VCBART: Bayesian Trees for Varying Coefficients

Sameer K. Deshpande[*], Ray Bai[†], Cecilia Balocchi[‡],
Jennifer E. Starling[§], and Jordan Weiss[¶]

Abstract. The linear varying coefficient models posits a linear relationship between an outcome and covariates in which the covariate effects are modeled as functions of additional effect modifiers. Despite a long history of study and use in statistics and econometrics, state-of-the-art varying coefficient modeling methods cannot accommodate multivariate effect modifiers without imposing restrictive functional form assumptions or involving computationally intensive hyperparameter tuning. In response, we introduce VCBART, which flexibly estimates the covariate effect in a varying coefficient model using Bayesian Additive Regression Trees. With simple default settings, VCBART outperforms existing varying coefficient methods in terms of covariate effect estimation, uncertainty quantification, and outcome prediction. We illustrate the utility of VCBART with two case studies: one examining how the association between later-life cognition and measures of socioeconomic position vary with respect to age and socio-demographics and another estimating how temporal trends in urban crime vary at the neighborhood level. An R package implementing VCBART is available at `https://github.com/skdeshpande91/VCBART`.

Keywords: tree ensembles, local linear modeling, posterior contraction.

MSC2020 subject classifications: Primary 62F15; secondary 91G60, 62F12, 62P25.

1 Introduction

1.1 Motivation

The linear varying coefficient model specifies a linear relationship between an outcome Y and p covariates $X_1, \ldots, X_p$ that is allowed to change according to the values of R *effect modifiers* $Z_1, \ldots, Z_R$. That is, the model asserts that

$$Y = \beta_0(\boldsymbol{Z}) + \beta_1(\boldsymbol{Z})X_1 + \cdots + \beta_p(\boldsymbol{Z})X_p + \varepsilon, \tag{1.1}$$

where the $\beta_0(\boldsymbol{Z}), \ldots, \beta_p(\boldsymbol{Z})$ are *functions* mapping $\mathbb{R}^R$ to $\mathbb{R}$ and the residual error ε has mean zero. In this paper, we use Bayesian Additive Regression Trees (BART; Chipman et al., 2010) to learn the covariate effect functions $\beta_j(\boldsymbol{Z})$, expressing each with

[*]Dept. of Statistics, University of Wisconsin–Madison, Madison USA, sameer.deshpande@wisc.edu
[†]Dept. of Statistics, University of South Carolina, Columbia, USA, RBAI@mailbox.sc.edu
[‡]School of Mathematics, University of Edinburgh, Edinburgh, UK, cecilia.balocchi@ed.ac.uk
[§]Mathematica Inc., Cambridge, USA, JStarling@mathematica-mpr.com
[¶]Dept. of Precision Medicine & Optimal Aging Institute, NYU Grossman School of Medicine, New York, USA, jordan.weiss@nyulangone.org

an ensemble of binary regression trees. Our proposed method, which we call VCBART, can produce extremely accurate estimates and well-calibrated uncertainty intervals of evaluations $\beta_j(\boldsymbol{z})$ and predictions $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}]$ without imposing rigid parametric assumptions or requiring computationally intensive tuning. VCBART further enjoys strong theoretical guarantees and scales gracefully to large datasets with tens of thousands of observations. The following applications motivate our work.

Socioeconomic position and cognition   A large body of evidence suggests that socioeconomic position (SEP) at different points in the life course is an important determinant of cognitive function in mid-life and older adulthood (Luo and Waite, 2005; Lyu and Burr, 2016; Marden et al., 2017; Greenfield and Moorman, 2019; Zhang et al., 2020). A critical open challenge in life course research involves estimating how the associations between cognition and various SEP measures evolve over time and with respect to sociodemographic characteristics. Typically, authors related cognitive outcome to several SEP variables with linear models that included pre-specified interactions between the SEP covariates and characteristics like age, gender, and race. For instance, both Lyu and Burr (2016) and Marden et al. (2017) modeled an interaction between SEP and age while Aartsen et al. (2019) introduced an additional interaction with age-squared. In doing so, these authors implicitly make strong parametric assumptions about how the associations of interest vary. Unfortunately such functional form assumptions may be inadequate as many of these associations can weaken or level off as people age (see Dupre, 2007, and references therein).

In Section 5.1, we use VCBART to estimate whether and how the associations between total score on a battery of cognitive assessments and measures of SEP in childhood, early adulthood, and later-life can vary with respect to sociodemographic factors like age, race, and gender. Our dataset comes from the Health and Retirement Study (HRS), which is a nationally representative longitudinal study of US-based adults, and contains $N = 67{,}988$ total observations of $n = 10{,}812$ subjects. Using VCBART, we found little temporal variation in the association between childhood SEP and later-life cognition. Further, after adjusting for SEP in childhood and late adulthood, the association between early adulthood SEP and later-life cognition varied with respect to race and gender but did not vary substantially over time.

Crime in Philadelphia   Over multiple studies of neighborhood-level crime, Balocchi and colleagues have discovered that (i) partially pooling data across adjacent neighborhoods often improves crime forecasts but (ii) ignoring potential spatial discontinuities can yield highly biased forecasts (Balocchi and Jensen, 2019; Balocchi et al., 2021, 2023). Spatial discontinuities not only occur along known geographic landmarks (e.g., highways, parks, or rivers) but can also coincide with less visible differences in neighborhood-level demographic and socioeconomic dimensions. Balocchi et al. (2023) fit linear models regressing a transformed crime density onto a time index in each of Philadelphia's 384 census tracts. Concerned about potential spatial discontinuities in the tract-specific slopes and intercepts, they estimated two partitions of the tracts, one for the slopes and one for the intercepts, in which parameter values were similar within clusters but not

across clusters. Working within a Bayesian framework, they identified several high posterior probability partition pairs with an ensemble optimization procedure that greedily searched over the space of pairs of partitions of census tracts.

Although they obtained promising predictive results, their analysis is limited by the assumption that the number of reported crimes within each census tract is monotonic in time. An arguably more realistic model would (i) incorporate higher-order trends to better capture potential non-linearities and (ii) allow the corresponding model coefficients to vary across tracts. That is, a more realistic model would allow for different non-monotonic crime trends in different parts of the city. Unfortunately, extending Balocchi et al. (2023)'s optimization procedure to cover higher-order models is computationally impractical, as it must search over a combinatorially vast product space of partitions. In Section 5.2, we recast their model as a linear varying coefficient model with a single categorical effect modifier Z recording the census tract. Using VCBART to fit several tract-specific polynomial models, we find that a quartic model of crime provides a better fit to data than their linear model.

1.2   Our contributions

In both problems, we wish to make minimal assumptions about the functional form of the covariate effects $\beta_j(\boldsymbol{Z})$. Additionally, at least for the HRS data, we wish to identify which elements of Z modify the effects of which elements of X. Further, in both applications, calibrated and coherent uncertainty quantification is imperative. For the HRS data, identifying disparities in the socioeconomic determinants of later-life cognition carries profound public health implications, especially in light of an aging population. And for the crime dataset, accurate assessments of the uncertainty about crime forecasts can inform future public safety policy decisions. Finally, we require a method that can scale to the size of the two motivating datasets, ideally without involving computationally intensive hyperparameter tuning.

Unfortunately, many existing state-of-the-art procedures for fitting varying coefficient models with $R > 1$ modifiers involve tuning several parameters with leave-one-out cross-validation or rigidly assume the $\beta_j(\boldsymbol{Z})$'s are additive in the Z_r's. Moreover, their default implementations often do not provide any out-of-sample uncertainty quantification.

We instead approximate each $\beta_j(\boldsymbol{Z})$ with a sum of regression trees. On synthetic data, our proposed procedure VCBART exhibits superior covariate effect recovery compared to the current state-of-the-art *without requiring any hand-tuning or strong structural assumptions.* We additionally show that, under mild conditions, the VCBART posterior concentrates at a near-optimal rate, even with correlated residual errors. To the best of our knowledge, our Theorem 3.1 is the first result demonstrating the theoretical near-optimality of Bayesian treed regression in settings with non-i.i.d. noise. VCBART also generalizes Hahn et al. (2020)'s Bayesian Causal Forests (BCF) model in order to estimate heterogeneous treatment effects in settings with multiple binary or continuous treatments.

Here is an outline for the rest of the paper. We briefly review relevant background on VC modeling and BART in Section 2. Then, in Section 3, we introduce VCBART, describe how to perform posterior inference, and state our asymptotic results. In Section 4, we demonstrate VCBART's excellent covariate effect recovery and predictive capabilities using synthetic data. We apply VCBART to our motivating datasets in Section 5 before outlining several avenues for future work in Section 6.

2 Background

2.1 Varying coefficient models

Since their introduction in Hastie and Tibshirani (1993), varying coefficient models have been extensively studied and deployed in statistics and econometrics. We give a very brief overview here; see Fan and Zhang (2008) and Franco-Villoria et al. (2019) for more comprehensive reviews.

When there is only a single modifier (i.e. $R = 1$), one popular approach to fitting the model in (1.1) is to express each $\beta_j(\boldsymbol{Z})$ as a linear combination of pre-specified basis functions (see, e.g., Hoover et al., 1998; Huang et al., 2002). Such a decomposition effectively reduces the functional regression problem to a high-dimensional linear regression problem for which numerous frequentist (see, e.g., Wang et al., 2008; Wang and Xia, 2009; Wei et al., 2011) and Bayesian (Bai et al., 2023) regularization techniques have been proposed. Kernel smoothing is another popular and theoretically supported alternative when $R = 1$ (Wu and Chiang, 2000).

Varying coefficient models are commonly encountered in spatial statistics (see, e.g., Gelfand et al., 2003; Finley and Banerjee, 2020). In these settings, there are usually $R = 2$ modifiers (space) or $R = 3$ modifiers (space and time) and the $\beta_j(\boldsymbol{Z})$'s are typically modeled with Gaussian processes. When the number of observations is large, full Bayesian inference with these models can be computationally demanding, necessitating the use of clever approximations or advanced techniques like distributed computing (Guhaniyogi et al., 2022). Outside of spatial contexts, to accommodate $R > 1$ modifiers, Tibshirani and Friedman (2020) and Lee et al. (2018) respectively constrained the $\beta_j(\boldsymbol{Z})$'s to be linear and additive functions of the Z_r's. In contrast, Li and Racine (2010) proposed a multivariate kernel smoothing estimator that imposes no rigid structural assumptions on the covariate effects. However, the default implementation of their procedure tunes several bandwidth parameters with leave-one-out cross-validation. When n, p, or R are large, their procedure is computationally prohibitive.

In the last decade, several authors have used regression trees to estimate the covariate effects. Bürgin and Ritschard (2015), for instance, modeled each $\beta_j(\boldsymbol{Z})$ with a single regression tree while Wang and Hastie (2012) and Zhou and Hooker (2022) both constructed separate ensembles for each $\beta_j(\boldsymbol{Z})$ using boosting. Unfortunately, existing tree-based procedures require substantial tuning and often do not automatically return any uncertainty estimates.

