the left, middle and right, we have $(z, z') \in \{(0, 0.5), (-0.5, 0), (-0.5, 0.5)\}$, respectively.

sented smooth curves, which are created using the reduced-rank approximation for the Gaussian Process priors on the leaf nodes. Both BART and bcf+ capture the general trend of the true functions, but in the case of smooth functions, the adaptability of bcf+ is clearly superior.

In general, BART can reasonably estimate the functions on Figure 4.3, while failing at properly estimating the functions on Figure 4.2. The possible reason is that since BART uses both Z and X_1 as possible variables for splitting, both the moderator and the treatment are being treated similarly. A BART model is not able to properly

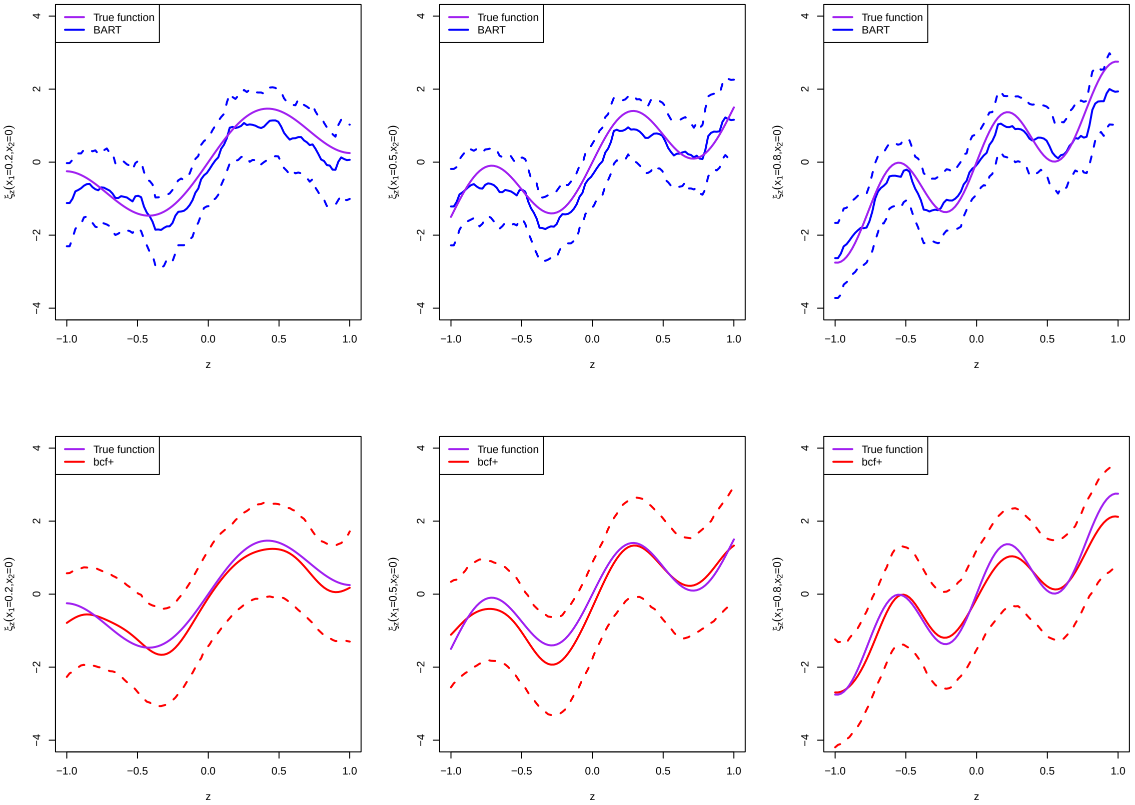


Figure 4.3: $\xi_z(\boldsymbol{x})$ curves over z. The control covariate X_2 was set to 0. The purple solid lines represent the true curves. On the top (bottom) graphs, the blue (red) solid lines represent BART (bcf+) posterior means, and the dashed lines represent their respective 95% credible intervals. On the left, middle and right, we have $X_1 \in \{0.2, 0.5, 0.8\}$, respectively.

estimate smooth functions, since these functions demand a higher number of splits. There might be a trade-off where while estimating the functions on Figure 4.2, the tree splits that would be required on the moderator X_1 are not available, leading to the poor result shown on Figure 4.3. It might be possible to partially address this issue with cross-validation, increasing the split probability and the number of trees in the model, but the issue of BART requiring too many tree resources to estimate a smooth curve remains. The bcf+ solves this issue by estimating functions that are smooth by definition, and therefore, the tree splits are available for the estimation of heterogeneity among the moderators.

4.4 Application

The National Medical Expenditure Survey (NMES) data was a survey designed to collect data about individuals and families regarding their health expenditure. In our case, the 1987 survey is used.

Using the data from this survey Imai and Van Dyk (2004) defined a variable to account for smoking exposure. This variable is a combination of the self-reported frequency and duration of smoking, called *packyears*, where

$$packyears = \frac{\text{\# of cigarettes per day}}{20} \times \text{\# of years smoked.}$$

Variable	**Description**	**Role**
log(`packyears`)	Log of lifetime smoking exposure	Treatment
log(`TOTALEXP`)	Log of total annual medical expenditure	Response
`AGESMOKE`	Age when the individual started smoking	Covariate
`LASTAGE`	Age at time of the survey	Covariate
`MALE`	Gender	Covariate
`RACE3`	Race	Covariate
`beltuse`	Frequency of seatbelt usage	Covariate
`educate`	Education level	Covariate
`marital`	Marital status	Covariate
`SREGION`	Region of residence	Covariate
`POVSTALB`	Poverty status	Covariate
`HSQACCWT`	Household sampling weight	Covariate

Table 4.1: Selected variables of NMES dataset

The variables selected as covariates in this analysis are the same selected by Imai and Van Dyk (2004), and a brief description of each variable is available on Table 4.1. All variables on Table 4.1 have been added as controls and moderators. The natural logarithm of the variables *TOTALEXP* and *packyears* have been used such the distribution of these variables is more well-behaved, resembling a bell-shaped curve. The $\log(packyear)$ (scaled to be between $[-1, 1]$) is the treatment variable in this section.

Furthermore, our sample has been restricted to smokers with a positive annual medical expenditure, as in Imai and Van Dyk (2004) and Hahn et al. (2020). Observations with *packyears* larger than 70 have also been removed from the sample as a way to mitigate some of the overlap issues that this dataset presents, because extremely larger values of *packyears* will not be observed for young people since they did not live long enough to have these levels of smoking exposure. Excluding young people from the sample is also an option, but since that may affect the estimated treatment effect for people with lower levels of smoking exposure, these individuals were kept in the sample. Our final sample size is $n = 7796$.

The $\mathbb{E}(Z|\boldsymbol{X} = \boldsymbol{x}_i)$ has also been added as a control since, as explained on Section 4.2.2, it can reduce the bias introduced by regularization induced confounding. $\mathbb{E}(Z|\boldsymbol{X} = \boldsymbol{x}_i)$ can be estimated using any chosen method, in this example the LineART from Chapter 3 with default parameters has been used, and all covariates described on Table 4.1 have been used as covariates. The first 50000 MCMC samples were treated as burn-in, and with a thinning of 5, 5000 samples were recorded. For the ith observation, $\mathbb{E}(Z|\boldsymbol{X} = \boldsymbol{x}_i)$ was estimated as the posterior mean for the given set of covariates $\boldsymbol{x}_i$.

Figure 4.4 represents the posterior sample of the length-scale related to the treatment effect term of bcf+. As in the original framework developed on Chapter 2, the possible values for the treatment variable have been restricted between $[-1, 1]$, which means that the sampled values of the parameter were focused on larger length-scale values (the mode, for example, is around 0.45, which is around 22.5% of the range of treatments), indicating that the treatment curves estimated by the Gaussian Processes on the leaf nodes are not necessarily wiggly.

