Estimating PATE under positivity violations: SBART+SPL for high-dimensional covariates

Lennard Maßmann LENNARD.MASSMANN@UNI-DUE.DE
Chair of Econometrics, Faculty of Business Administration and Economics, University of Duisburg-Essen, Universitätsstraße 12, 45117 Essen, Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

The positivity assumption is a fundamental requirement for causal inference in the potential outcomes framework, ensuring that all individuals have a positive probability of receiving each treatment option. However, real-world datasets often violate this assumption, particularly in regions of weak overlap where one treatment group is underrepresented or entirely absent for certain combinations of confounding variables. Traditional approaches, such as trimming and weighting, address these violations but typically modify the target population, potentially introducing bias. The Bayesian Additive Regression Trees with Spline Models (BART+SPL) has been proposed as a solution to this issue. This approach combines Bayesian Additive Regression Trees (BART) for imputation in regions of overlap with spline models (SPL) to extrapolate into regions of weak overlap, thereby preserving the initial target population. While delivering precise results when considering low-dimensional covariates, the performance of BART+SPL is compromised under high-dimensional covariates. To address this limitation, we propose SBART+SPL, an extension of the BART+SPL framework that integrates SoftBART into the estimation procedure. SoftBART generalizes BART by implementing smooth decision rules and sparsity-inducing splitting probabilities. A simulation study demonstrates that SBART+SPL yields better precision and improved coverage compared to BART+SPL when estimating population average treatment effects (PATE) in the presence of high-dimensional covariates and violations of the positivity assumption. The applicability of SBART+SPL is illustrated by re-analyzing an empirical study that evaluates the impact of exposure to natural gas compressor stations on cancer mortality rates across U.S. counties.

Keywords: Bayesian causal inference, regression trees, overlap, extrapolation

1. Introduction

One of the crucial assumptions to draw causal inference from observational data when using the potential outcome framework is the positivity assumption (Rubin, 2005; Hernán and Robins, 2020; Li et al., 2023). The assumption postulates that every individual within the considered population possesses a positive probability of receiving either treatment status. Oftentimes, the similar concept of overlap is used to justify the positivity assumption in non-parametric treatment effect estimation by comparing confounder distributions for both treatment groups: the exposure group and the control group in the case of a binary treatment. Many real-world datasets suffer from large regions of non-overlap, which constitutes a violation of the positivity assumption. Non-overlapping confounder distributions emerge when, at random, no or just a small number of individuals belonging to one treatment group are observed in a specific confounder region (Westreich and Cole, 2010). The majority of methods to tackle this issue are based on specific modeling assumptions. Usual approaches, like trimming (removing observed data in non-overlap regions) or weighting (reducing the influence

of non-overlap observations by decreasing their weights), change the underlying population considered by the initial estimand. For instance, if the initial estimand was the average treatment effect (ATE), those methods are only able to identify the ATE for the trimmed or re-weighted population. Furthermore, approaches like inverse probability weighting can lead to weight instability in regions of non-overlap, as the estimated probabilities may be close to zero (Crump et al., 2009; Stürmer et al., 2010; Li et al., 2018; Zhu et al., 2021).

Nethery et al. (2019) introduce BART+SPL, a combination of Bayesian Additive Regression Trees (BART) and a spline model (SPL), as an approach to differentiate between observations within overlap and non-overlap regions based on estimated propensity scores. BART+SPL reliably estimates the population average treatment effect (PATE) by taking into account both of these regions. It is characterized by substantially lower model dependence as well as suitable uncertainty quantification in the non-overlap region via a new Bayesian two-stage procedure. In the imputation phase, BART (Chipman et al., 2010) is used to estimate individual treatment effects in the overlap region. In the subsequent smoothing phase, more model dependence is introduced by extrapolating unit-level effects from the overlap region into the non-overlap region, relying on a flexible spline model (Wahba, 1990). The individual treatment effects of both regions are then aggregated into the final PATE estimate. Simulation studies in Nethery et al. (2019) demonstrate desirable results when faced with lower-dimensional data and different degrees of non-overlap. However, when many covariates are irrelevant to the potential outcomes, BART+SPL suffers from increasing bias and severe undercoverage, making it ill-suited for high-dimensional settings.