Before proceeding, we pause to highlight an important difference between VCBART and Bai et al. (2023). That work is limited to settings with only one effect modifier

(i.e., $R = 1$) and where interest lies primarily in determining which X_j's affect Y (i.e., whether $\beta_j(\boldsymbol{z}) = 0$ for all $\boldsymbol{z}$). In contrast, VCBART allows for an arbitrary number of effect modifiers, estimates individual evaluations $\beta_j(\boldsymbol{z})$, and determines which Z_r's drive variation in each function $\beta_j(\boldsymbol{Z})$.

2.2 Bayesian Additive Regression Trees

Chipman et al. (2010) introduced BART in the context of the standard nonparametric regression problem: given n observations from the model $y \sim \mathcal{N}(f(\boldsymbol{x}), \sigma^2)$, they estimated f with a sum of M piecewise constant step functions, which they represented as binary regression trees. Formally, they computed a posterior distribution over regression tree ensembles, which induces an approximate posterior distribution over regression function evaluations $f(\boldsymbol{x})$. Their regression tree prior, which we adopt and detail in Section 3.1, introduces strong regularization and encourages trees to be "weak learners" in the sense that no one tree explains too much variation in the observed y's.

The basic BART model has been extended successfully to survival analysis (Sparapani et al., 2016), multiple imputation (Xu et al., 2016), log-linear models (Murray, 2021), semi-continuous responses (Linero et al., 2020), and causal inference (Hill, 2011; Hahn et al., 2020). With conceptually simple modifications, BART can also recover smooth (Linero and Yang, 2018; Starling et al., 2020) and monotonic (Starling et al., 2019; Chipman et al., 2022) functions. In each setting, BART-based methods often substantially outperform existing procedures in terms of function recovery, prediction, and ease-of-use. Indeed, nearly every BART extension provides default hyperparameters that yield generally excellent performance "off-the-shelf." Further, recent results in Ročková and Saha (2019) and Ročková and van der Pas (2020) demonstrate BART's theoretical near-optimality under very mild assumptions. See Tan and Roy (2019) and Hill et al. (2020) for more detailed reviews of BART and its many extensions.

3 The VCBART procedure

Without loss of generality, suppose that we have R_{cont} continuous modifiers, each of which has been re-scaled to lie in the unit interval $[0, 1]$, and R_{cat} categorical or discrete modifiers. For $r = 1, \ldots, R_{\mathrm{cat}}$, we will assume that the r-th categorical modifier can take K_r distinct values contained in some discrete set $[K_r]$. We further concatenate our modifiers into a vector $\boldsymbol{Z}$ whose first R_{cont} entries record the values of the continuous modifiers and whose last R_{cat} entries record the values of the categorical modifiers. In other words, each observed modifier vector $\boldsymbol{z}$ lies in the product space $\mathcal{Z} = [0, 1]^{R_{\mathrm{cont}}} \times [K_1] \times \cdots \times [K_{R_{\mathrm{cat}}}]$.

Both motivating applications involve repeated measurements over time: in the HRS dataset, we have between four and eight observations per subject and in the crime dataset, we have yearly crime densities in each census tract. For simplicity, we will refer to the observed units (i.e., HRS participants and Philadelphia census tracts) as "subjects." For each subject $i = 1, \ldots, n$, we observe n_i triplets $(\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, y_{it})$ of covariates

$\boldsymbol{x}$, modifiers $\boldsymbol{z}$, and outcome y. For all $i = 1, \ldots, n$, and $t = 1, \ldots, n_i$, we model

$$y_{it} = \beta_0(\boldsymbol{z}_{it}) + \sum_{j=1}^{p} \beta_j(\boldsymbol{z}_{it})x_{itj} + \sigma\varepsilon_{it}. \tag{3.1}$$

We further assume $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^\top \sim \mathcal{N}_{n_i}(\boldsymbol{0}_{n_i}, \boldsymbol{\Sigma}_i(\rho))$ where $\boldsymbol{\Sigma}_i(\rho)$ is a correlation matrix with off-diagonal elements equal to $0 \leq \rho < 1$. In other words, we assume an exchangeable or compound symmetry correlation structure for each subject's errors. We assume that the noise vectors $\boldsymbol{\varepsilon}_i$'s are independent across subjects. We do not assume that we observe each subject an equal number of times nor do we assume that the observation times are equally spaced.

The key idea of VCBART is to approximate each function $\beta_j(\boldsymbol{Z})$ with its own regression tree ensemble. In Section 3.1 we describe the prior over the regression trees and in Section 3.2, we outline strategies for posterior computation and modifier selection.

3.1   Regression tree prior

To set our notation, let T be a binary decision tree that consists of a collection of internal nodes and a collection of terminal or *leaf nodes*. Each internal node of T is associated with a decision rule $\{Z_r \in \mathcal{C}\}$. When Z_r is a continuous variable, $\mathcal{C}$ is a half-open interval of the form $[0, c)$ and when Z_r is categorical, $\mathcal{C}$ is a subset of $[K_{r-R_{\text{cont}}}]$.

Given T and any vector $\boldsymbol{z}$, we can imagine tracing a path from the root down the tree by following the decision rules. Specifically, any time the path encounters the rule $\{Z_r \in \mathcal{C}\}$, it proceeds to the left child if $z_r \in \mathcal{C}$. Otherwise, it proceeds to the right child. Such decision-following paths continue until they reach a leaf node. It is not difficult to verify that every such path terminates in a single leaf node, implying that T induces a partition of $\mathcal{Z}$ into disjoint regions, one for each leaf. By associating each leaf node ℓ of T with a scalar *jump* μ_ℓ, the pair $(T, \boldsymbol{\mu})$ represents a piecewise constant function over $\mathcal{Z}$, where $\boldsymbol{\mu}$ denotes the collections of jumps (see Figure 1 for an example).

For each $j = 0, \ldots, p$, we introduce an ensemble $\mathcal{E}^{(j)} = \{(T_m^{(j)}, \boldsymbol{\mu}_m^{(j)})\}_{m=1}^{M}$ of M regression trees to approximate the function $\beta_j(\boldsymbol{Z})$. We model each ensemble independently and place independent and identical priors on the individual regression trees within each ensemble. Following Chipman et al. (2010), we specify a scaled inverse-χ^2 prior for the residual error variance $\sigma^2 \sim \lambda\nu\chi_\nu^{-2}$. Like that work, we take $\nu = 3$ and set λ so that there is 90% prior probability on the event that σ is less than the standard deviation of the observed responses Y. We complete our prior specification with a uniform prior on the autocorrelation ρ.

We specify the prior for an individual tree $(T_m^{(j)}, \boldsymbol{\mu}_m^{(j)})$ in the j-th ensemble compositionally with a marginal prior for the decision tree and a conditional prior for the jumps given the tree. Just like in the original BART model, we model the jumps in $\boldsymbol{\mu}_m^{(j)}$ as *a priori* independent $\mathcal{N}(0, \tau_j^2)$ random variables, where τ_j is a positive hyperparameter. Thus, for each $\boldsymbol{z} \in \mathcal{Z}$, the implied marginal prior of $\beta_j(\boldsymbol{z})$ is $\mathcal{N}(0, M\tau_j^2)$. We have found taking $\tau_j = 0.5/\sqrt{M}$ to be an effective default choice; see Section S2 of
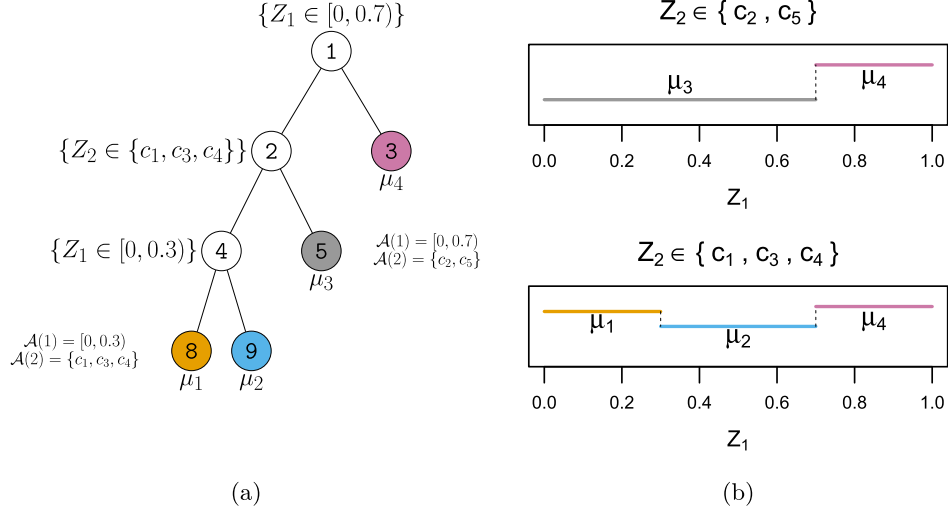
Figure 1: A regression tree defined over $\mathcal{Z} = [0,1] \times \{c_1, \ldots, c_5\}$ (a) and its step function representation (b). In (a), we also report the sets of available values of Z_1 and Z_2 at nodes 5 and 8.

the Supplementary Materials (Deshpande et al., 2024) for a hyperparameter sensitivity analysis.

We specify the decision tree prior implicitly by describing how to sample from the prior. First, we draw the overall graphical structure with a branching process that starts from a root node, which is initially considered terminal. Whenever a new terminal node is created at depth d, we attach two child nodes to it with probability $0.95(1 + d)^{-2}$. The quadratic decay ensures that the process terminates at a finite depth.

Given the graphical structure, we draw decision rules at each non-terminal node in two steps. First, we sample the splitting variable index $v \sim \text{Multinomial}(\theta_{j1}, \ldots, \theta_{jR})$ and θ_{jr} is the probability of selecting Z_r. We place a symmetric Dirichlet(η_j/R) prior on the vector of splitting probabilities and place a further Beta prior over $\eta_j/(\eta_j + R)$. The implied prior density of η_j is proportional to $(R + \eta_j)^{-(R+1)}$. Although the η_j prior favors sparsity, it has the capacity to include more variables as needed (Linero, 2018, §3.3). As evidenced by the simulation studies in Linero (2018), the prior hierarchy on the splitting variable works well in both the sparse setting, when the function depends on only a few variables, and in the non-sparse setting.

Then, conditional on r, we set $\mathcal{C}$ to be a random subset of $\mathcal{A}(r)$, the set of available Z_r values at the current non-terminal node (see annotations in Figure 1a). When Z_r is continuous, $\mathcal{A}(r)$ is an interval and we set $\mathcal{C} = [0, c)$ where c is drawn uniformly from $\mathcal{A}(r)$. When Z_r is categorical, $\mathcal{A}(r)$ is a discrete set. How we draw the random subset $\mathcal{C}$ depends on whether Z_r displays network structure like the census tracts in our crime data. If not, we assign each element of $\mathcal{A}(r)$ to $\mathcal{C}$ with probability 0.5. Otherwise, we

form $\mathcal{C}$ by drawing a random network partition. Briefly, we partition the vertices of the subgraph induced by $\mathcal{A}(r)$ by deleting a random edge from a random spanning tree of the subgraph; see Section 5.2 below and Deshpande (2023) for more details.