In their analysis Imai and Van Dyk (2004) conclude that doubling *packyears* leads to an expected multiplicative factor of 1.04 on the expected annual medical

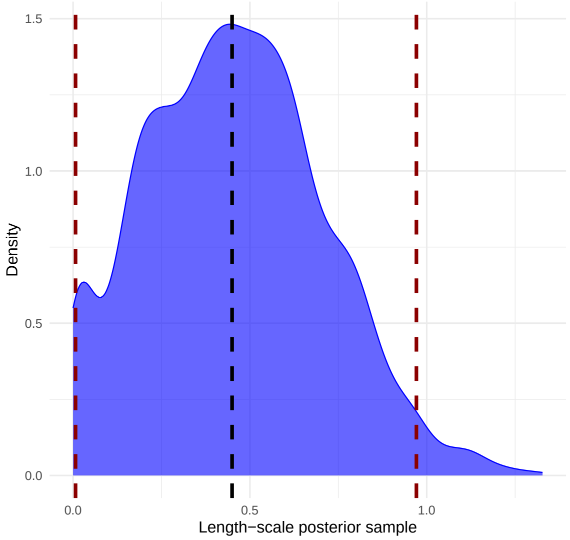


Figure 4.4: Posterior samples of treatment effect length-scale. The black dashed line represents the posterior mean, while the red dashed lines represent the 2.5% and 97.5% quantiles.

expenditure. Our approach calculates the conditional average treatment effect of doubling *packyears* for a number of different values, as seen in Figure 4.5. It makes sense that for larger values of *packyears* the observed multiplicative factor is larger, since the more an individual smokes, there is a common expectation that a higher number of health related issues will arise, which, ultimately, could lead to higher expected annual medical expenses.

For the largest gap in treatment (32 vs 64 *packyears*) the estimated conditional average treatment effect presented a higher level of uncertainty on the right tail of the density than the second largest gap in treatment (16 vs 32 *packyears*). This behavior could be a signal of a higher level of uncertainty since the number of observed subjects decreases as *packyears* increases, and this lack of data can be translated to higher levels of uncertainty, or that some people in the group of 32 vs 64 *packyears* had more health related issues, extending the right tail of the density to higher values.

When analyzing the estimated $ATE_{(z,z')}$ between low levels of treatments,

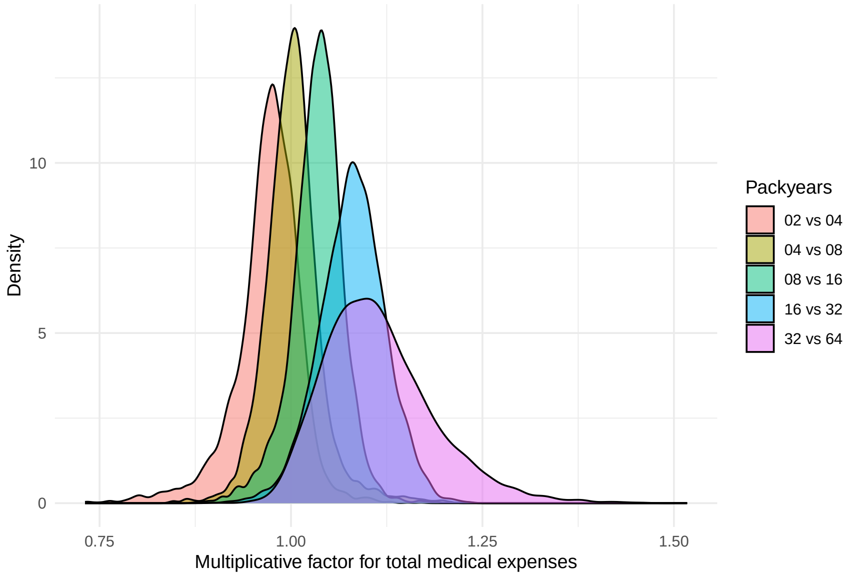


Figure 4.5: Comparison of the exponential of estimated $ATE_{(z,z')}$ between different treatments values ($packyears$). The exponential was utilized for interpretability. The y-axis represents the density of the exponential of estimated $ATE_{(z,z')}$, while the x-axis represents the multiplicative factor on expected annual medical expenditure between two different values of $packyears$.

such as 2 and 4, or 4 and 8, the lifetime exposure to smoking has likely been too low to cause a relevant change in the expected annual medical expenditure, which would explain the densities being around the multiplicative factor of 1, representing basically no change in the response.

Figure 4.6 is a representation of the impact on annual medical expenditures of a heavy smoker (64 $packyears$) in comparison to a light smoker (8 $packyears$). The distribution of the exponential of estimated $ATE_{(8,64)}$ indicates that there is an increase in annual medical expenditures with an average of around 23.8%.

Still comparing light (8 $packyears$) and heavy (64 $packyears$) smokers, the exponential of estimated treatment effects seems to be decreasing as age increases, as seem in Figure 4.7. This kind of behavior is probably due to survivor bias (as noted by Imai and Van Dyk (2004) and Hahn et al. (2020) in their analysis), since older

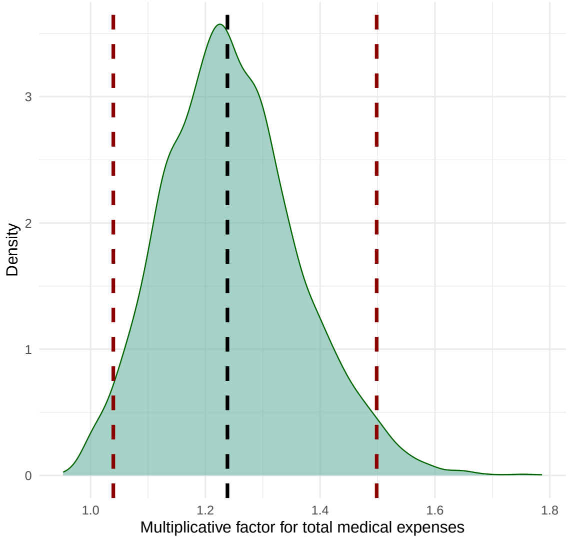


Figure 4.6: Exponential of estimate $ATE_{(8,64)}$. The black dashed line represents the mean of the sample, while the red dashed lines represent the 2.5% and 97.5% percentiles.

people with more health issues (which are the people with higher levels of medical expenditure) are more prone to pass away. Another possible explanation is that since older people have a higher baseline of medical expenditures in comparison to younger people, the multiplicative factor of being a heavy smoker is not as high as the multiplicative factor estimated for younger people.

Since all covariates were included in the bcf+ model, it is possible to perform subgroup analysis for heterogeneous treatment effects. Figure 4.8 represents the CART tree created by using the exponential of estimated individual treatment effects as the response variable, along with all covariates used in the bcf+ model. The depth of the tree was fixed to two, keeping the subgroups more interpretable. The general order of subgroups is similar to what was analyzed by Hahn et al. (2020), with Age and Gender playing an important role, where young males apparently had a larger treatment effect in comparison to older women.