The contribution of this paper revolves around this issue and proposes a new method that builds upon the framework of BART+SPL. First, instead of using BART for the imputation stage, the SoftBART algorithm introduced by Linero and Yang (2018) is used, as it has been shown that covariate dimensions can increase nearly exponentially with the sample size for SoftBART. SoftBART is a generalization of BART that uses smooth decision rules instead of hard ones and sparsity-inducing splitting probabilities instead of uniformly distributed ones. This enhances the ability to model smoothness in the data-generating process, improves uncertainty quantification, and supports the selection of relevant covariates by shrinking irrelevant covariates. Second, SoftBART's ability to capture increased uncertainty and reduce bias when extrapolating beyond the overlap region eliminates the need for BART+SPL's unidentifiable variance inflation parameter (Nethery et al., 2019). SBART+SPL can therefore be viewed as extending BART+SPL to high-dimensional covariates when estimating the PATE under overlap violations. This generalization is particularly relevant for the empirical application in Nethery et al. (2019), which involves many covariates.

Related work proposes different Bayesian modeling approaches to estimate PATE while retaining the initial target estimand (Gutman and Rubin, 2015; Li et al., 2018, 2019; Nethery et al., 2019; Zhu et al., 2023; Wang et al., 2024). Gutman and Rubin (2015) develop Multiple Imputation with Two Subclassification Splines (MITSS) to estimate average treatment effects with multiple covariates. MITSS extends the method in Gutman and Rubin (2013), which only considers binary outcomes and one covariate for treatment effect estimation. Gutman and Rubin (2013) and Gutman and Rubin (2015) partition observations into subclasses of units with similar covariate values and estimate the PATE by averaging across estimates within these subclasses. Subclasses are related to each other by placing the knots of two regression splines (Wahba, 1990) at the boundaries of each subclass. The splines are estimated separately for each treatment group considering the distributions of potential outcomes conditional on covariates, respectively. Compared to Gutman and Rubin (2013), MITSS in Gutman and Rubin (2015) allows to analyze continuous outcomes and many covariates.

Their simulation study implies that MITSS is generally a valid and accurate approach, regardless of whether covariates are scalar or multivariate. However, when some units have non-overlapping covariate values across groups, MITSS relies on splines to implicitly extrapolate to regions without observed data for a given treatment group and introduces greater bias and improper coverage rates. Zhu et al. (2023) develop a Bayesian model based on non-parametric Gaussian Process (GP) priors. The method takes into account the full covariate space and does not need to differentiate between regions of overlap a-priori. Instead, they incorporate the amount of non-overlap into the GP itself and extrapolate with the covariate kernel of the GP. Wang et al. (2024) apply a local extrapolation method to the Accelerated Bayesian Causal Forest (XBCF) of Krantsevich et al. (2023) by integrating GP into XBCF's leaf nodes, creating a model termed XBCF-GP. This hybrid approach allows for more accurate predictions and better uncertainty quantification for data points outside the training range. Inference on treatment effects is illustrated using simulation data with extreme non-overlap regions wherein either only exposed or unexposed individuals are observed (Wang et al., 2024).

The paper is structured as follows: Section 2 introduces necessary causal inference notation, defines the assumptions of the potential outcome framework and discusses different notions when contrasting between the region of overlap and non-overlap. Section 3 contrasts the conventional BART algorithm of Chipman et al. (2010) with the SoftBART algorithm of Linero and Yang (2018) and describes the modified two-stage procedure of SBART+SPL based on the BART+SPL algorithm of Nethery et al. (2019). The simulation study focuses on two high-dimensional covariate setups (Nethery et al., 2019; Krantsevich et al., 2023; Wang et al., 2024) and is presented in Section 4. Section 5 compares SBART+SPL to BART+SPL and baseline methods in an empirical analysis of the effect of exposure to natural gas compressor stations on mortality rates in U.S. mid-western counties (Mokdad et al., 2017; Nethery et al., 2019).