3.2   Posterior computation and modifier selection

Conditional regression tree updates   Because the VCBART posterior is analytically intractable, we simulate posterior draws using a Gibbs sampler. Briefly, we sequentially update each tree $(T_m^{(j)}, \boldsymbol{\mu}_m^{(j)})$ conditionally on the other $Mp - 1$ regression trees using two steps. First, we update the decision tree structure $T_m^{(j)}$ with a Metropolis-Hastings step and transition kernel that randomly grows or prunes the tree. Like Chipman et al. (2010)'s original BART sampler, this update depends on the data and other regression trees through a particular partial residual. Then, given $T_m^{(j)}$, we draw the jumps $\boldsymbol{\mu}_m^{(j)}$ from its conditional posterior distribution. Essentially, this amounts to fitting an intercept-free conjugate Bayesian linear regression model in each leaf of the tree. Whereas the jumps in the original BART model are conditionally independent, the correlated errors in (3.1) render the jumps in VCBART conditionally dependent. After updating all trees in the ensemble for β_j, we perform a conjugate Dirichlet-Multinomial update for the vector of splitting probabilities $\boldsymbol{\theta}_j$. After updating the trees in the jth ensemble and $\boldsymbol{\theta}_j$, we update η_j using an independence Metropolis step. After sweeping over all ensembles, we perform a conditionally conjugate Inverse Gamma update for σ^2 and sample a new ρ using a random-walk Metropolis update on the unconstrained quantity $\log(\frac{\rho}{1-\rho})$. See Section S4 of the Supplementary Materials (Deshpande et al., 2024) for a detailed derivation of our Gibbs sampler.

Modifier selection   In our sampler, in order to update the vector of splitting index probabilities $\boldsymbol{\theta}_j$, we keep track of the number of times each modifier Z_r is used in a decision rule in the ensemble $\mathcal{E}_j$. Using these counts, we can estimate probability that each Z_r is selected at least once in the ensemble $\mathcal{E}_j$ used to estimate $\beta_j(\boldsymbol{Z})$. These selection probabilities are analogous to the posterior inclusion probabilities encountered in Bayesian sparse linear regression. Based on this interpretation, for each ensemble $\mathcal{E}_j$, we construct an analog of Barbieri and Berger (2004)'s median probability model by reporting those modifiers Z_r whose selection probability exceeds 0.5.

Remark 1 (Identifiability). *For each realization of the modifiers $\boldsymbol{z}$, the vector of covariate function evaluations $\boldsymbol{\beta}(\boldsymbol{z}) = (\beta_0(\boldsymbol{z}), \ldots, \beta_p(\boldsymbol{z}))$ is identifiable as long as $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top | \boldsymbol{Z} = \boldsymbol{z}]$ is positive definite (Huang and Shen, 2004). The individual regression trees, however, are not identified. This is because we can switch the order of the trees without changing the predictions. Nevertheless, in our simulation studies, we did not experience any difficulties recovering the true covariate effects.*

In light of Remark 1, we do not suggest diagnosing Markov chain convergence using the samples of individual trees. We instead recommend tracking convergence using samples of σ, the residual standard deviation. In our experiments, we have found that although a large number of MCMC iterations (i.e., 20,000 or 50,000) may be needed for

adequate mixing (e.g, Gelman-Rubin $\hat{R} < 1.1$), one can often obtain excellent predictions and well-calibrated uncertainty intervals using many fewer iterations (i.e., 2,000); see Section 4 and Section S2.1 of the Supplementary Materials (Deshpande et al., 2024).

3.3   Posterior contraction of VCBART

With some minor modifications, VCBART enjoys essentially the same favorable theoretical properties as BART — namely, the posterior concentrates at a nearly minimax optimal rate. To facilitate our theoretical analysis of the VCBART posterior, we assume that all modifiers lie in $[0, 1]^R$. In practice, we can always rescale continuous modifiers to lie in the unit interval, and with one-hot encoding, we can represent categorical modifiers with multiple binary indicators. However, our main theorem still holds as long when all modifiers in $\boldsymbol{Z}$ are uniformly bounded. We additionally assume that the observed modifiers $\boldsymbol{z}_{it} \in \mathbb{R}^p, i = 1, \ldots, n, t = 1, \ldots, n_i$, are fixed and follow a regular design (Definition 3.3 in Ročková and van der Pas (2020)). We refer readers to Ročková and van der Pas (2020) and Ročková and Saha (2019) for a technical definition. At a high level, design regularity implies that a k-d tree partition (Bentley, 1975) of the dataset results in cells with similar diameters (i.e., similar maximal distance between two points in a cell).

Like Ročková and Saha (2019), we modify the decision tree prior so that the probability that the branching process continues growing at depth d is γ^d for some fixed $N^{-1} < \gamma < 1/2$. We assume that for each $i = 1, \ldots, n$ and $t = 1, \ldots, n_i$, the ground truth varying coefficient model is

$$y_{it} = \beta_{0,0}(\boldsymbol{z}_{it}) + \sum_{j=0}^{p} \beta_{0,j}(\boldsymbol{z}_{it})x_{itj} + \sigma_0\varepsilon_{it}, \qquad (3.2)$$

where each $\beta_{0,j}(\boldsymbol{Z}), j = 0, \ldots, p$, is α_j-Hölder continuous with $0 < \alpha_j \le 1$, and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}_{n_i}, \boldsymbol{\Sigma}_i(\rho_0))$, where the $(j, k)^{\text{th}}$ entry of $\boldsymbol{\Sigma}_i(\rho_0)$ is $\mathbb{1}(j = k) + \rho_0\mathbb{1}(j \neq k)$.

Recall that $N = \sum_{i=1}^{n} n_i$ is the total number of observations for all n subjects, R is the number of effect modifiers, and p is the number of covariates. Let $n_{\max} = \max\{n_1, \ldots, n_n\}$ denote the maximum number of within-subject observations. For two nonnegative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ to denote $0 < \liminf_{n\to\infty} a_n/b_n \le \limsup_{n\to\infty} a_n/b_n < \infty$. Meanwhile, $a_n = O(b_n)$ means that for sufficiently large n, there exists a constant $C > 0$ independent of n such that $a_n \le Cb_n$. We make the following assumptions.

(A1) The true varying coefficients satisfy $\max_{\boldsymbol{Z} \in [0,1]^R} |\beta_{0,j}(\boldsymbol{Z})| < \infty$ for all $0 \le j \le p$.

(A2) There is a constant $D > 1$ so that $|x_{itj}| \le D$ for all $1 \le i \le n$, $1 \le t \le n_i$, $1 \le j \le p$.

(A3) R, p, and $n_{\max}$ satisfy $R = O((\log N)^{1/2})$, $p = O(1)$, and $n_{\max} \asymp N/n$.

Assumption (A1) assumes that the true varying coefficients $\beta_j(\boldsymbol{Z})$'s are uniformly bounded. This assumption is also made in Zhou and Hooker (2022) and is likely to

be satisfied in practice since the $\boldsymbol{Z}$'s are uniformly bounded. Assumption (A2) assumes that all of the covariates $\boldsymbol{X}$ are also uniformly bounded. Finally, Assumption (A3) specifies appropriate rates of growth for R, p, and $n_{\max}$.

Let $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ be $N \times (p+1)$ matrices whose respective (i, j)-th entries are $\beta_j(\boldsymbol{z}_{it})$ and $\beta_{0,j}(\boldsymbol{z}_{it})$. Let $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_N^2 = N^{-1} \sum_{i=i}^n \sum_{t=1}^{n_i} \sum_{j=0}^p [\beta_j(\boldsymbol{z}_{it}) - \beta_{0,j}(\boldsymbol{z}_{it})]^2$ be the squared empirical ℓ_2 norm. If we knew the true smoothness levels α_j, the minimax rate for estimating $\boldsymbol{\beta}_0$ in $\|\cdot\|_N^2$ would be $\sum_{j=0}^p N^{-2\alpha_j/(2\alpha_j+R)}$ (Ročková and van der Pas, 2020). This is because the minimax estimation rate for each individual α_j-Hölder smooth function $\beta_{0,j}(\boldsymbol{Z})$ is $N^{-2\alpha_j/(2\alpha_j+R)}$ (Stone, 1982) and we do not assume that any of the $\beta_{0,j}(\boldsymbol{Z})$'s are zero functions. In the absence of knowledge of the true α_j's, Theorem 3.1 shows that VCBART can estimate the varying coefficients $\boldsymbol{\beta}_0$ at nearly this rate, sacrificing only a logarithmic factor.

Theorem 3.1. *Under* (3.2)*, suppose we endow* $(\boldsymbol{\beta}, \sigma^2, \rho)$ *with the following VCBART prior: (a) For the BART priors on the varying coefficients* β_j*'s, the probability of splitting a tree node at depth* d *is* $q(d) = \gamma^d$ *for some* $1/N < \gamma < 1/2$ *and the splitting variables are chosen uniformly at random; (b)* $\sigma^2 \sim \lambda\nu\chi_\nu^{-2}$*, where* $\lambda, \nu > 0$*; and (c)* $\rho \sim Uniform(0, 1)$*. Assume that assumptions (A1)–(A3) hold. Then for some constant* $\widetilde{C} > 0$ *and* $r_N^2 = \log N \times \sum_{j=0}^p N^{-2\alpha_j/(2\alpha_j+R)}$*,*

$$\Pi\left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_N > \widetilde{C} r_N | \boldsymbol{Y}\right) \to 0, \tag{3.3}$$

in $\mathbb{P}_{\boldsymbol{\beta}_0}^{(N)}$*-probability as* $N, R \to \infty$*.*

The proof of Theorem 3.1 is given in Section S1 of the Supplementary Materials (Deshpande et al., 2024).

We now pause to highlight the novelties of Theorem 3.1. First, most existing theoretical results for BART assume independent observations. Theorem 3.1 appears to be the first to establish the near-optimality of BART under *non*-independent errors. Secondly, within the broader literature on varying coefficient models, there are many theoretical results when there is only a *single* modifier, i.e. $R = 1$ (Huang et al., 2004; Wang et al., 2008; Wang and Xia, 2009; Wei et al., 2011; Xue and Qu, 2012). However, to our knowledge, Zhou and Hooker (2022) are the only other authors who have derived asymptotic theory for varying coefficient models under *multivariate* effect modifiers with $R > 1$. Our Theorem 3.1 strengthens the result of Zhou and Hooker (2022) in two ways. First, Zhou and Hooker (2022) only proved estimation consistency for their (frequentist) treed varying coefficient model. We sharpen their result by deriving a nearly-minimax posterior *convergence rate*. Second, we allow the number of effect modifiers R to *diverge* with sample size in Assumption (A3), whereas the dimension R is fixed in Zhou and Hooker (2022).