By analyzing the difference between those subgroups, as seen in Figure 4.9, the

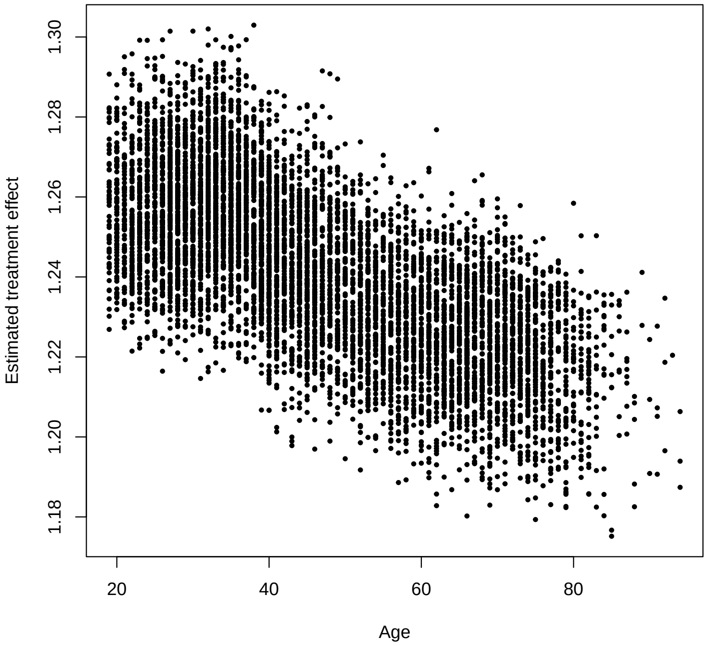


Figure 4.7: Posterior mean of the exponential of estimated individual treatment effects by age. Light smokers (8 *packyears*) vs heavy smokers (64 *packyears*).

mode is around zero, but the probability that the difference is larger than 0 is 82.8%. It must be noted that Hahn et al. (2020) estimated the treatment effect of older women to be practically zero when comparing light and heavy smokers in a binary treatment setting, while our analysis show a decrease in the multiplicative factor from around 1.27 for young men to around 1.22 for older women. This is a considerable change in the estimated subgroups of conditional average treatment effects in comparison to the the analysis of Hahn et al. (2020), where the treatment effect for the subgroup of older women was close to zero in the comparison between light and heavy smokers. It must be noted, however, that in their analysis, the authors transformed *packyears* in a binary treatment, with individuals that had *packyears* over 17 being assigned a treatment equal to 1. Their way of dealing with overlap also differed, since individuals younger than 28 were excluded from the sample, while keeping all values of *packyears*.

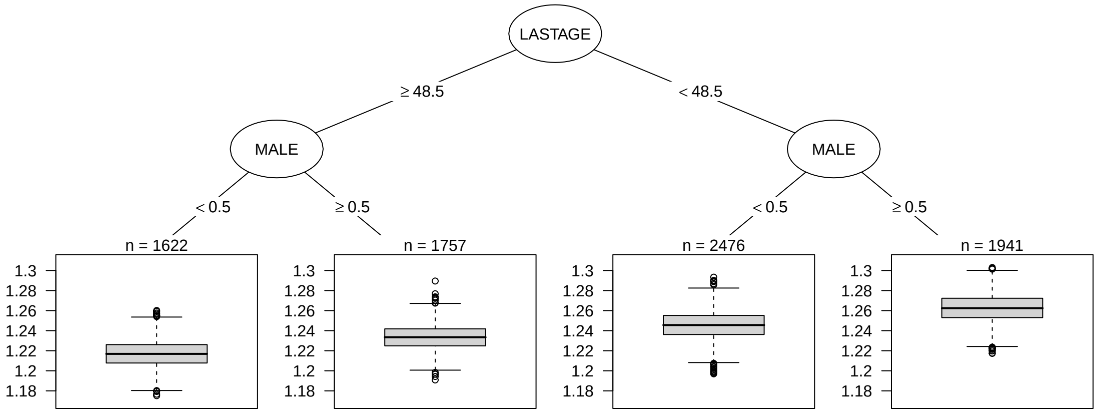


Figure 4.8: CART model. Response: Posterior mean of the exponential of estimated individual treatment effects.

4.5 Discussion

Our simulation of an experimental study toy example presented promising results, with the bcf+ model being capable of properly estimating the $CATE_{(z,z')}(\boldsymbol{x})$ for different values of z and z', when the SUTVA, strong ignorability and positivity assumptions hold.

For observational studies, as in the NMES Smoke data application, the results were in line with what was observed by Imai and Van Dyk (2004) and Hahn et al. (2020). Increases in *packyears* leads to an increase in the multiplicative factor on expected annual medical expenditures, and this effect generally increases as the gap in *packyears* widens. In a comparison between light smokers (8 *packyears*) vs heavy smokers (32 *packyears*), the estimated increase in annual medical expenditures had an average of 23.4%. Heterogeneity has been observed, specially between Age and Gender, but this effect may be due to survivor bias or higher medical expenditure baseline.

In general, the bcf+ presented results aligned with out expectations. Further analysis on controlled settings under targeted selection and the use of the general-

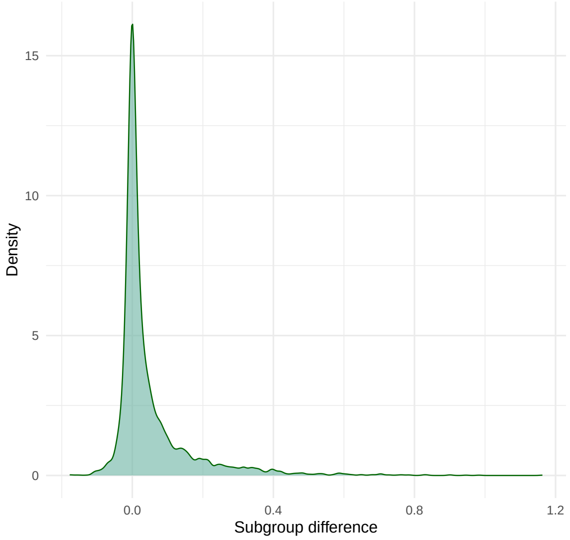


Figure 4.9: Subgroup difference - Men younger than 37.5 years vs women older than 58.5 years.

ized propensity score as a covariate still need to be performed, especially with more simulations where the RIC phenomenon might impact the estimation of treatment effects.

Chapter 5: Discussion

In this dissertation, three novel tree-based Bayesian methods were introduced. These methods are BART extensions that are providing possible solutions to open problems. LineART provides a BART alternative with locally adaptive linear regressions that can more easily adapt to smoother curves than a step function, while maintaining almost the same time efficiency as the original BART. The scalable tsBART solves the time efficiency issue that was originally present in tsBART when the targeted smoothing variable presented a large number of unique values. And the bcf+ addresses the issue of estimating treatment effects with continuous treatments for smooth curves by using a scalable tsBART prior.

Among the future extensions, the use of the approach introduced by Albert and Chib (1993) to extend both the LineART prior and the scalable tsBART prior to binary responses is straightforward, due to the use of latent variables, and basically no changes to the priors should be necessary. Another possible extension that can be explored is the use of a LineART prior to estimate the prognostic effect on bcf+, since LineART showed promising results in comparison to BART and the novel prior should perform well in most settings where BART struggles to smooth curves. All approaches can be studied under settings with sparsity, where variable selection, such as using a Dirichlet hyperprior (Linero, 2018), can be applied. Furthermore, the use of stochastic trees with Accelerated BART (He et al., 2019) would increase the time efficiency of models while maintaining most of the priors untouched.

In the observational studies, it is usually assumed that strong ignorability and positivity hold, however these assumptions are very strong. It is unlikely that there are no unmeasured confounders but even in the case where this is true, positivity might still pose a challenge. In the application of Chapter 4, for example, it is unlikely that young adults will have high values for the treatment variable *packyears*, but removing all of young adults from the dataset would remove a decent chunk of the data, which impacts the estimation of the treatment curve around lower values for *packyears*. There is no right answer about how to deal with the overlap issue in this example, and in many settings with continuous treatments, this issue might be hard to overcome.

In conclusion, we expect that the development of these novel methods will contribute to current and future data analysis problems, providing accessible software to practitioners in a multitude of fields, but especially in the causal inference.