2. Potential Outcomes and Region of Overlap

Let Y_i^{obs} be the observed continuous outcome of individual i, with individuals $i = 1, ..., n$. The binary treatment status of individual i is $D_i = d \in \{0, 1\}$ and let $\boldsymbol{X}_i$ be the $\mathcal{P}$-dimensional vector of confounding values that have been observed for individual i. Consequently, $\boldsymbol{X}$ is defined as the $n \times \mathcal{P}$-dimensional matrix of confounders for all individuals. Using the Stable Unit Treatment Value Assumption (SUTVA), the potential outcome notation is introduced by interpreting the potential outcomes $Y_i(1)$ and $Y_i(0)$ as the values that would have been realized had one observed either treatment status $D_i = 1$ or $D_i = 0$ for individual i. Implicitly, one states with SUTVA that only one potential outcome is realized for individual i such that the non-realized potential outcome is a missing data point. Consequently, one has $Y_i^{obs} = D_i Y_i(1) + (1 - D_i) Y_i(0)$ and $Y_i^{mis} = (1 - D_i) Y_i(1) + D_i Y_i(0)$ with the latter being the missing potential outcome of individual i. That is, one is faced with the fundamental problem of causal inference by having two potential outcomes $Y_i(0)$ and $Y_i(1)$ but only observing one realization Y_i^{obs} of $(Y_i(0), Y_i(1))$ (Rubin, 2005; Hernán and Robins, 2020; Li et al., 2023). We follow Li et al. (2023) and define the individual (τ_i), conditional average $(\tau(\boldsymbol{X}_i))$, and population average (τ_P) treatment effects as follows:

$$\tau_i \coloneqq Y_i(1) - Y_i(0), \quad \tau\left(\boldsymbol{X}_i\right) \coloneqq \mathbb{E}[\tau_i | \boldsymbol{X}_i], \quad \tau_P \coloneqq \mathbb{E}_{\boldsymbol{X}}[\tau(\boldsymbol{X}_i)]. \tag{1}$$

To identify τ_P with observational data, usually two additional assumptions next to the above-stated SUTVA are invoked (Li et al., 2023). Unconfoundedness assumes that the treatment assignment is approximately randomized conditional on a sufficiently informative $\boldsymbol{X}_i$, mimicking a completely

randomized controlled trial, such that

$$(Y_i(1), Y_i(0)) \perp D_i | \boldsymbol{X}_i. \tag{2}$$

The positivity assumption postulates that every unit in the considered population possesses a non-zero probability of being assigned to both treatment groups by

$$0 < Pr(D_i = 1 | \boldsymbol{X}_i) < 1, \tag{3}$$

with $p.sc(\boldsymbol{X}_i) = Pr(D_i = 1|\boldsymbol{X}_i)$ being the propensity score. Unconfoundedness in (2) and positivity in (3) allow us to view the whole dataset intuitively as a collection of many small $\boldsymbol{X}_i$-indexed randomized trials. Unconfoundedness ensures conditionally exogenous treatment assignment while positivity ensures that randomization actually occurs in the data. Under (2), we receive for the population average treatment effect,

$$\tau_P = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[Y_i(1) - Y_i(0)|\boldsymbol{X}_i]] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[Y^{obs}|\boldsymbol{X}_i, D_i = 1] - \mathbb{E}_{\boldsymbol{X}}[Y^{obs}|\boldsymbol{X}_i, D_i = 0]]. \tag{4}$$

To identify the conditional expectations in (4) non-parametrically, assumption (3) is needed. Note that there exists a tension between the assumptions of unconfoundedness and positivity. As the number of covariates increases, the unconfoundedness assumption becomes more plausible but with a rising number of covariates it is also more likely to have sparse data for one or both treatment groups in certain regions of $\boldsymbol{X}$ (Li et al., 2023).

The positivity assumption in (3) can be violated both structurally and randomly. Structural positivity is violated if a unit of interest is not able to obtain treatment such that enlarging the sample size does not alleviate the issue (D'Amour et al., 2021) and will not be the focus of this paper. In contrast, a random positivity violation allows treatment assignment to the unit of interest theoretically, but one cannot (or only rarely) observe this type of treatment empirically within the analyzed data. Increasing the sample size n may mitigate this type of positivity violation. However, increasing the number of covariates $\mathcal{P}$ in $\boldsymbol{X}$ makes the occurrence of a random overlap violation more probable as the likelihood of observing similar units of the exposure and control group for a given covariate combination decreases.