We also highlight a few limitations of Theorem 3.1. First, the restriction that $0 < \alpha_j \le 1$ is an inevitable consequence of using hard decision trees (i.e., step functions) to estimate the varying coefficient functions (Ročková and Saha, 2019; Ročková and van der Pas, 2020). This fundamental limitation could be overcome if we replaced the hard decision rules with soft probabilistic decision rules (Linero and Yang, 2018). In this

case, the restriction $0 < \alpha_j \leq 1$ could be relaxed to $\alpha_j > 0$ for all $j = 0, \ldots, p$. Such an extension of VCBART is the subject of ongoing work (see Section 6).

We also assume that $O(p) = 1$ in Assumption (A3), which implies that the number of covariates p is bounded. This assumption was also made in Zhou and Hooker (2022) and is necessary to ensure that the posterior contraction rate $r_N^2 = \log N \times \sum_{j=0}^{p} N^{-2\alpha_j/(2\alpha_j+R)}$ decays to zero as $N \to \infty$. If we were to allow p to diverge with N, then we would also need to assume sparsity in the varying coefficients (i.e., $\beta_{0,j}(\boldsymbol{Z}) = 0$ for many $j \in \{0, 1, \ldots, p\}$). In this sparse set-up, let $S_0 \subset \{0, 1, \ldots, p\}$ denote the set of indices of the non-zero varying coefficients, with cardinality $|S_0| = s_0$. Under this sparse, high-dimensional varying coefficient model, the minimax estimation rate improves to $\widetilde{r}_N^2 = (s_0 \log p)/N + \sum_{j \in S_0} N^{-2\alpha_j/(2\alpha_j+R)}$ (Yang and Tokdar, 2015). In this case, we could also allow p and s_0 to diverge at the rates of $\log p = O(N^\xi), \xi \in (0, 1)$, and $s_0 = O(\log N)$ respectively (Yang and Tokdar, 2015). VCBART does not perform *selection* from the varying coefficient functions $\beta_j(\boldsymbol{Z})$'s, and we do *not* assume sparsity in the $\beta_{0,j}(\boldsymbol{Z})$'s. The result in Theorem 3.1 thus applies only to the "small p" case. Nevertheless, given the scarce literature on varying coefficient models with *multivariate* effect modifiers $(R > 1)$, Theorem 3.1 is still noteworthy and serves as an important first step towards further theoretical developments. We are currently working to extend VCBART to sparse, high-dimensional varying coefficient models where only a handful of the $\beta_j(\boldsymbol{Z})$'s are nonzero, and we will also investigate the posterior contraction rate in this context. This extension is briefly discussed in Section 6.

Finally, in Theorem 3.1, VCBART is only (nearly) minimax-optimal over the class of additive varying coefficient models where the individual functions have isotropic Hölder smoothness. This assumption may be restrictive. Recently, Jeong and Rockova (2023) and Ročková and Rousseau (2023) have established the near-minimax optimality of Bayesian treed regression under even broader function classes which allow for heterogeneous smoothness, discontinuities, and spatial inhomogeneities. Investigating the asymptotic properties of VCBART under substantially relaxed smoothness assumptions is an interesting problem for future work.

4 Simulation studies

We performed two simulation studies to understand VCBART's ability to (i) estimate covariate effect functions and predict future outcomes; (ii) identify which elements of $\boldsymbol{Z}$ modify the effects of which elements of X; and (iii) scale to large datasets. We generated data for these experiments from the model in (3.1) with $p = 5$ correlated covariates, $R = 20$ potential effect modifiers, $\sigma = 1$, and independent within-subject errors. The correlated covariates were drawn from a multivariate normal distribution with mean zero and a covariance matrix with entries $0.5^{|i-j|}$. The modifiers were drawn uniformly from the interval $[0, 1]$. To assess VCBART's ability to select relevant modifiers (i.e., identify the support of the $\beta_j(\boldsymbol{Z})$'s), we used the following covariate effect functions, which depend only five of the $R = 20$ modifiers, in all experiments:

$$\beta_0(\boldsymbol{z}) = 3z_1 + \big(2 - 5\mathbb{1}(z_2 > 0.5)\big) \sin(\pi z_1) - 2 \times \mathbb{1}(z_2 > 0.5)$$
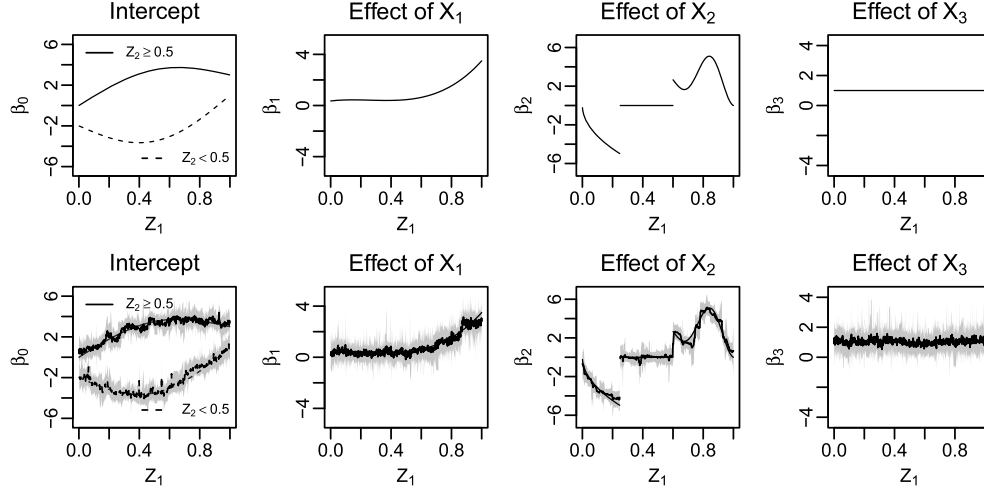
Figure 2: **Top row.** True functions $\beta_0, \ldots, \beta_3$. **Bottom row.** VCBART's posterior mean (dark line) and point-wise 95% credible interval (shaded) for each function.

$$\beta_1(\boldsymbol{z}) = \frac{\sin(2z_1 + 0.5)}{4z_1 + 1} + (2z_1 - 0.5)^4$$

$$\beta_2(\boldsymbol{z}) = \left(3 - 3z_1^2 \cos(6\pi z_1)\right) \times \mathbb{1}(z_1 > 0.6) - 10\sqrt{z_1} \times \mathbb{1}(z_1 < 0.25)$$

$$\beta_3(\boldsymbol{z}) = 1$$

$$\beta_4(\boldsymbol{z}) = 10 \sin(\pi z_1 z_2) + 20(z_3 - 0.5)^2 + 10z_4 + 5z_5$$

$$\beta_5(\boldsymbol{z}) = \exp\left\{\sin\left((0.9(z_1 + 0.48))^{10}\right)\right\} + z_2 z_3 + z_4.$$

The top row of Figure 2 shows the functions $\beta_0, \ldots, \beta_3$. The bottom row of Figure 2 superimposes the VCBART posterior mean and 95% credible intervals for $\beta_0, \ldots, \beta_3$ computed using a single dataset from our first experiment, which comprised four observations of $n = 250$ subjects for a total of $N = 1{,}000$ observed $(\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, y_{it})$'s. The posterior mean and uncertainty bands were computed after running four independent chains of VCBART for 2,000 iterations, discarding the first 1,000 samples of each as burn-in. We approximated each $\beta_j(\boldsymbol{Z})$ with an ensemble of $M = 50$ trees and set each $\tau_j = 0.5/\sqrt{M}$.

Although the covariate effect functions have very different shapes, VCBART recovered each remarkably well: the posterior means of each $\beta_j(\boldsymbol{Z})$ closely tracked the shape of the true functions and, for the most part, the true function values were within the shaded point-wise 95% credible intervals. We observed similar behavior for the functions $\beta_4(\boldsymbol{Z})$ and $\beta_5(\boldsymbol{Z})$ that depend on multiple effect modifiers (see Figure S4 in the Supplementary Materials (Deshpande et al., 2024)). For these data, the Gelman-Rubin $\hat{R}$ value computed using the samples of σ was around 1.2, which is somewhat higher than the heuristic convergence threshold of 1.1. Figure S1 of the Supplementary Materials (Deshpande et al., 2024) shows a representative traceplots for σ. While running each

chain for many more iterations produced smaller $\hat{R}$ values and ostensibly better-mixing chains, we found that additional iterations did not meaningfully change the quality of VCBART's predictions and uncertainty intervals (Figure S2 of the Supplementary Materials (Deshpande et al., 2024)). Consequently, for the remainder of this section, we report the performance of VCBART based on running four chains for 2,000 iterations.

Repeating this experiment 25 times — that is, fixing the $\boldsymbol{x}_{it}$'s and $\boldsymbol{z}_{it}$'s but simulating the y_{it}'s anew — the mean coverage of the 95% posterior credible intervals, averaged across all the β_j's, exceeded 99%. Further, averaging across all functions and replications, the mean sensitivity, specificity, precision and F1-score[1] for modifier selection with our median probability model were 0.85, 0.99, 0.95, and 0.90. Recalling that each $\beta_j(\boldsymbol{Z})$ depended on only a subset of the $R = 20$ modifiers, these results indicate excellent support recovery performance in the presence of several inactive variables. In fact, VCBART retains its excellent ability to identify the relevant modifiers when R is large and the effect modifiers are correlated and even in the presence of moderate-to-high correlation among the X_j's; see Sections S3.4 and S3.5 of the Supplementary Materials (Deshpande et al., 2024). Averaging across 25 replications, sequentially running four MCMC chains for 2,000 iterations each took under one minute on an M1 Mac.

We compared VCBART's ability to estimate covariate effects out-of-sample to those of the following methods: the standard linear model (`lm`); Li and Racine (2010)'s kernel smoothing (`KS`; implemented in Hayfield and Racine (2008)'s **np** package); Bürgin and Ritschard (2017)'s tree-based varying coefficient procedure (`TVC`); and Zhou and Hooker (2022)'s boosted tree procedure method (`BTVCM`). `BTVCM` is not a boosted version of `TVC`: the latter grows a single deep tree for each $\beta_j(\boldsymbol{Z})$ and then prunes each tree using cross-validation while the former sets a fixed tree depth in each round of boosting. By default, `KS` respectively perform cross-validation to learn kernel bandwidth parameters. We used the default cross-validation settings for `TVC` implemented in the R package **vcrpart** (Bürgin and Ritschard, 2017). The default implementation of `BTVCM`, on the other hand, does not tune the number of trees or learning rate; in our experiments, we ran the method with 200 trees and with the learning rate used in Zhou and Hooker (2022)'s experiments. Section S3.1 of the Supplementary Materials (Deshpande et al., 2024) contains additional details about the implementation of VCBART and these competitors.