Appendix A: Supplemental materials for Chapter 2

A.1 Details on the tree prior

Each tree prior is given by the same way as in Chipman et al. (2010), such that for a regression tree T_j,

$$\mathbb{P}(T_j) = \prod_{\gamma \notin \mathcal{L}} \mathbb{P}_{Split}(\gamma) \prod_{\gamma \in \mathcal{L}} (1 - \mathbb{P}_{Split}(\gamma)) \prod_{\gamma \notin \mathcal{L}} \mathbb{P}_{Rule}(\gamma), \tag{A.1}$$

where $\mathcal{L}$ is the set of the leaf nodes; $\mathbb{P}_{Split}(\gamma) = \alpha(1 + d_\gamma)^{-\beta}$, with α and β being hyperparameters, and d_γ being the depth of the node γ; $\mathbb{P}_{Rule}(\gamma)$ is the probability of choosing a variable among the available variables for the split on the node γ times the probability of choosing a cutpoint among the available cutpoints for the chosen variable on node γ, which in our case, both follow an Uniform distribution.

The prior for σ is also set following the guidelines of Chipman et al. (2010), such that

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}, \tag{A.2}$$

where $\nu = 3$ and λ is picked so that $\mathbb{P}(\sigma < \hat{\sigma}) = 0.90$, which are the default recommendations from the authors.

A.2 Further details on the Hilbert space Gaussian process approximations

As presented on Section 2.3.2 the tsBART prior rely on Gaussian Processes (Rasmussen and Williams, 2006) to create smooth curves over the selected variable t. In general, a Gaussian Process only depends of its covariance kernel C_θ to create the smooth curves. In our case, each entry of C_θ is given by $C_\theta(t, t') = \exp\left(\frac{(t-t')^2}{2\theta}\right)$ since our kernel is based into the squared exponential kernel, but other kernels could be used as well.

By the Weiner-Khintchim theorem (Rasmussen and Williams, 2006) we have that covariance function C_θ can be represented by using its spectral density form, such that

$$C_\theta(t, t') \coloneqq C_\theta(\boldsymbol{t}) = \frac{1}{2\pi} \int S(\boldsymbol{u}) \exp\left(i\boldsymbol{u}^T \boldsymbol{t}\right) du, \tag{A.3}$$

which means that there is a one to one relationship between the covariance kernel and its associated spectral density. Riutort-Mayol et al. (2020) use this result, viewing the covariance operator as a pseudo-differential operator, and by representing it as a formal series of Laplace operators they perform a Hilbert-space approximation of the series expansion.

The selected method is a reduced-rank approximation, such that the Matrix Inversion Lemma can be used to speed-up the computation of the solution to a Gaussian Process regression, which as noted by Riutort-Mayol et al. (2020), requires $\mathcal{O}\left(|\mathcal{T}|^3\right)$ operations. The approximation explored in this section allows the solution to be handled in $\mathcal{O}\left(|\mathcal{T}|m^2\right)$ operations, $m < |\mathcal{T}|$, where m is the number of basis functions used in the approximation.

This method is based on an extended domain of $\mathcal{T}$ such that the error of the approximation for values between $\min(\mathcal{T})$ and $\max(\mathcal{T})$ is controlled. This domain extension is done by multiplying the size of the domain (which is centered at zero) by the variable c. Since the original domain can shifted and scaled at will with appropriate modification to a (shift-invariant) kernel, here the domain of the data is always considered $[-1, 1]$.

The final approximation is of the form

$$C_\theta(t, t') \approx C^m_{\theta c}(t, t') = \sum_{q=1}^{m} \delta_{\theta c}\left(q\right) \phi_{qc}(t) \phi_{qc}(t'),$$

where

$$\phi_{qc}(t) = \sin\left(\frac{\pi q(t + c)}{2c}\right), \qquad \delta_{\theta c}\left(q\right) = \frac{\theta}{\sqrt{c}} \exp\left(-\frac{\theta^2 \left(\frac{\pi q}{2c}\right)^2}{2}\right).$$

Intuitively, the function $\delta_{\theta c}(.)$ works as a way to create weights to the basis functions that are used in the approximation. As it can be seen on Figure A.1, approximations with low values of θ will make the function $\delta_{\theta c}(.)$ decrease faster to zero due to the fact that θ defines how wiggly the Gaussian Process is, and therefore, the less wiggly a Gaussian Process is, the number of basis functions required to represent it decreases, while Gaussian Processes with low values of θ will require a higher order of basis functions to be represented (this can intuitively be seen by analyzing Figure A.2).

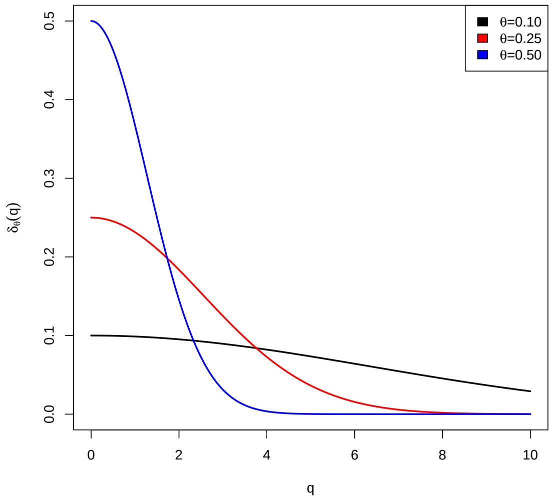


Figure A.1: The effect of θ over the function $\delta_{\theta c}(q)$; c=1.

It is also possible to perform a rigorous error analysis by controlling the selection of the hyperparameters m and c using the relative error in comparison with the true squared exponential kernel as shown on Equation (2.14).

Finally, one interesting result is that the method converges to the true covariance function as the values of c and m increase. As noted on Theorem 1 of Riutort-Mayol et al. (2020),

$$\lim_{c\to\infty}\left[\lim_{m\to\infty} C^{m}_{\theta c}(t,t')\right] = C_{\theta}(t,t'). \tag{A.4}$$

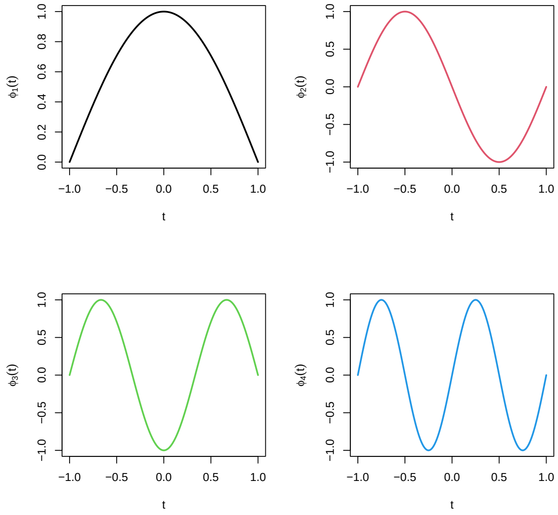


Figure A.2: $\phi_{qc}(t)$ for $q = \{1, 2, 3, 4\}$; c=1.

A.3 Algorithms to select m and c

Analyzing the two different examples presented on Figures 2.4 and 2.5, it is clear that once a value for m have been selected, it is possible to find the value of c which minimizes e_θ for its specific curve, therefore selecting the second hyperparameter needed to perform the proposed approximation. Using Algorithm 1 the practitioner can find m and c by just specifying the value for θ, $\epsilon_{\max}$ and *step* (the rate which c increases in the algorithm). We recommend $step = 0.01$.