This paper examines the issue of random positivity violations within the potential outcome framework used for causal inference in observational studies for increasing $\mathcal{P}$. Random positivity is closely related to the concept of overlap where we define a region of overlap as a region with covariate values being present in the exposure as well as in the control group. Hence, a region of non-overlap is characterized by covariate values for units that do not exist in a sufficient amount in both groups. Random positivity can be evaluated by assessing overlap through the inspection of propensity score estimates $\widehat{p.sc}(\boldsymbol{X}_i)$ (Zhu et al., 2021). However, with $\widehat{p.sc}(\boldsymbol{X}_i) \approx 0$ the likelihood of observing a unit with these covariate values in the exposure group is low. Similarly, with $\widehat{p.sc}(\boldsymbol{X}_i) \approx 1$ the likelihood of observing a unit with these covariate values in the control group is low. For weighting methods like standard inverse probability weighting (Hernán and Robins, 2006; Austin, 2011), extreme propensity score estimates might raise the issue of almost infinite weights by dividing by $\widehat{p.sc}(\boldsymbol{X}_i)$. Furthermore, trimming and weighting both relocate the inferential target from the initial population at hand to the trimmed or matched population to circumvent possible overlap issues (Zhu et al., 2021).

For BART+SPL, Nethery et al. (2019) extrapolate from pre-specified regions of overlap to the regions of non-overlap. In contrast to trimming or matching, extrapolation based on the distinction

between those regions allows statements about treatment effects concerning the original study sample, as the initial inferential target population is not relocated. We adopt the definitions of the region of overlap O and non-overlap $O^{\neg}$ from Nethery et al. (2019). A unit i with estimated propensity score $\widehat{p.sc}(\boldsymbol{X}_i)$ belongs to O if, in both treatment groups, more than $b_O = 7$ units have estimated propensity scores sufficiently close to $\widehat{p.sc}(\boldsymbol{X}_i)$ such that the range of the set of these scores is less than $a_O = 0.1$. Full computational details are given in Appendix A.

To illustrate, suppose unit 1 has $\widehat{p.sc}(\boldsymbol{X}_1) = 0.20$ with many estimated propensity scores of control units nearby, yet fewer than b_O estimated propensity scores of exposed units fall within a range of a_O; then $\widehat{p.sc}(\boldsymbol{X}_1) \in O^{\neg}$. Conversely, if unit 2 has $\widehat{p.sc}(\boldsymbol{X}_2) = 0.45$ and in both groups more than b_O neighbors can be found with a range below a_O, then $\widehat{p.sc}(\boldsymbol{X}_2) \in O$. This shows that overlap is assessed locally and separately within each group, permitting non-overlapping regions anywhere in the estimated propensity score distribution and not only in the tails, as under the trimming rule of Crump et al. (2009). The local criterion may also be more tractable under high-dimensional $\boldsymbol{X}$ than the approach of Hill and Su (2013). We index units with estimated propensity scores belonging to O by $o = 1, \ldots, n_O$ and those in $O^{\neg}$ by $o^{\neg} = 1, \ldots, n_{O^{\neg}}$, such that $n = n_O + n_{O^{\neg}}$. Given these regions, BART+SPL in Nethery et al. (2019) and SBART+SPL, as its proposed extension in Section 3.1, take observations in the overlap region to extrapolate treatment effect patterns into the non-overlap region using either BART or SoftBART in combination with a spline model. The following Section 3 motivates the usage of SoftBART as a suitable extension of BART before contrasting SBART+SPL with BART+SPL.

3. SBART+SPL for PATE Estimation

Consider the semiparametric Gaussian regression problem in Linero and Yang (2018) by

$$Y_i = f_0(\boldsymbol{X}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where f_0 is an unknown regression function predicting outcomes Y_i given covariates $\boldsymbol{X}_i \in \mathbb{R}^{\mathcal{P}}$, and ϵ_i denotes a Gaussian error term with variance σ^2. The BART algorithm of Chipman et al. (1998, 2010) models f_0 as a sum of J regression trees $f(\boldsymbol{X}_i = \boldsymbol{x}) = \sum_{j=1}^{J} g(\boldsymbol{X}_i = \boldsymbol{x}; \mathcal{T}_j, \mathcal{M}_j)$, where each single-tree contribution $g(\cdot)$ is defined by a tree structure $\mathcal{T}_j$, leaf node parameters $\mathcal{M}_j = (\mu_{j1}, \ldots, \mu_{jL_j})$, and hard decision rules that partition the covariate space into axis-aligned rectangular regions. At each branch node b, BART selects a predictor index $p_b \in \{1, \ldots, \mathcal{P}\}$ and a cutpoint C_b, then routes observations according to whether $x_{p_b} \leq C_b$ where x_{p_b} denotes the p_b-th element in $\boldsymbol{x}$. Standard notation, prior specifications, and posterior inference for BART are given in Appendix B.