We generated 25 different datasets comprising 250 training subjects and 25 testing subjects, each contributing four observations. That is, we repeatedly trained each method using 1,000 total simulated observations and evaluated performance using an additional 100 observations. As with Figure 2, we ran four chains VCBART for 2,000 iterations, discarding the first 1,000 as "burn-in," with $M = 50$ trees per ensemble and setting each $\tau_j = 0.5/\sqrt{M}$. Since the off-the-shelf implementations of `KS`, `TVC`, and `BTVCM` do not return standard errors of the estimated covariate effects, we formed 95% bootstrap confidence intervals for each $\beta_j(\boldsymbol{z})$ using 50 bootstrap resamples. We ran these experiments in a shared high-throughput computing cluster (Center for High Throughput Computing, 2006). Figures 3a and 3b compare the out-of-sample mean square error

---

[1]Letting TP, TN, FP, and FN be the numbers of true positive, true negative, false positive, and false negative identifications in the supports of the β_j's, sensitivity is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, specificity is $\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, precision is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and the F1 score is $2 \times \mathrm{TP}/(2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN})$.
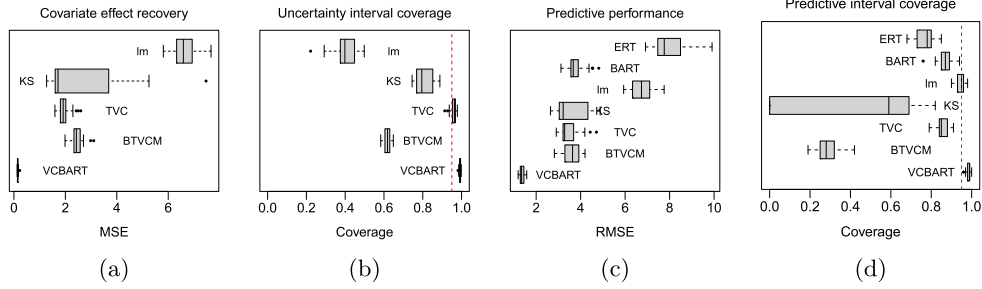
Figure 3: (a) Average mean square error for estimating evaluations $\beta_j(\boldsymbol{z})$. (b) Average coverage of 95% uncertainty intervals for evaluations $\beta_j(\boldsymbol{z})$. (c) Predictive root mean square error. (d) Coverage of 95% prediction intervals. All measures reported over 25 testing datasets.

for estimating $\beta_j(\boldsymbol{z})$ and uncertainty interval coverages, averaged across all six functions and over all 100 testing set observations. We report function-by-function performance in Section S3.2 of the Supplementary Materials (Deshpande et al., 2024).

In terms of estimating out-of-sample evaluations $\beta_j(\boldsymbol{z})$, `lm`, which assumes that each covariate effect is constant with respect to Z, performed the worst while VCBART appeared the best. VCBART additionally produced much more accurate predictions and better-calibrated uncertainty intervals than `BTVCM`, which also approximates each $\beta_j(\boldsymbol{Z})$ with a regression tree ensemble. The higher coverage of VCBART's uncertainty intervals was not a by-product of exceedingly wide intervals. In fact, VCBART's credible intervals for evaluations $\beta_j(\boldsymbol{z})$ tended to be shorter than the bootstrapped confidence intervals produced by `KS`, `BTVCM`, and `TVC`.

Sequentially running four VCBART chains took, on average, about three minutes on our shared cluster. Both `lm` and `BTVCM` were faster for these data, with `lm` running nearly instantaneously and `BTVCM` producing point estimates $\hat{\beta}_j(\boldsymbol{z})$ in about 90 seconds. That said, both `lm` and `BTVCM` produced substantially worse point estimates than VCBART. Although a carefully tuned `BTVCM` could yield better performance, we expect such tuning would take much longer. We additionally found that obtaining point estimate $\hat{\beta}_j(\boldsymbol{z})$ with `TVC` and `KS` respectively about 1 hour and 9.6 hours, on average. Sequentially running four chains of VCBART was substantially faster and produced better-calibrated uncertainty intervals than performing 50 bootstrap resamples for `BTVCM`, `KS`, and `TVC`. In summary, compared to existing methods for fitting varying coefficient models, VCBART was faster and produced more accurate predictions and better-calibrated uncertainty intervals off-the-shelf.

Intuitively, we should expect any method that reliably recovers individual out-of-sample evaluations $\beta_j(\boldsymbol{z})$ to yield highly accurate predictions of the outcome Y. And indeed, Figure 3c shows that the out-of-sample predictive root mean square error (RMSE) of VCBART and the aforementioned competitors generally track with their estimation error. Figure 3c additionally shows that VCBART achieved smaller predictive RMSE

than two fully non-parametric methods, `BART` (implemented in McCulloch et al. (2018)'s **BART** package) and extremely randomized trees (`ERT`; implemented in Wright and Ziegler (2017)'s **ranger** package). In our experiments, we allowed `BART` and `ERT` to use both X and Z to predict Y. Although `BART` and `ERT` are facially more flexible than the other competitors, their predictive performance was worse than the correctly-specified varying coefficient methods. Interestingly, however, `BART` and `ERT` produced predictive intervals with better coverage than the well-specified `BTVCM` and `KS` (Figure 3d). That said, VCBART tended to return predictive intervals with better coverage than both `BART` and `ERT`.

To better understand how well VCBART scales to larger datasets, we conducted a second simulation study. Similar to our first study, we generated 25 synthetic datasets with n training subjects and 250 testing subjects, each contributing four observations. We varied n between 25 and 12,500, allowing us to see how VCBART performed when trained on as few as 100 and as many as 50,000 total observations. For brevity, we summarize the main findings of that study here and defer more detailed results to Section S3.3 of the Supplementary Materials (Deshpande et al., 2024). Unsurprisingly, as N increased, VCBART's estimation and prediction error decreased and the frequentist coverage of its credible intervals and posterior predictive intervals remained high. In fact, VCBART produced uncertainty intervals with higher-than-nominal frequentist coverage when fit to as few as $N = 100$ total observations. We additionally found that the median probability model's operating characteristics generally improved a N increased. For instance, the F1 score was 0.68 when $N = 200$ and exceeded 0.85 for $N > 1000$. Averaging over 25 simulated datasets, sequentially running four VCBART chains for 2,000 iterations on an M1 Mac took around 55 seconds for $N = 1,000$ observations, 8.5 minutes for $N = 10,000$ observations, and around 48 minutes with $N = 50,000$ observations.

5   Real data examples

5.1   HRS cognition data

To investigate how associations between later-life cognition and socioeconomic position at multiple life stages may vary over time and across sociodemographic characteristics, we analyzed publicly available data from the Health and Retirement Study (HRS). The HRS is a nationally representative longitudinal survey of US adults over the age of 50 and their spouses of any age. Since 1992, the HRS has biennially assessed the economic, health, and social implications of aging through its core survey with response rates greater than 85% in each wave. The HRS is sponsored by the National Institute on Aging (NIA; U01AG009740) and is conducted by the University of Michigan. See Sonnega et al. (2014) for additional details about the HRS.

In each wave, the HRS consistently administered a battery of cognitive tests that include tasks like listening to a series of 10 words and recalling as many possible immediately and several minutes later. The HRS constructed a summary score that ranges from 0 to 35, with higher scores reflecting better cognitive functioning. In our analysis,

we use the total cognitive score as our outcome. Our covariate set includes measures of SEP at three distinct stages of life. The first is a composite childhood SEP index, developed and validated by Vable et al. (2017), which is based on the educational attainment and occupation of each subject's parents and their overall financial well-being in childhood. Higher scores on this index correspond to higher SEP. We additionally included indicators of high school and college completion as proxies of SEP in early adulthood. Finally, as measures of SEP in older adulthood, we used log-income and an indicator of labor force participation. Our modifiers included age in months, gender, race, the U.S. Census division of subject's birthplace, and an indicator of whether the subject identified as Hispanic.

We restricted our analysis to HRS subjects who (i) had complete covariate, modifier, and outcome data in at least four HRS waves and (ii) were dementia-free in the first wave with complete data. This resulted in analytic sample of $n = 10,812$ subjects who contributed a total of $N = 67,988$ person-years over the study period. In total, we had $p = 5$ covariates and $R = 5$ potential modifiers. Using the same hyperparameter setting as in Section 4 (i.e. $M = 50, \tau_j = 0.5/\sqrt{M}$), we simulated four Markov chains of 2,000 iterations each. After discarding the first 1,000 iterations of each chain as "burn-in", we obtained a total of 4,000 MCMC samples from the VCBART posterior. The $\hat{R}$ value based on the samples of σ was 1.01.

Based on these samples, for each combination of r and j, we computed the posterior probability that the modifier Z_r was used in at least one decision rule in the ensemble $\mathcal{E}_j$. We then computed the median probability model for each $\beta_j(z)$ by thresholding these probabilities at 0.5, as described in Section 3.2. The median probability model included gender for all $\beta_j(Z)$ and age for all coefficient functions except for the one associated with college completion. It also identified race as modifying the effects of childhood SEP, high school completion, and college completion but not later-life income or labor force participation. Similarly, it identified birth place as modifying the effects of just childhood SEP and high school completion. Taken together, these results suggest that, after adjusting for SEP in childhood and early-adulthood, difference in later-life cognitive score driven by differences in later-life SEP appear not to vary with respect to race and birthplace.

Further inspection revealed that the marginal posterior distributions of the estimated partial effects of later-life income for subjects in our dataset tended to concentrate within the interval $[-0.3, 0.9]$. For reference, the sample standard deviation of cognitive scores was about 4 points on a scale ranging from 0 to 35 points. This would suggest that, after adjusting for SEP in childhood and early adulthood, later-life income has a small predictive effect, if any, on later-life cognitive scores.

Figure 4 shows the posterior mean and point-wise 95% credible intervals of the predicted intercept and partial effects of childhood SEP and high school completion for two baseline individuals, one Black and one White. Both individuals are females born in the East North Central Census Division of the U.S., which includes the states of Illinois, Indiana, Michigan, Ohio, and Wisconsin. We can interpret the intercept as an average cognitive score for these baseline individuals if they (i) did not complete high
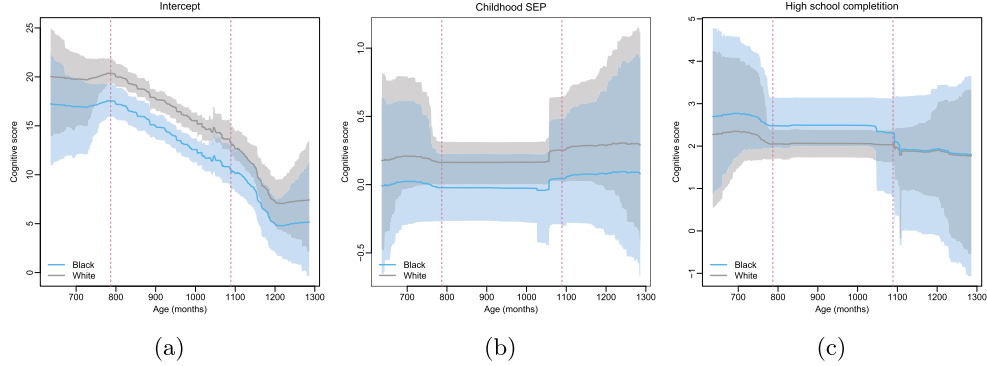
Figure 4: Posterior mean and pointwise 95% credible intervals of the intercept (a) and partial effects of childhood SEP (b) and high school completion (c) on later-life cognitive score as functions of age. Dashed vertical lines delineate the 2.5% and 97.5% quantiles of the observed ages in our dataset.

school or college and (ii) did not report any earned income or labor force participation in older adulthood.