Algorithm 1 Selecting m and c for a fixed θ

```
Require: $\epsilon_{\max} > 0$;  $step > 0$;  $\theta > 0$
Ensure: $m \in \mathbb{N}^*$;  $c \geq 1$                              ▷ Valid values for $m$ and $c$
  $m = 1$                                                               ▷ Start $m$
  $c = 1$                                                               ▷ Start $c$
  $\epsilon_{current} = e_\theta(m, c)$                                 ▷ Start $\epsilon_{current}$
  while $\epsilon_{current} \geq \epsilon_{\max}$ do                    ▷ While $\epsilon_{current}$ is not small enough
      $\epsilon_{last} = \epsilon_{current}$                            ▷ Update $\epsilon_{last}$
      $c = c + step$                                                    ▷ Update $c$
      $\epsilon_{current} = e_\theta(m, c)$                             ▷ Update $\epsilon_{current}$
      if $\epsilon_{current} > \epsilon_{last}$ then                    ▷ If $\epsilon_{current}$ is increasing, we passed the minimum
          $m = m + 1$                                                   ▷ Update $m$
          $c = 1$                                                       ▷ Restart $c$
          $\epsilon_{current} = e_\theta(m, c)$                         ▷ Update $\epsilon_{current}$
      end if
  end while
  return $(m; c)$                                                       ▷ Return the selected hyperparameters
```

Algorithm 2 basically expands the procedure implemented in Algorithm 1. In general, Algorithm 1 is used only for the selection of m, with the difference that instead of using a fixed θ, now θ_{smin} is used instead. After that, a value of θ close to zero, named $\theta_{\min}$, must be select. This value is used to create a sequence which increases by $step_\theta$. For every value of the sequence, the value c which minimizes e_θ for a given m is estimated, and the sequence will stop when c achieves a value $c_{\max}$ that is high enough to cover almost every reasonable scenario. We suggest $\theta_{\min} = 0.01$, $step_\theta = 0.01$, and $c_{\max} = 10$, which should cover almost all scenarios taking into consideration the fact that the domain of the data is always scaled between $[-1, 1]$.

Algorithm 2 Selecting m and defining the best c for different values of θ

```
Require: $\epsilon_{\max} > 0$; $step > 0$; $step_\theta > 0$; $\theta_{smin}$; $\theta_{\min}$; $c_{\max}$
Ensure: $m \in \mathbb{N}^*$; $c \geq 1$                                    ▷ Valid values for $m$ and $c$
  Select $m$ using Algorithm 1 with $\epsilon_{\max}$, $step$, and $\theta_{smin}$ for fixed $\theta$   ▷ $m$ is fixed
  $i = 1$
  $\theta = \theta_{\min}$
  $c = 1$
  $\epsilon_{current} = e_\theta(m, c)$
  while $c < c_{\max}$ do
      $\epsilon_{last} = \epsilon_{current}$                                  ▷ Update $\epsilon_{last}$
      $c = c + step$                                                ▷ Update $c$
      $\epsilon_{current} = e_\theta(m, c)$                                ▷ Update $\epsilon_{current}$
      if $\epsilon_{current} > \epsilon_{last}$ then     ▷ If $\epsilon_{current}$ is increasing, we passed the minimum
          $\theta_i = \theta$                                              ▷ Save $\theta$
          $c_i = c - step$                               ▷ Save $c$ for our minimum
          $\theta = \theta + step_\theta$                                      ▷ Update $\theta$
          $c = 1$                                                   ▷ Update $c$
          $i = i + 1$                                               ▷ Update $i$
      end if
  end while
  return $(m; (c_1, \theta_1); (c_2, \theta_2); ...)$         ▷ Return the selected hyperparameters
```

Appendix B: Supplemental materials for Chapter 3

B.1 Predictive performance table

Dataset	BART	LineART	LineART-CV	SoftBART
Abalone	0.36102(0.01935)	0.35667(0.01962)	0.35525(0.01998)	0.34929(0.01965)*
Ais	0.03830(0.01062)	0.01906(0.00770)	0.01800(0.00688)*	0.02174(0.00409)
Alcohol	0.95056(0.04305)	0.95293(0.04272)	0.95376(0.04308)	0.95009(0.04389)*
Amenity	0.27148(0.02584)	0.28340(0.02827)	0.27832(0.02915)	0.26542(0.02425)*
Attend	0.21614(0.03556)	0.20321(0.03090)	0.19396(0.03654)*	0.22343(0.02989)
Baseball	0.17723(0.05844)	0.17295(0.05881)	0.17182(0.05862)	0.16619(0.05251)*
Baskball	0.79292(0.32061)	0.78707(0.31351)	0.76930(0.32234)*	0.77629(0.31412)
Boston	0.16545(0.05032)	0.17175(0.04117)	0.18898(0.08014)	0.16438(0.04054)*
Budget	0.00111(0.00034)	0.00139(0.00125)	0.00107(0.00064)*	0.00281(0.00081)
Cane	0.54323(0.03193)	0.54376(0.03416)	0.52402(0.03212)*	0.55932(0.02969)
Cardio	0.83673(0.26855)	0.82812(0.26354)	0.82171(0.24752)*	0.82742(0.25867)
College	0.17209(0.02026)	0.16813(0.01763)	0.17157(0.01778)	0.16417(0.01784)*
Cps	0.61881(0.11344)*	0.62312(0.10951)	0.62491(0.11261)	0.61922(0.11800)
Cpu	0.02022(0.01498)	0.00467(0.00401)	0.00453(0.00367)	0.00223(0.00247)*
Deer	0.19516(0.06506)	0.25315(0.11550)	0.23405(0.15417)	0.18556(0.08201)*
Diabetes	0.40577(0.11214)	0.41035(0.12025)	0.42141(0.12001)	0.38946(0.11147)*
Diamond	0.00207(0.00104)*	0.00337(0.00206)	0.00295(0.00209)	0.00297(0.00120)
Edu	0.74506(0.06499)	0.74882(0.06859)	0.74014(0.06355)*	0.75029(0.06194)
Enroll	0.17683(0.05147)	0.17685(0.05107)	0.17324(0.04792)	0.16681(0.04961)*
Fame	0.04500(0.00941)	0.03224(0.00982)	0.03008(0.00587)	0.02446(0.00401)*
Fat	0.30057(0.06019)	0.28430(0.04807)*	0.29029(0.05018)	0.28777(0.05340)
Fishery	0.40996(0.03348)	0.46794(0.16740)	0.41051(0.03841)*	0.46406(0.03102)
Hatco	0.07373(0.02919)*	0.11167(0.05637)	0.09544(0.04075)	0.08178(0.03461)
Insur	0.02029(0.00224)	0.02104(0.00395)	0.02209(0.00890)	0.01936(0.00222)*
Laheart	0.47497(0.07939)	0.48944(0.08249)	0.47554(0.09714)	0.46210(0.08732)*
Medicare	0.73719(0.03804)	0.73703(0.03431)	0.74142(0.03794)	0.73543(0.03883)*
Mpg	0.12149(0.04712)	0.11772(0.04435)	0.11505(0.04202)	0.10763(0.03784)*
Mumps	0.14429(0.01557)	0.15282(0.01616)	0.12857(0.01376)*	0.19777(0.01805)
Mussels	0.12869(0.04155)	0.14606(0.06850)	0.13829(0.05689)	0.11682(0.04266)*
Ozone	0.26258(0.06529)	0.25468(0.05752)*	0.26228(0.05942)	0.25599(0.06067)
Price	0.15335(0.10057)	0.16437(0.11812)	0.15104(0.11323)*	0.16683(0.11922)
Rate	0.57846(0.15466)	0.63458(0.17538)	0.56862(0.12657)*	0.56909(0.15678)
Rice	0.34505(0.11453)	0.34117(0.12203)*	0.34527(0.12137)	0.37085(0.11472)
Scenic	0.48765(0.13072)	0.46514(0.14413)	0.47581(0.14580)	0.46095(0.13395)*
Servo	0.10229(0.13032)	0.11432(0.13271)	0.12093(0.14865)	0.07494(0.12619)*
Smsa	0.11812(0.05559)	0.08659(0.04073)	0.08788(0.04671)	0.08462(0.04047)*
Strike	0.32889(0.03965)*	0.33099(0.03972)	0.33103(0.04064)	0.33165(0.03845)
Tecator	0.02668(0.00540)	0.03354(0.02452)	0.02216(0.01256)	0.01501(0.00433)*
Tree	0.44994(0.17939)	0.39319(0.13278)	0.38107(0.12754)*	0.40597(0.14690)
Triazine	0.68949(0.27466)	0.66069(0.27719)	0.65940(0.26318)	0.63579(0.26511)*
Wage	0.50159(0.05144)*	0.50231(0.05081)	0.50306(0.05138)	0.50254(0.05132)