A key limitation of BART is its reliance on step-wise constant functions, which produce non-smooth predictions and, critically for causal inference, fail to reflect increased uncertainty when extrapolating into regions of weak overlap (Hahn et al., 2020). The SoftBART algorithm of Linero and Yang (2018) addresses both shortcomings through two modifications. First, building on the Dirichlet Additive Regression Tree (DART) algorithm of Linero (2018), SoftBART places a sparsity-inducing Dirichlet prior on the splitting probabilities $\boldsymbol{s} = [s_1, \ldots, s_{\mathcal{P}}]'$ that govern variable selection at each branch node, enabling effective regularization when the number of relevant predictors is much smaller than $\mathcal{P}$. Second, SoftBART replaces the hard decision rules $\mathbb{1}(x_{p_b} \leq C_b)$ with smooth logistic gating functions $\psi\big((x_{p_b} - C_b)/\tau_j^{bw}\big)$, where ψ is a logistic link function, and each tree $\mathcal{T}_j$ is assigned its own exponentially-distributed bandwidth parameter τ_j^{bw}. The formal definitions of

the Dirichlet prior on $\boldsymbol{s}$, the soft splitting rules, and all remaining prior specifications are given in Appendix C.

Two properties of SoftBART are central to the SBART+SPL algorithm proposed in Section 3.1. As illustrated in Appendix D.1, SoftBART achieves lower RMSE than BART when predicting smooth functions, as hard splits introduce non-smooth jumps around the true data-generating process. More importantly for our setting, Appendix D.2 shows that SoftBART properly inflates posterior variance in regions of non-overlap when estimating treatment effects, whereas BART predicts a constant treatment effect with no increase in uncertainty. These improvements in precision and uncertainty quantification motivate the replacement of BART with SoftBART within the BART+SPL framework.

3.1. SBART+SPL

To generate draws from the posterior density of τ_P using SBART+SPL, a simplified pseudo-code for the Markov Chain Monte Carlo (MCMC) algorithm in the style of Nethery et al. (2019) is described in Algorithm 2 of Appendix E.1. The main differences compared to the original BART+SPL algorithm are two-fold. In the imputation phase within the region of overlap (steps 1 to 5 in Algorithm 2), SoftBART (Linero and Yang, 2018; Linero, 2022) replaces BART (Chipman et al., 2010) for Bayesian backfitting and the imputation of missing potential outcomes by applying SoftBART to Equation (21), which targets a regression problem similar to Equation (5). In the extrapolation phase (steps 6 to 9 in Algorithm 2), patterns of the estimated individual treatment effects in the region of overlap are extrapolated into the region of non-overlap. Here, SBART+SPL does not rely on the introduced variance inflation parameter in BART+SPL for spline extrapolation. As indicated in Section 3 and Appendix D.2, SoftBART improves uncertainty quantification reasonably compared to BART, thereby allowing us to avoid BART+SPL's unidentifiable variance inflation parameter in SBART+SPL.

3.1.1. IMPUTATION USING SOFTBART

Steps 1 to 3 of Algorithm 2 perform the Bayesian backfitting for the SoftBART algorithm as described in Linero and Yang (2018) using observed data from the region of overlap. The region of overlap is specified using the approach of Nethery et al. (2019) described in Appendix A. Let us use $\boldsymbol{Y}_O^{obs} = [Y_1^{obs}, ..., Y_{n_O}^{obs}]'$ as the vector of observed outcomes of individuals o in the region of overlap. Moreover, $\boldsymbol{X}_O$ is the $n_O \times \mathcal{P}$-dimensional covariate matrix. Note that the posterior distribution for a given tree of the SoftBART algorithm conditions on the imputation noise variance σ^2_{Imp} with

$$\pi\left(\mathcal{T}_j, \mathcal{M}_j | \boldsymbol{Y}_O^{obs}, \sigma^2_{Imp}, \tau_j^{bw}, \mathcal{T}_{(j)}, \mathcal{M}_{(j)}\right), \tag{6}$$

where $\mathcal{T}_{(j)} \equiv \{\mathcal{T}_v : 1 \leq v \leq J,\ v \neq j\}$ is the set of all tree structures except $\mathcal{T}_j$ and, similarly defined, $\mathcal{M}_{(j)}$ is the set of all leaf node parameters except $\mathcal{M}_j$. Equation (6) can be written more compactly using partial residuals. Let the partial residuals be $\boldsymbol{V}_{Oj} = \boldsymbol{Y}_O^{obs} - \sum_{v \neq j} g(\boldsymbol{X}_O; \mathcal{T}_v, \mathcal{M}_v)$. Then, $\boldsymbol{V}_{Oj}$ allows for Bayesian backfitting where all remaining tree parameters are held constant while drawing from one particular tree.