The dashed vertical lines in each panel show the 2.5% and 97.5% quantiles of the distribution of ages in our dataset. Unsurprisingly, there is much more uncertainty about the estimated effects outside of this age range. In Figure 4a, we observe a steady decline in cognitive scores over time for both baseline individuals as well as a persistent gap between them, with the White individual having slightly higher baseline cognitive scores. Figure 4b shows that the posterior distributions of the partial effect of childhood SEP for these individuals is concentrated on small positive and negative effect sizes. In sharp contrast, however, we see in Figure 4c that high school completion is associated with higher cognitive scores for both baseline individuals. Interestingly, we see that the estimated increase in cognitive score associated with high school completion is slightly higher for the Black baseline individual. Figure S12 in the Supplementary Materials (Deshpande et al., 2024) plots the difference between these curves.

Recall that our median probability model detected temporal variation in the partial effect of high school completion. At least for the two baseline individuals considered in Figure 4c, that temporal variation appears to manifest mostly in the lower tail of the distribution. Between the ages of 65 and 90 years old, which covers approximately the middle 95% of observed ages in our dataset, the partial effect appears, for all practical purposes, constant over time.

5.2 Philadelphia crime data

The Philadelphia police department releases the latitude and longitude of every reported crime in the city at `opendataphilly.org`. In each of the $n = 384$ census tracts in the city (Figure 5a), we computed the yearly crime density, defined as number of crimes per

(a)   (b)
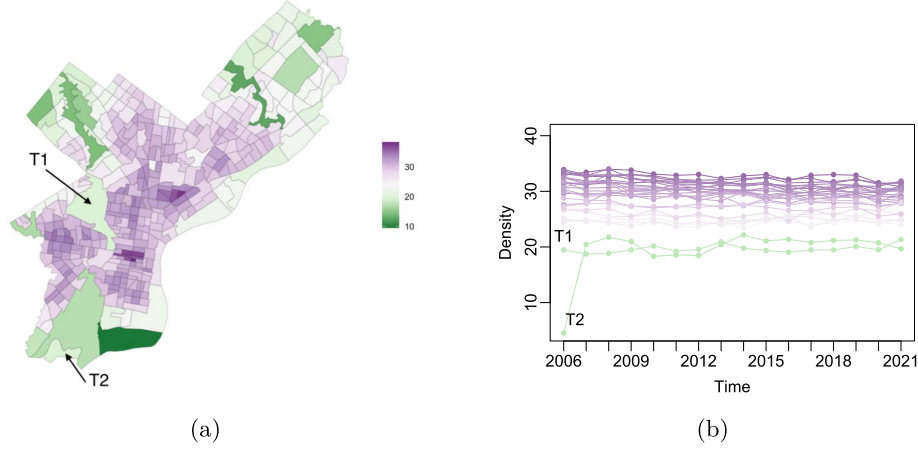
Figure 5: Census tracts of Philadelphia colored according to the averaged transformed crime density (a). Trajectories of crime densities for tract labelled "T1" in (a), its neighbors, and the tract labelled "T2" (b).

square mile, and applied an inverse hyperbolic sine transformation to counteract the considerable skewness. Let $y_{v,t}$ be the transformed crime density in census tract v at time t, with $t = 1$ corresponding to 2006 and $t = 16$ corresponding to 2021. Balocchi et al. (2023) modeled $y_{v,t} \sim \mathcal{N}(f^{(v)}(t), \sigma^2)$ and computed first-order approximations of each $f^{(v)}$. That is, they estimated tract-specific parameters $\beta_0^{(v)}$ and $\beta_1^{(v)}$ such that $f^{(v)}(t) \approx \beta_0^{(v)} + \beta_1^{(v)}\tilde{t}$, where $\tilde{t}$ is a centered and re-scaled version of the time index. Rather than separately estimate each tract's parameters, they assumed that the intercepts and slopes were spatially clustered. To estimate the latent clusterings, they ran several greedy searches across the space of pairs of census tract partitions.

Despite the reasonably accurate predictions obtained with their first-order approximation, a potential drawback of their analysis is the assumed monotonicity of crime over time. An arguably more realistic model would allow for crime potential increases and decreases within each census tract. Although it is conceptually easy to elaborate their model with higher-order terms, fitting such a model is computationally impractical. Specifically, extending their approach to find an order-d polynomial approximation involves searching over the product space of d partitions. The combinatorial vastness of this space renders efficient search all but infeasible.

We can instead view their model and higher-order elaborations thereof as varying coefficient models in which the coefficients vary across the vertices of a network determined by census tract adjacency (Figure S13 in the Supplementary Materials (Deshpande et al., 2024)). Specifically, for a positive integer d, we can approximate

$$f^{(v)}(t) \approx \beta_0(v) + \beta_1(v)\tilde{t} + \beta_2(v)\tilde{t}^2 + \cdots + \beta_d(v)\tilde{t}^d, \qquad (5.1)$$

where we have suggestively expressed the β_d's as functions of the census tract label v.

For a given approximation order d, fitting the model in (5.1) is relatively straightforward with VCBART. To encourage spatial smoothness in the $\beta_j(v)$'s, we used a version of the BART prior that recursively partitions the tracts into spatially contiguous clusters. This prior is identical to the one described in Section 3.1 except in how it draws the set $\mathcal{C}$ for decision rules involving the census tract label. To draw such $\mathcal{C}$'s, we first compute the set $\mathcal{A}$ of available census tracts and then form a network whose vertices are the tracts in $\mathcal{A}$ and whose edges encode spatial adjacency relationships between those tracts. Next, we sample a spanning tree of this network uniformly at random before deleting one edge at random, which partitions the network into two connected components. These connected components correspond to spatially contiguous clusters of census tracts and we set $\mathcal{C}$ to be the tracts in one of these clusters. See Deshpande (2023, §3.3) for more details about the prior.

We used VCBART to fit several versions of the model in (5.1) with different approximation orders $d \in \{1, \ldots, 4\}$. We also consider two within-tract error correlation structure for each order d; the first structure assumes independence across observations while the second assumes that the errors are equally correlated across time points within each tract (i.e. compound symmetry). To compare the performance of each version, we formed 50 random training-testing splits and computed the in- and out-of-sample prediction error for each version (Table 1). Each training split consisted of 90% of the available data and the testing observations were held out uniformly at random. We additionally fit a basic BART model using the implementation provided by **BART**. Running four MCMC chains for 2,000 iterations each for each VCBART model yielded Gelman-Rubin $\hat{R}$ values between 1.05 to 1.15. Increasing the total number of iterations to 5,000 per chain yielded $\hat{R}$ values between 1.01 and 1.08.

	RMSE (train)	RMSE (test)	Run time (hours)
$d = 1$, `ind`	0.60	0.69	4.55
$d = 2$, `ind`	0.53	0.65	7.61
$d = 3$, `ind`	0.51	0.64	10.44
$d = 4$, `ind`	**0.49**	**0.63**	14
$d = 1$, `cs`	0.63	0.71	4.68
$d = 2$, `cs`	0.56	0.67	8.35
$d = 3$, `cs`	0.54	0.66	11.01
$d = 4$, `cs`	0.51	0.66	12.90
`BART`	0.68	0.75	**2.4**

Table 1: In- and out-of-sample root mean square prediction error and run time (in hours) averaged over 50 replications in which 90% of the data were used for training and 10% for testing. Best performance in each column is bolded. `ind` stands for independent errors and `cs` stands for compound symmetry errors.

As we might expect, for both error structures, in-sample RMSE decreased with the approximation order d. The out-of-sample RMSE also decreased as we increased d. We further found that the compound symmetry model for the residual errors tended to produce worse in- and out-of-sample predictions. Compared to BART, all of our VCBART models impose rather strong functional form assumptions about the temporal

trend of crime rates. On this view, we might regard BART as being much more flexible than VCBART since it does not make any such restrictions. Nevertheless, we find that for these data, each of our VCBART models fits the data better than BART. For all methods considered, the 95% predictive intervals achieved coverages between 96% and 98% on both the training and testing data (results not shown).
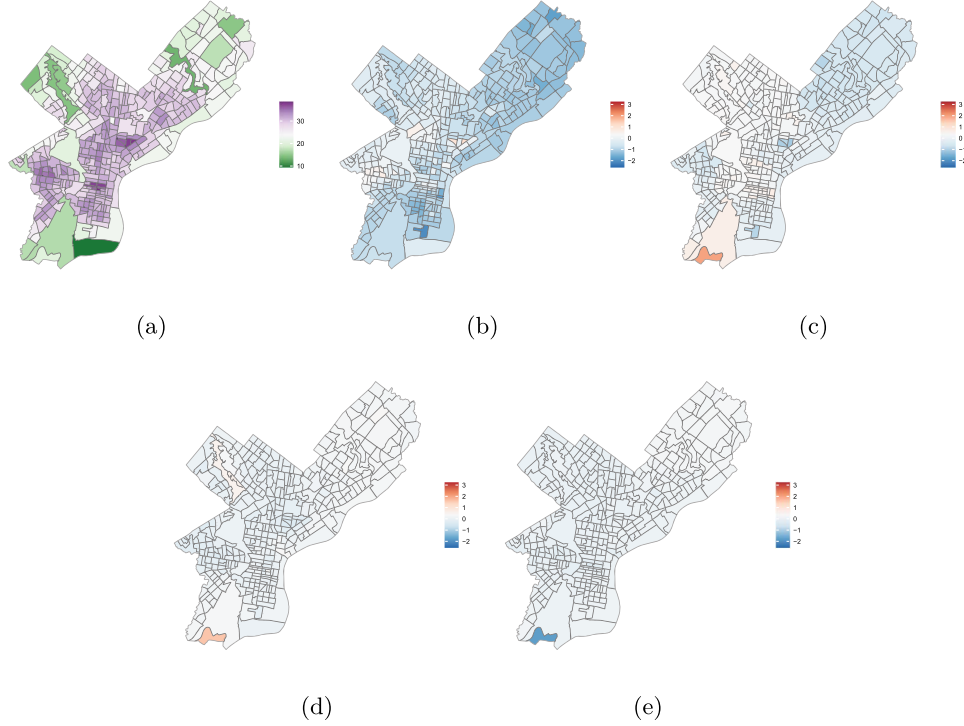

(a)   (b)   (c)

(d)   (e)

Figure 6: Posterior mean estimates of the varying coefficients in the best-fitting fourth-order approximation.

Figure 6 shows the posterior means of the coefficients from the best-fitting model, which was a fourth-order model. The posterior mean of the first-order term β_1 was negative in 354 out of the 384 census tracts (92%) and the posterior placed 95% probability on negative β_1 values in 293 census tracts (76%). This finding is consistent with the general observation that crime fell city-wide in the early-half of the 2010's. With one exception, the posterior distributions of β_3 and β_4 concentrated on very small values. This finding is not entirely unexpected, in light of the relatively flat trace plots in Figure 5b and for the other census tracts (not shown). The one exception is a tract near the south of the city (labelled "T2" in Figure 5), which experienced a dramatic increase in reported crimes from 2006 to 2007 and a somewhat oscillatory pattern over the next several years. So while a quadratic polynomial was sufficient to capture the temporal variation in the rest of the city, a higher-order approximation was needed for this particular census tract.