Table B.1: Average out-of-sample MSE by dataset (20 simulations each). Standard deviation is in parenthesis. Stars mark the competitor with the lowest average for each dataset.

Dataset	p-value
Abalone	0.003
Ais	<0.001
Alcohol	<0.001
Amenity	0.004
Attend	0.006
Baseball	0.336
Baskball	0.539
Boston	0.389
Budget	0.847
Cane	0.851
Cardio	0.401
College	0.106
Cps	0.095
Cpu	<0.001
Deer	0.025
Diabetes	0.63
Diamond	0.003
Edu	0.243
Enroll	0.998
Fame	<0.001
Fat	0.003
Fishery	0.098
Hatco	0.004
Insur	0.278
Laheart	0.123
Medicare	0.989
Mpg	0.52
Mumps	0.001
Mussels	0.153
Ozone	0.061
Price	0.55
Rate	0.006
Rice	0.502
Scenic	0.023
Servo	0.144
Smsa	<0.001
Strike	0.429
Tecator	0.639
Tree	0.001
Triazine	0.249
Wage	0.543

Table B.2: P-values of two-sided t-test for the average values of the ratio $\log\left(RMSE_{LineART}/RMSE_{BART}\right)$. Null hypothesis: $Average = 0$.

B.2 Likelihood

For simplification, the b index is omitted in the section. We can write the log-likelihood of the leaf node γ as

$$l\left(\gamma\right) = \log\left(\left(2\pi\right)^{-\frac{n}{2}} \det\left(\sigma^2\mathbf{I} + \Omega\Sigma\Omega^T\right)^{-\frac{1}{2}} \exp\left(\mathbf{y}^T\left(\sigma^2\mathbf{I} + \Omega\Sigma\Omega^T\right)^{-1}\mathbf{y}\right)\right).$$

By applying the results of Section B.5, we have

$$\begin{aligned} l\left(\gamma\right) = &\log\Big(\left(2\pi\right)^{-\frac{n}{2}}\left(\det\left(\sigma^2\mathbf{I}\right)\det\left(\mathbf{I} + \Sigma\Omega^T\Omega\right)\right)^{-\frac{1}{2}} \times \\ &\exp\left(\mathbf{y}^T\left(\sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{I}\Omega\left(\Sigma^{-1} + \Omega^T\sigma^{-2}\mathbf{I}\Omega\right)^{-1}\Omega^T\sigma^{-2}\mathbf{I}\right)\mathbf{y}\right)\Big), \end{aligned}$$

$$l\left(\Theta\right) = -\frac{n}{2}\log\left(2\pi\right) - n\log\left(\sigma\right) - \frac{1}{2}\det\left(\mathbf{I} + \frac{\Sigma\Omega^T\Omega}{\sigma^2}\right) - \frac{\mathbf{y}^T\mathbf{y}}{\sigma^2} - \frac{\mathbf{y}^T\Omega}{\sigma^2}\left(\Sigma^{-1} + \frac{\Omega^T\Omega}{\sigma^2}\right)^{-1}\frac{\Omega^T\mathbf{y}}{\sigma^2}.$$

It is easy to see that

$$\Omega^T\Omega = \begin{bmatrix} n & \sum_{i=1}^n x_{di} \\ \sum_{i=1}^n x_{di} & \sum_{i=1}^n x_{di}^2 \end{bmatrix}, \qquad \Omega^T\boldsymbol{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{di}y_i, \end{bmatrix}, \qquad \boldsymbol{y}^T\boldsymbol{y} = \sum_{i=1}^n y_i^2,$$

are the sufficient statistics for this leaf node.

B.3 The Dirichlet hyperprior

Linero (2018) introduced the DART model, which included a Dirichlet hyperprior when selecting the variables for splits. In our case, we have

$$s \sim \mathcal{Dirichlet}\left(\frac{\alpha}{P}, ..., \frac{\alpha}{P}\right),$$

where s is a vector of probabilities of size P. Initially, the implementation can set $\alpha = P$.

To sample s, the parameters of the Dirichlet can be updated by using categorical variables, and in this case m_j are the number of times that variable j have been used in the model. Therefore,

$$s \sim \mathcal{D}irichlet\left(\frac{\alpha}{P} + m_1, ..., \frac{\alpha}{P} + m_P\right).$$

For sampling α a Beta prior can be used, such that

$$\frac{\alpha}{\alpha + \rho} \sim \mathcal{B}eta\,(a, b)\,,$$

where $\rho = P$, $a = 0.5$, $b = 1$. We have used a slice sampler in this case.

For the Grow proposal, first the available variables are selected, and sampled according to the probabilities in s.

One of the strategies suggested by Linero (2018) for the software implementation is to not update the probabilities for the first few iterations (25% is suggested as a guideline). This allows the trees to grow before the variable selection is conducted and can help with the mixing of the model. This prior is readily available for implementation and improvement is expected specially under sparsity.

B.4 Metropolis-Hastings Step

The Metropolis-Hastings step is used to sample the new trees from the posterior distribution, and will be approached in further detail. This section has used as base the work of Kapelner and Bleich (2016).

The Metropolis ratio is given by

$$\begin{aligned}\alpha &= \frac{\mathbb{P}(T^* \to T)\mathbb{P}(T^*|R,\sigma^2)}{\mathbb{P}(T \to T^*)\mathbb{P}(T|R,\sigma^2)}\\ &= \frac{\mathbb{P}(T^* \to T)\mathbb{P}(R|\sigma^2,T^*)\mathbb{P}(T^*)}{\mathbb{P}(T \to T^*)\mathbb{P}(R|\sigma^2,T)\mathbb{P}(T)}\\ &= \frac{\mathbb{P}(T^* \to T)}{\mathbb{P}(T \to T^*)} \times \frac{\mathbb{P}(R|\sigma^2,T^*)}{\mathbb{P}(R|\sigma^2,T)} \times \frac{\mathbb{P}(T^*)}{\mathbb{P}(T)},\end{aligned}$$

where the first term is the transition ratio, the second term is the likelihood ratio, and the third term is the prior ratio.

The likelihood is the same as defined on Section B.2, and the tree prior is given by

$$\mathbb{P}(T) = \prod_{\gamma \notin \mathcal{L}} \mathbb{P}_{Split}(\gamma) \prod_{\gamma \in \mathcal{L}} (1 - \mathbb{P}_{Split}(\gamma)) \prod_{\gamma \notin \mathcal{L}} \mathbb{P}_{Rule}(\gamma) \prod_{\gamma \in \mathcal{L}} \mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathfrak{R}}(\gamma)},$$

where $\mathcal{L}$ is the set of the leaf nodes; $\mathfrak{R}$ is the set of the root nodes; $\mathbb{I}_{\gamma \in \mathfrak{R}}(\gamma)$ is an indicator function which is 1 when γ is a root node; $\mathbb{P}_{Split}(\gamma) = \frac{\alpha}{(1+d_\gamma)^\beta}$, with α and β being hyperparameters, and d_γ being the depth of the node γ; $\mathbb{P}_{Rule}(\gamma)$ is the probability of choosing a variable among the available variables for the split on the node γ times the probability of choosing a cutpoint among the available cutpoints for the chosen variable on node γ; and $\mathbb{P}_{Slope}(\gamma)$ is the probability of choosing a variable which will define the slope of the root node.