Step 1 mainly updates the tree structures $\mathcal{T}_j$ and the leaf parameters $\mathcal{M}_j$ for each tree j, similarly to Appendix B for the BART algorithm. Updating $\mathcal{T}_j$ and τ_j^{bw} is solved via Metropolis-Hastings. In contrast to BART+SPL, the update for the bandwidth parameters τ_j^{bw} is necessary in SBART+SPL

to allow for flexible gating function shapes as outlined in Section 3. The necessary marginalization over $\mathcal{M}_j$ is performed in closed form due to the Gaussian error assumption. Choosing the number of trees $J = 20$ to be much smaller than in the standard BART algorithm (there we have $J = 200$) balances the increase in computational time due to this marginalization step while $\mathcal{M}_j$ can be drawn from a normal distribution. After iterating over all J trees, the splitting probabilities $\boldsymbol{s}$ are updated via Gibbs sampling using a Dirichlet distribution as indicated in step 2 of Algorithm 2 (see Appendix C.1 for the prior specification). In step 3, one draws the imputation noise variance σ^2_{Imp} and the leaf variance parameter σ^2_μ from Inverse-Gamma distributions, as discussed in Appendix B.1 and Appendix C.1. Moreover, the sparsity-control parameter α from the Dirichlet prior on $\boldsymbol{s}$ is retrieved using slice sampling by Neal (2003). Therefore, steps 1 to 3 are a replacement of the BART algorithm with the SoftBART algorithm in the original BART+SPL algorithm of Nethery et al. (2019) to perform a smoothed imputation for a specific Y_o^{mis} in step 4. Step 5 retrieves an individual treatment effect $\tilde{\Delta}_o$ for an observation $o \in \{1, ..., n_O\}$ in the region of overlap by computing the difference between the observed outcome Y_o^{obs} and the imputed value $\tilde{Y}_o^{mis}$ for the missing outcome Y_o^{mis}. This gives a vector of estimated individual treatment effects for the overlap region, $\tilde{\boldsymbol{\Delta}}_O = [\tilde{\Delta}_1, ..., \tilde{\Delta}_{n_O}]'$. Steps 4 and 5 are equivalent to the derivations in Nethery et al. (2019), we explain their computations in Appendix E.2.

3.1.2. EXTRAPOLATION WITHOUT VARIANCE INFLATION PARAMETER

Steps 6 to 9 in Algorithm 2 for SBART+SPL are similar to BART+SPL except that the tuning parameter for variance inflation in the non-overlap region drops out. The proposed smoothing procedure in Nethery et al. (2019) for BART+SPL estimates individual treatment effects in the region of non-overlap based on $\tilde{\boldsymbol{\Delta}}_O$ retrieved from the imputation phase in Section 3.1. Extrapolation in the non-overlap region relies on a parametric model that extracts treatment effect patterns in the region of overlap by

$$\tilde{\Delta}_o = \boldsymbol{W}'_o \boldsymbol{\beta}_{Smo} + \epsilon_o \, , \epsilon_o \sim \mathcal{N}\left(0, \sigma^2_{Smo}\right). \tag{7}$$