6 Discussion

We have introduced VCBART for fitting linear varying coefficient models that retain the interpretability of linear models while displaying predictive accuracy on par with more flexible, blackbox regression models. On simulated data, VCBART displayed covariate effect function recovery and predictive performance superior to existing state-of-the-art VC methods *without intensive problem-specific tuning or imposing any structural assumptions about the functions* $\beta_j(\boldsymbol{Z})$. It moreover returned coherent and generally well-calibrated uncertainty quantifications. Theorem 3.1 shows that, up to a logarithmic factor, the VCBART posterior concentrates at minimax optimal rate.

In our analysis of the HRS data, we did not explicitly adjust for potential selection biases and confounding. Indeed, it is possible that subjects with higher childhood SEP tended to be better educated and more well-off in later-life. For a more rigorous study of causal effects of SEP on cognitive trajectories, we could use inverse-probability weights as in Marden et al. (2017); indeed, incorporating observation-specific weights to the basic VCBART algorithm is relatively straightforward. Relatedly, the Bayesian causal forest (BCF) models of Hahn et al. (2020) and Woody et al. (2020) are special cases of our more general VCBART model with single covariate. Studying VCBART's ability to estimate the simultaneous effects of multiple, possibly continuous, treatments under suitable identifying assumptions is an exciting avenue for future development.

In our motivating applications, the number of covariates p was relatively small, compared to the number of total observation N. One can easily imagine, however, settings in which p is comparable to or larger than N. In such settings, running VCBART is challenging as the coefficient functions are no longer identified without further assumptions like sparsity. Under the assumption that only a handful of $\beta_j(\boldsymbol{Z})$'s are truly non-zero functions, we can estimate a sparse varying coefficient model by replacing the normal prior on tree jumps μ_ℓ with spike-and-slab priors. Doing so involves introducing an ensemble-specific indicator and, conditionally on the relevant indicator, modeling all jumps in an ensemble as drawn from, for instance, a mean-zero Gaussian distribution with either a very small or very large variance. Such model elaboration inspires a natural Gibbs sampler that iterates between updating these indicators and updating trees within each ensemble. Unfortunately, we would expect this Gibbs sampler to inherit the slow mixing of Gibbs samplers for spike-and-slab regression and for treed regression (Ronen et al., 2022; Kim and Ročková, 2023). A potentially more promising computational solution would combine existing fast mode-finding algorithms like EMVS (Ročková and George, 2014) for sparse regression and the accelerated BART and BCF algorithms (He et al., 2019; Krantsevich et al., 2023) for treed regression. Extending these accelerated algorithms to the general varying coefficient model setting is the subject of on-going work.

Because our implementation of VCBART utilizes hard decision trees, in which observations follow deterministic paths from the root to the leafs, it is unable to exactly recover smooth functions. As demonstrated by Linero (2022), however, one can implement VCBART using soft decision trees where, for each $\boldsymbol{z}$, the observations follow probabilistic paths from the root to the leafs (see, e.g., Linero and Yang, 2018, and references therein). Whereas a hard tree will output a single jump for a given $\boldsymbol{z}$, a soft tree

will output a weighted average of all jumps, thereby introducing additional smoothness. Extending our implementation of VCBART to support soft decision trees is the subject of ongoing work.

VCBART assumes homoscedastic, Gaussian residual errors and does not account for potential measurement errors of covariates. However, if these assumptions are violated, then the performance of VCBART may deteriorate. In the frequentist literature, there have been extensions to varying coefficient models that properly account for heteroscedasticity, skewed and/or heavy-tailed data, and measurement error (Shen et al., 2014; Xiong et al., 2023; Dong et al., 2022). VCBART can similarly be extended to these scenarios. For example, implementing a heteroskedastic version of VCBART in which the residual errors have variances of the form $\sigma^2(\boldsymbol{x}, \boldsymbol{z})$ is relatively straightforward. Indeed, following Pratola et al. (2019), one could express log-variance $\log \sigma^2(\boldsymbol{x}, \boldsymbol{z})$ with its own ensemble of regression trees. Then in each MCMC iteration, one would update the trees in the variance ensemble conditionally on the ensembles for the mean function (i.e., the $\beta_j(\boldsymbol{Z})'s$) and vice versa, just like the algorithm in Pratola et al. (2019). Following Um et al. (2023), we could also implement a version of VCBART that is more robust to outliers by replacing the multivariate Gaussian residual error in (3.1) with a multivariate skew-normal distribution.

While VCBART was shown to scale reasonably with sample size in Section 4, further improvements in scalability are desirable. Recently, Luo and Pratola (2023) introduced a formal Bayesian procedure for dividing data into individual "shards," fitting BART models to each shard, and combining predictions across shards. We anticipate that a similarly sharded VCBART would be capable of scaling to even larger datasets than those considered here. Finally, throughout this work we have assumed that the sets of covariates X and effect modifiers Z was fixed and known. While domain knowledge and theory often suggest natural choices of effect modifiers, we can readily envision scenarios in which it is not immediately clear whether a particular predictor should enter (1.1) as a covariate, modifier, or both. We leave automatic determination of the appropriate varying coefficient model specification to future work.

Supplementary Material

Supplementary Materials (DOI: 10.1214/24-BA1470SUPP; .pdf). Proofs of main results and results from hyperparameter sensitivity analysis and additional simulation studies

References

Aartsen, M. J., Cheval, B., Sieber, S., Van der Linden, B. W., Gabriel, R., Courvoisier, D. S., Guessous, I., Burton-Jeangros, C., Blane, D., Ihle, A., Kliegel, M., and Cullati,

S. (2019). "Advantaged socioeconomic conditions in childhood are associated with higher cognitive functioning but strong cognitive decline in older age." *Proceedings of the National Academy of Sciences*, 116(2): 5478–5486. 2

Bai, R., Boland, M. R., and Chen, Y. (2023). "Scalable high-dimensional Bayesian varying coefficient models with unknown within-subject covariance." *Journal of Machine Learning Research*, 24: 1–49. MR4664696. 4

Balocchi, C., Deshpande, S. K., George, E. I., and Jensen, S. T. (2023). "Crime in Philadelphia: Bayesian clustering with particle optimization." *Journal of the American Statistical Association*, 118(542): 818–829. MR4595458. doi: https://doi.org/10.1080/01621459.2022.2156348. 2, 3, 18

Balocchi, C., George, E. I., and Jensen, S. T. (2021). "Clustering areal units at multiple levels of resolution to model crime incidence in Philadelphia." *arXiv preprint arXiv:2112.02059.* 2

Balocchi, C. and Jensen, S. T. (2019). "Spatial modeling of trends in crime over time in Philadelphia." *Annals of Applied Statistics*, 13(4): 2235–2259. MR4037429. doi: https://doi.org/10.1214/19-aoas1280. 2

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *Annals of Statistics*, 32(3): 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 8

Bentley, J. L. (1975). "Multidimensional binary search trees used for associative searching." *Communications of the ACM*, 18(9): 509–517. 9

Bürgin, R. and Ritschard, G. (2015). "Tree-based varying coefficient regression for longitudinal ordinal responses." *Computational Statistics and Data Analysis*, 86: 65–80. MR3312738. doi: https://doi.org/10.1016/j.csda.2015.01.003. 4

Bürgin, R. and Ritschard, G. (2017). "Coefficient-wise tree-based varying coefficient regression with vcrpart." *Journal of Statistical Software*, 80(6): 1–33. 13

Center for High Throughput Computing (2006). "Center for High Throughput Computing." URL https://chtc.cs.wisc.edu/ 13

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics*, 4(1): 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 1, 5, 6, 8

Chipman, H. A., George, E. I., McCulloch, R. E., and Shively, T. S. (2022). "mBART: Multidimensional monotone BART." *Bayesian Analysis*, 17(2): 515–544. MR4483229. doi: https://doi.org/10.1214/21-ba1259. 5

Deshpande, S. K. (2023). "**flexBART**: Flexible Bayesian regression trees with categorical predictors." *arXiv preprint arXiv:2211.04459.* 8, 19

Deshpande, S. K., Bai, R., Balocchi, C., Starling, J. E., and Weiss, J. (2024). "Supplementary Material for "VCBART: Bayesian trees for varying coefficients"." *Bayesian Analysis.* doi: https://doi.org/10.1214/24-BA1470SUPP. 7, 8, 9, 10, 12, 13, 14, 15, 17, 18

Dong, H., Otsu, T., and Taylor, L. (2022). "Estimation of varying coefficient models with measurement error." *Journal of Econometrics*, 230(2): 388–415. MR4466730. doi: https://doi.org/10.1016/j.jeconom.2020.12.013. 22

Dupre, M. E. (2007). "Educational differences in age-related patterns of disease: Reconsidering the cumulative disadvantage and age-as-leveler hypotheses." *Journal of Health and Social Behavior*, 48(1): 1–15. 2

Fan, J. and Zhang, W. (2008). "Statistical methods with varying coefficient models." *Statistics and its Interface*, 1: 179–195. MR2425354. doi: https://doi.org/10.4310/SII.2008.v1.n1.a15. 4

Finley, A. O. and Banerjee, S. (2020). "Bayesian spatially varying coefficient models in the spBayes R package." *Environmental Modelling and Software*, 125. 4

Franco-Villoria, M., Ventrucci, M., and Rue, H. (2019). "A unified view on Bayesian varying coefficient models." *Electronic Journal of Statistics*, 13(2): 5334–5359. MR4047589. doi: https://doi.org/10.1214/19-EJS1653. 4

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). "Spatial modeling with spatially varying coefficient processes." *Journal of the American Statistical Association*, 98(462): 387–396. MR1995715. doi: https://doi.org/10.1198/016214503000170. 4

Greenfield, E. A. and Moorman, S. M. (2019). "Childhood socioeconomic status and later life cognition: Evidence from the Wisconsin Longitudinal Study." *Journal of Aging and Health*, 31(9): 1589–1615. 2

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2022). "Distributed Bayesian varying coefficient modeling using a Gaussian Process prior." *Journal of Machine Learning Research*, 23(84): 1–59. MR4576669. 4

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression models for causal inference: regularization, confounding, and heterogeneous effects." *Bayesian Analysis*, 15(3): 965–1056. MR4154846. doi: https://doi.org/10.1214/19-BA1195. 3, 5, 21

Hastie, T. and Tibshirani, R. (1993). "Varying-coefficient models." *Journal of the Royal Statistical Society Series B*, 55(4): 757–796. MR1229881. 4

Hayfield, T. and Racine, J. S. (2008). "Nonparametric econometrics: The np package." *Journal of Statistical Software*, 27(5). 13