For different types of proposal, each term will be different. In this case, the focus is on the proposals GROW and PRUNE, that are the mostly commonly used proposals in the software implementations of Bayesian regression trees. It should be noted that the basic form of the Metropolis Ratio is almost the same as in the regular BART, with the difference being one term that is added to the Metropolis Ratio.

B.4.1 GROW

The proposal GROW consists of choosing randomly one of the leaf nodes of the current tree, and then growing two new leaf nodes under the chosen node.

For the transition rate, first let γ be the node chosen to be split. In this case, the transition ratio is given by

$$\frac{\mathbb{P}(T^* \to T)}{\mathbb{P}(T \to T^*)},$$

where,

$$\mathbb{P}(T \to T^*) = \mathbb{P}(GROW)\mathbb{P}(\text{Selecting } \gamma \text{ among the available nodes to grow})\mathbb{P}_{Rule}(\gamma),$$

$$\begin{aligned}\mathbb{P}(T^* \to T) =& \mathbb{P}(PRUNE)\mathbb{P}(\text{Selecting } \gamma \text{ among the available nodes to prune})\times \\ & \mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathcal{R}(\gamma)}}.\end{aligned}$$

The likelihood ratio (which is determined solely by the leaf nodes), can be simplified, since most part of the tree is exactly the same, with the exception of the modified terminal nodes. Then,

$$\frac{\mathbb{P}(R|\sigma^2, T^*)}{\mathbb{P}(R|\sigma^2, T)} = \frac{\mathbb{P}(R_{\gamma_L}|\sigma^2)\mathbb{P}(R_{\gamma_R}|\sigma^2)}{\mathbb{P}(R_\gamma|\sigma^2)} = \exp\left(l(\gamma_L) + l(\gamma_R) - l(\gamma)\right).$$

Like the likelihood ratio, most part of the tree prior can be simplified due to the fact that the two trees differ only by the modified terminal nodes. Let γ be the node chosen to be splitted, and let γ_L and γ_R be the new nodes that have grown from that split. Therefore,

$$\begin{aligned}\frac{\mathbb{P}(T^*)}{\mathbb{P}(T)} &= \frac{(1 - \mathbb{P}_{Split}(\gamma_L))(1 - \mathbb{P}_{Split}(\gamma_R))\mathbb{P}_{Split}(\gamma)\mathbb{P}_{Rule}(\gamma)}{(1 - \mathbb{P}_{Split}(\gamma))\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathcal{R}(\gamma)}}} \\ &= \frac{\left(1 - \frac{\alpha}{(1+d_{\gamma_L})^\beta}\right)\left(1 - \frac{\alpha}{(1+d_{\gamma_R})^\beta}\right)\left(\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Rule}(\gamma)}{\left(1 - \frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathcal{R}(\gamma)}}} \\ &= \frac{\left(1 - \frac{\alpha}{(1+d_\gamma+1)^\beta}\right)\left(1 - \frac{\alpha}{(1+d_\gamma+1)^\beta}\right)\left(\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Rule}(\gamma)}{\left(1 - \frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathcal{R}(\gamma)}}} \\ &= \frac{\left(1 - \frac{\alpha}{(2+d_\gamma)^\beta}\right)^2\left(\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Rule}(\gamma)}{\left(1 - \frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathcal{R}(\gamma)}}}.\end{aligned}$$

B.4.2 PRUNE

The PRUNE proposal consists of collapsing two leaf nodes, in such a way that their parent node will now become a leaf node. This is done by choosing randomly among the available nodes that can become a new leaf (nodes that are parents, but are not grandparents).

Let γ be the node chosen to become a leaf. In this case, the transition ratio is given as

$$\frac{\mathbb{P}(T^* \to T)}{\mathbb{P}(T \to T^*)},$$

where

$$\begin{aligned}\mathbb{P}(T \to T^*) =&\mathbb{P}(PRUNE)\mathbb{P}(\text{Selecting } \gamma \text{ among the available nodes to prune})\times\\ &\times \mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma \in \mathcal{R}}(\gamma)}\end{aligned}$$

$$\mathbb{P}(T^* \to T) = \mathbb{P}(GROW)\mathbb{P}(\text{Selecting } \gamma \text{ among the available nodes to grow})\mathbb{P}_{Rule}(\gamma).$$

The likelihood ratio (which is determined solely by the leaf nodes), can be simplified, since most part of the tree is exactly the same, with the exception of the modified terminal nodes. The ratio can be expressed as

$$\frac{\mathbb{P}(R|\sigma^2, T^*)}{\mathbb{P}(R|\sigma^2, T)} = \frac{\mathbb{P}(R_\gamma|\sigma^2)}{\mathbb{P}(R_{\gamma_L}|\sigma^2)\mathbb{P}(R_{\gamma_R}|\sigma^2)} = \exp\left(l(\gamma) - l(\gamma_L) - l(\gamma_R)\right).$$

Most part of the tree prior can be simplified due to the fact that the two trees differ only by the modified terminal nodes. Let γ be the node chosen to become a new leaf node, and let γ_L and γ_R be the new nodes that will be collapsed, then

$$
\begin{aligned}
\frac{\mathbb{P}(T^*)}{\mathbb{P}(T)} &= \frac{(1-\mathbb{P}_{Split}(\gamma))\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma\in\mathfrak{R}}(\gamma)}}{(1-\mathbb{P}_{Split}(\gamma_L))(1-\mathbb{P}_{Split}(\gamma_R))\mathbb{P}_{Split}(\gamma)\mathbb{P}_{Rule}(\gamma)} \\
&= \frac{\left(1-\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma\in\mathfrak{R}}(\gamma)}}{\left(1-\frac{\alpha}{(1+d_{\gamma_L})^\beta}\right)\left(1-\frac{\alpha}{(1+d_{\gamma_R})^\beta}\right)\left(\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Rule}(\gamma)} \\
&= \frac{\left(1-\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma\in\mathfrak{R}}(\gamma)}}{\left(1-\frac{\alpha}{(1+d_\gamma+1)^\beta}\right)\left(1-\frac{\alpha}{(1+d_\gamma+1)^\beta}\right)\left(\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Rule}(\gamma)} \\
&= \frac{\left(1-\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Slope}(\gamma)^{\mathbb{I}_{\gamma\in\mathfrak{R}}(\gamma)}}{\left(1-\frac{\alpha}{(2+d_\gamma)^\beta}\right)^2\left(\frac{\alpha}{(1+d_\gamma)^\beta}\right)\mathbb{P}_{Rule}(\gamma)}.
\end{aligned}
$$

B.5 Useful Results

In order to calculate the likelihood more efficiently, we can use the Matrix Inversion Lemma, and the Sylvester Determinant Theorem.

B.5.1 Matrix Inversion Lemma

Let $\mathbf{A}$ be a $n \times n$ matrix, $\mathbf{U}$ is a $n \times k$ matrix, $\mathbf{C}$ is a $k \times k$ matrix, $\mathbf{V}$ is a $k \times n$ matrix, then

$$
(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1}+\mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1}. \tag{B.1}
$$

B.5.2 Sylvester Determinant Theorem

Let $\mathbf{X}$ be a $n \times n$ matrix, $\mathbf{A}$ is a $n \times k$ matrix, $\mathbf{B}$ is a $k \times n$ matrix, then

$$
\det(\mathbf{X}+\mathbf{AB}) = \det(\mathbf{X})\det\left(\mathbf{I}+\mathbf{B}\mathbf{X}^{-1}\mathbf{A}\right). \tag{B.2}
$$

Works Cited

James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88 (422):669–679, 1993. doi: 10.1080/01621459.1993.10476321.

L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. *Classification and Regression Trees.* The Wadsworth Statistics/Probability Series, 1984.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443): 935–960, 1998.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1): 266–298, 2010.

Singfat Chu. Pricing the c's of diamond stones. *Journal of Statistics Education*, 9, 2001.

R. Dennis Cook and Sanford Weisberg. *An Introduction to Regression Graphics.* Addison-Wesley Professional, 1994.

Sameer K. Deshpande, Ray Bai, Cecilia Balocchi, Jennifer E. Starling, and Jordan Weiss. Vcbart: Bayesian trees for varying coefficients, 2024. URL `https://arxiv.org/abs/2003.06416`.

Peng Ding and Fan Li. Causal Inference: A Missing Data Perspective. *Statistical Science*, 33(2):214–237, 2018. doi: 10.1214/18-STS645. URL `https://doi.org/10.1214/18-STS645`.

Jacob Feldmesser. Computer Hardware. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5830D.

Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, 1991. doi: 10.1214/aos/1176347963. URL `https://doi.org/10.1214/aos/1176347963`.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 2001.

Jerome H. Friedman and Bernard W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21, 1989. ISSN 00401706.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006.

P. Richard Hahn, Carlos M. Carvalho, David Puelz, and Jingyu He. Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, 13(1):163 – 182, 2018. doi: 10.1214/16-BA1044. URL `https://doi.org/10.1214/16-BA1044`.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965 – 2020, 2020. doi: 10.1214/19-BA1195. URL `https://doi.org/10.1214/19-BA1195`.

J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black. *Multivariate Data Analysis*. Prentice Hall, Upper Saddle River, NJ, 1998.

Jingyu He, Saar Yalov, and P. Richard Hahn. XBART: Accelerated Bayesian Additive Regression Trees. *arXiv e-print, arXiv: 1810.02215v3*, 2019.

Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(Volume 7, 2020):251–278, 2020. ISSN 2326-831X. doi: https://doi.org/10.1146/annurev-statistics-031219-041110. URL `https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-031219-041110`.

Jennifer L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Keisuke Hirano and Guido W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley Sons, Ltd, 2004. ISBN 9780470090459. doi: https://doi.org/10.1002/0470090456.ch7. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/0470090456.ch7`.

Kosuke Imai and David A. Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

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

C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer. The ‘Trier Social Stress Test’–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, (28):76–81, 1993.

M. F. Kratz. Level crossings and other level functionals of stationary Gaussian processes. *Probab. Surv.*, (3):230–288, 2006.

Miguel Lázaro-Gredilla, Joaquin Quiñnero-Candela, Carl Edward Rasmussen, and An237;bal R. Figueiras-Vidal. Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 11(63):1865–1881, 2010. URL `http://jmlr.org/papers/v11/lazaro-gredilla10a.html`.

Antonio R. Linero. Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, 2018. doi: 10.1080/01621459.2016.1264957.

Antonio R. Linero. Softbart: Soft bayesian additive regression trees, 2022. URL `https://arxiv.org/abs/2210.16375`.

Antonio R. Linero and Yun Yang. Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5):1087–1110, 09 2018. ISSN 1369-7412. doi: 10.1111/rssb.12293. URL `https://doi.org/10.1111/rssb.12293`.

Antonio R. Linero, Piyali Basak, Yinpu Li, and Debajyoti Sinha. Bayesian Survival Tree Ensembles with Submodel Shrinkage. *Bayesian Analysis*, 17(3): 997 – 1020, 2022. doi: 10.1214/21-BA1285. URL `https://doi.org/10.1214/21-BA1285`.

E. I. George M. T. Pratola, H. A. Chipman and R. E. McCulloch. Heteroscedastic bart via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417, 2020. doi: 10.1080/10618600.2019.1677243. URL `https://doi.org/10.1080/10618600.2019.1677243`.

Jared S. Murray. Log-linear bayesian additive regression trees for multinomial logistic and count regression models, 2019. URL `https://arxiv.org/abs/1701.01503`.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–480, 1923.

English translation published in 1990. Originally in *Roczniki Nauk Rolniczych Tom X* [in Polish].

Estevão B. Prado, Rafael A. Moral, and Andrew C. Parnell. Bayesian additive regression trees with model trees. *Statistics and Computing*, 31(20), 2021. doi: 10.1007/s11222-021-09997-3.

Estevão B. Prado, Andrew C. Parnell, Rafael A. Moral, Nathan McJames, Ann O'Shea, and Keefe Murphy. Accounting for shared covariates in semiparametric Bayesian additive regression trees. *The Annals of Applied Statistics*, 19(1):302 – 328, 2025. doi: 10.1214/24-AOAS1960.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005.

C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning, 2006.

Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. Practical hilbert space approximate bayesian gaussian processes for probabilistic programming, 2020. URL https://arxiv.org/abs/2004.11408.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66(5):688–701, 1974. ISSN 0022-0663.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880.

Alex Smola and Peter Bartlett. Sparse greedy gaussian process regression. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 30(2):419–446, Mar 2020. ISSN 1573-1375. doi: 10.1007/s11222-019-09886-w. URL `https://doi.org/10.1007/s11222-019-09886-w`.

Rodney A. Sparapani, Brent R. Logan, Robert E. McCulloch, and Purushottam W. Laud. Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in Medicine*, 35(16):2741–2753, 2016. doi: https://doi.org/10.1002/sim.6893. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6893`.

Rodney A Sparapani, Lisa E Rein, Sergey S Tarima, Tourette A Jackson, and John R Meurer. Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics*, 21 (1):69–85, 07 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy032. URL `https://doi.org/10.1093/biostatistics/kxy032`.

Jennifer E. Starling, Jared S. Murray, Carlos M. Carvalho, Radek K. Bukowski, and James G. Scott. BART with Targeted Smoothing: An analysis of patient-specific stillbirth risk. *arXiv e-print, arXiv: 1805.07656v7*, 2019.

Jennifer E. Starling, Jared S. Murray, Patricia A. Lohr, Abigail R. A. Aiken, Carlos M. Carvalho, and James G. Scott. Targeted smooth bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation, 2020. URL `https://arxiv.org/abs/1905.09405`.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1775–1784, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/wilson15.html.

Spencer Woody, Carlos M. Carvalho, P. Richard Hahn, and Jared S. Murray. Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime, 2020. URL https://arxiv.org/abs/2007.09845.

David S. Yeager, Christopher J. Bryan, James J. Gross, Jared S. Murray, Danielle Krettek Cobb, Pedro H. F. Santos, Hannah Gravelding, Meghann Johnson, and Jeremy P. Jamieson. A synergistic mindsets intervention protects adolescents from stress. *Nature*, 607(7919):512–520, Jul 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04907-7.

Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 6:233–243, 1986.

Vita

Pedro Henrique Filipini dos Santos was born in Brazil. He earned a Bachelor of Science in Economics from the Federal University of São Paulo in 2015, followed by a Master of Science in Statistics from the University of São Paulo in 2019.

In the same year, he began his doctoral studies in the Statistics group within the Department of Information, Risk, and Operations Management (IROM) at the University of Texas at Austin. His research interests include Bayesian statistics, regression trees, and causal inference.

Address: pedro.santos@mccombs.utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.