He, J., Yalov, S., and Hahn, P. R. (2019). "XBART: Accelerated Bayesian Additive Regression Trees." In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 1130–1138. 21

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 5

Hill, J. L., Linero, A. R., and Murray, J. S. (2020). "Bayesian Additive Regression Trees: A review and look forward." *Annual Review of Statistics and its Applications*,

7(1): 251–278. MR4104193. doi: https://doi.org/10.1146/annurev-statistics-031219-041110.   5

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data." *Biometrika*, 85(4): 809–822. MR1666699. doi: https://doi.org/10.1093/biomet/85.4.809.   4

Huang, J. Z. and Shen, H. (2004). "Functional coefficient regression models for nonlinear time series: A polynomial splines approach." *Scandinavian Journal of Statistics*, 31: 515–534. MR2101537. doi: https://doi.org/10.1111/j.1467-9469.2004.00404.x.   8

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). "Varying-coefficient models and basis function approximations for the analysis of repeated measurements." *Biometrika*, 89(1): 111–128. MR1888349. doi: https://doi.org/10.1093/biomet/89.1.111.   4

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). "Polynomial spline estimation and inference for varying coefficient models with longitudinal data." *Statistica Sinica*, 14: 763–788. MR2087972.   10

Jeong, S. and Rockova, V. (2023). "The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension." *Journal of Machine Learning Research*, 24(337): 1–65. MR4690286. doi: https://doi.org/10.5540/tcam.2023.024.02.00337.   11

Kim, J. and Ročková, V. (2023). "On mixing rates for Bayesian CART." *arXiv preprint arXiv:2306.00126*.   21

Krantsevich, N., He, J., and Hahn, P. R. (2023). "Stochastic tree ensembles for estimating heterogeneous effects." In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 6120–6131.   21

Lee, K., Lee, Y. K., Park, B. U., and Yang, S. J. (2018). "Time-dynamic varying coefficient models for longitudinal data." *Computational Statistics and Data Analysis*, 123: 50–65. MR3777085. doi: https://doi.org/10.1016/j.csda.2018.01.016.   4

Li, Q. and Racine, J. S. (2010). "Smooth varying-coefficient estimation and inference for qualitative and quantitative data." *Econometric Theory*, 26(6): 1607–1637. MR2738011. doi: https://doi.org/10.1017/S0266466609990739.   4, 13

Linero, A. R. (2018). "Bayesian regression trees for high-dimensional prediction and variable selection." *Journal of the American Statistical Association*, 113(522): 626–636. MR3832214. doi: https://doi.org/10.1080/01621459.2016.1264957.   7

Linero, A. R. (2022). "SoftBart: Soft Bayesian Additive Regression Trees." *arXiv preprint arXiv:2210.16375*.   21

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2020). "Semiparametric mixed-scale models using shared Bayesian forests." *Biometrics*, 76(1): 131–144. MR4098550. doi: https://doi.org/10.1111/biom.13107.   5

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to

smoothness and sparsity." *Journal of the Royal Statistical Society Series B*, 80(5): 1087–1110. MR3874311. doi: https://doi.org/10.1111/rssb.12293.   5, 10, 21

Luo, H. and Pratola, M. T. (2023). "Sharded Bayesian Additive Regression Trees." *arXiv preprint arXiv:2306.0036v1*. MR4336676.   22

Luo, Y. and Waite, L. J. (2005). "The impacts of childhood and adult SES on physical, mental, and cognitive well-being in later life." *The Journals of Gerontology: Series B*, 60(2): S93–S101.   2

Lyu, J. and Burr, J. (2016). "Socioeconomic status across the life course and cognitive function among older adults: An examination of the latency, pathways, and accumulate hypotheses." *Journal of Aging Health*, 28(1): 40–67.   2

Marden, J., Tchetgen Tchetgen, E. J., Kawachi, I., and Glymour, M. (2017). "Contribution of socioeconomic status and 3 life-course periods to late-life memory function and decline: early and late predictors of dementia risk." *American Journal of Epidemiology*, 186(7): 805–814.   2, 21

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C., and Pratola, M. (2018). ***BART**: Bayesian Additive Regression Trees*. R package version 2.1.   15

Murray, J. S. (2021). "Log-linear Bayesian Additive Regression Trees for categorical and count responses." *Journal of the American Statistical Association*, 116(534): 756–769. MR4270022. doi: https://doi.org/10.1080/01621459.2020.1813587.   5

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2019). "Heteroscedastic BART using multiplicative regression trees." *Journal of Computational and Graphical Statistics*. MR4116052. doi: https://doi.org/10.1080/10618600.2019.1677243.   22

Ročková, V. and Rousseau, J. (2023). "Ideal Bayesian spatial adaptation." *Journal of the American Statistical Association*. MR4797924. doi: https://doi.org/10.1080/01621459.2023.2241705.   11

Ronen, O., Saarinen, T., Tan, Y. S., Duncan, J., and Yu, B. (2022). "A mixing time lower bound for a simplified version of BART." *arXiv preprint arXiv:2210.09352v1*.   21

Ročková, V. and George, E. I. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109(506): 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223.   21

Ročková, V. and Saha, E. (2019). "On theory for BART." In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2839–2848.   5, 9, 10

Ročková, V. and van der Pas, S. (2020). "Posterior concentration for Bayesian regression trees and forests." *Annals of Statistics*, 48(4): 2108–2131. MR4134788. doi: https://doi.org/10.1214/19-AOS1879.   5, 9, 10

Shen, S.-L., Cui, J.-L., Mei, C.-L., and Wang, C.-W. (2014). "Estimation and inference of semi-varying coefficient models with heteroscedastic errors." *Journal of Multivariate*

Analysis, 124: 70–93. MR3147312. doi: `https://doi.org/10.1016/j.jmva.2013.10.010`. 22

Sonnega, A., Faul, J., Ofstedal, M., Langa, K. M., Phillips, J. W., and Weird, D. R. (2014). "Cohort profile: The Health and Retirement Study (HRS)." *International Journal of Epidemiology*, 43(2): 576–585. 15

Sparapani, R., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). "Nonparametric survival analysis using Bayesian Additive Regression Trees." *Statistics in Medicine*, 35(16): 2741–2753. MR3513715. doi: `https://doi.org/10.1002/sim.6893`. 5

Starling, J. E., Aiken, C. E., Murray, J. S., Nakimuli, A., and Scott, J. G. (2019). "Monotone function estimation in the presence of extreme data coarsening: Analysis of preeclampsia and birth weight in urban Uganda." *arXiv preprint arXiv:1912.06946*. 5

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R., and Scott, J. G. (2020). "BART with targeted smoothing: An analysis of patient-specific stillbirth risk." *Annals of Applied Statistics*, 14(1): 28–50. MR4085082. doi: `https://doi.org/10.1214/19-AOAS1268`. 5

Stone, C. J. (1982). "Optimal global rates of convergence for nonparametric regression." *The Annals of Statistics*, 10(4): 1040–1053. MR0673642. 10

Tan, Y. V. and Roy, J. (2019). "Bayesian Additive Regression Trees and the General BART model." *Statistics in Medicine*, 38(25): 5048–5069. MR4022845. doi: `https://doi.org/10.1002/sim.8347`. 5

Tibshirani, R. and Friedman, J. (2020). "A pliable lasso." *Journal of Computational and Graphical Statistics*, 29(1): 215–225. MR4085876. doi: `https://doi.org/10.1080/10618600.2019.1648271`. 4

Um, S., Linero, A. R., Sinha, D., and Bandyopadhyay, D. (2023). "Bayesian Additive Regression Trees for multivariate skewed responses." *Statistics in Medicine*, 42(3): 246–263. MR4537847. doi: `https://doi.org/10.1002/sim.9613`. 22

Vable, A. M., Gilsanz, P., Nguyen, T. T., Kawachi, I., and Glymour, M. M. (2017). "Validation of a theoretically motivated approach to measuring childhood socioeconomic circumstances in the Health and Retirement Study." *PLoS ONE*, 12(10): e0185898. 16

Wang, H. and Xia, Y. (2009). "Shrinkage estimation for the varying coefficient model." *Journal of the American Statistical Association*, 104(486): 747–757. MR2541592. doi: `https://doi.org/10.1198/jasa.2009.0138`. 4, 10

Wang, J. C. and Hastie, T. (2012). "Boosted varying-coefficient regression models for product demand prediction." *Journal of Computational and Graphical Statistics*, 23(2): 361–382. MR3215815. doi: `https://doi.org/10.1080/10618600.2013.778777`. 4

Wang, L., Li, H., and Huang, J. Z. (2008). "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements." *Journal of the American*

Statistical Association, 103(484): 1556–1569. MR2504204. doi: `https://doi.org/10.1198/016214508000000788`.   4, 10

Wei, F., Huang, J., and Li, H. (2011). "Variable selection and estimation in high-dimensional varying-coefficient models." *Statistica Sinica*, 21: 1515–1540. MR2895107. doi: `https://doi.org/10.5705/ss.2009.316`.   4, 10

Woody, S., Carvalho, C. M., Hahn, P. R., and Murray, J. S. (2020). "Estimating heterogeneous effects of continuous exposures using Bayesian tree ensembles: Revising the impact of abortion rates on crime." *arXiv preprint arXiv:2007.09845v1*.   21

Wright, M. N. and Ziegler, A. (2017). "ranger: A fast implementation of random forests for high dimensional data in C++ and R." *Journal of Statistical Software*, 77(1): 1–17. 15

Wu, C. O. and Chiang, C.-T. (2000). "Kernel smoothing on varying coefficient models with longitudinal dependent variable." *Statistica Sinica*, 10: 433–456. MR1769751.   4

Xiong, W., Tian, M., Tang, M., and Pan, H. (2023). "Robust and sparse learning of varying coefficient models with high-dimensional features." *Journal of Applied Statistics*, 50(16): 3312–3336. MR4665314. doi: `https://doi.org/10.1080/02664763.2022.2109129`.   22

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). "Sequential BART for imputation of missing covariates." *Biostatistics*, 17(3): 589–602. MR3603956. doi: `https://doi.org/10.1093/biostatistics/kxw009`.   5

Xue, L. and Qu, A. (2012). "Variable selection in high-dimensional varying-coefficient models with global optimality." *Journal of Machine Learning Research*, 13(1): 1973–1998. MR2956349.   10

Yang, Y. and Tokdar, S. T. (2015). "Minimax-optimal nonparametric regression in high dimensions." *The Annals of Statistics*, 43(2): 652–674. MR3319139. doi: `https://doi.org/10.1214/14-AOS1289`.   11

Zhang, Z., Liu, H., and Choi, S.-w. (2020). "Early-life socioeconomic status, adolescent cognitive ability, and cognition in late midlife: Evidence from the Wisconsin Longitudinal Study." *Social Science & Medicine*, 244: 112575.   2

Zhou, Y. and Hooker, G. (2022). "Decision tree boosted varying coefficient models." *Data Mining and Knowledge Discovery*, 36: 2237–2271. MR4510524. doi: `https://doi.org/10.1007/s10618-022-00863-y`.   4, 9, 10, 11, 13