Note that σ^2_{Smo} is the noise variance for the smoothing model which can be different from the noise variance in Equation (6) for the imputation model. Moreover, $\tilde{\Delta}_o$ relates to an estimated individual treatment effect for the overlap region from the imputation phase, and $\boldsymbol{W}_o$ is a vector that holds covariates $\boldsymbol{X}_o$ and restricted cubic splines applied to the imputed or observed outcome and the estimated propensity score of unit $o \in \{1, ..., n_O\}$. An extrapolated individual treatment effect for a unit $o^\neg$ in the non-overlap region, $\tilde{\Delta}_{o^\neg}$, is a draw from the corresponding posterior predictive distribution for individual treatment effects in the non-overlap region. This draw uses updates of $\{\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}\}$ from the overlap region and information about covariates, outcomes, and estimated propensity scores from the non-overlap region. Appendix E.3 presents more details for the estimation of $\tilde{\boldsymbol{\Delta}}_{O^\neg} = [\tilde{\Delta}_1, ..., \tilde{\Delta}_{n_{O^\neg}}]'$ for units $o^\neg \in \{1, ..., n_{O^\neg}\}$. Based on estimates of $\tilde{\boldsymbol{\Delta}}_{O^\neg}$ from Section 3.1.2 and $\tilde{\boldsymbol{\Delta}}_O$ from Section 3.1.1, step 9 in Algorithm 2 draws the PATE estimate $\hat{\Delta}_P$ via Bayesian bootstrap (Rubin, 1981; Wang et al., 2015; Nethery et al., 2019) to account for uncertainty in effect modifier selection (see Appendix E.4 for implementation details).

For BART+SPL, Nethery et al. (2019) introduced an unidentifiable variance inflation parameter for observations in the region of non-overlap by adding a variance component to σ^2_{Smo}. More precisely, they replace the noise term in (7) with

$$\epsilon_o \sim \mathcal{N}\left(0, \sigma^2_{Smo} + \mathbb{1}(\widehat{p.sc}(\boldsymbol{X}_o) \in O^\neg) \cdot \text{var}_{\text{infl}}\right). \tag{8}$$

The idea of this proposal is to satisfy the need for an increase in uncertainty where overlap is weak due to sparse data and the fact that BART itself does not properly account for this need, as pointed out in Section 3. The BART algorithm does not indicate higher uncertainty when there is weak overlap, such that var_{infl} is proposed as a remedy for this issue in Nethery et al. (2019). The recommended default specification for the variance inflation parameter in BART+SPL in Equation (8) depends on the distance from the unit's estimated propensity score to the nearest estimated propensity score in the region of overlap and the range of the estimated individual effects in the region of overlap, range($\tilde{\boldsymbol{\Delta}}_O$) (Nethery et al., 2019). This choice reflects an increase in uncertainty in individual causal effect estimates as observations move farther from the region of overlap while remaining anchored to the scale of $\tilde{\boldsymbol{\Delta}}_O$ supported by the data. Moreover, their construction induces uncertainty that grows linearly with distance into the region of non-overlap. However, var_{infl} remains unidentifiable and is treated as a hyperparameter in Nethery et al. (2019). In contrast, Appendix D.2 suggests that the SoftBART algorithm is capable of reflecting the higher uncertainty in regions of weak overlap for the univariate case. Therefore, the SBART+SPL algorithm removes this unidentifiable variance inflation parameter in step 8 of Algorithm 2. Thus, it does not use var_{infl} to properly account for the increase in variance in the region of non-overlap, which is equivalent to using $\text{var}_{\text{infl}} = 0$ in (8) by default.

In Figure 1, we illustrate the influence of var_{infl} by comparing BART+SPL with and without var_{infl} against SBART+SPL for one simulation dataset based on the setup in Appendix F.1. In the non-overlap region with estimates for $\tilde{\boldsymbol{\Delta}}_{O^\neg}$, BART+SPL degrades sharply in precision. Regardless of whether variance inflation is applied, BART+SPL's credible intervals are too small to cover the true individual effects. In comparison, SBART+SPL decisively reduces bias for $\tilde{\boldsymbol{\Delta}}_{O^\neg}$, translating directly into improved coverage and precision for the PATE estimation. Appendix G further elaborates on this illustration.

4. Simulation Study

The behavior of SBART+SPL compared to BART+SPL for inference regarding τ_P is investigated by following the high-dimensional covariates setting of the simulation study in Nethery et al. (2019). As highlighted in Section 3, a key property of SoftBART as an integral part of SBART+SPL is variable selection in sparse datasets which should give SBART+SPL an edge over BART+SPL in terms of precision concerning inference on τ_P. Added to that, Nethery et al. (2019) indicate severe undercoverage and increasing bias for BART+SPL when the number of covariates rises. SBART+SPL is expected to improve credible interval coverage compared to BART+SPL as elaborated in Section 3.1.2.

The target estimand τ_P is estimated by its corresponding population average treatment effect estimate, $\hat{\Delta}_P$. We compute its estimate by averaging over the $m \in \{1, ..., M\}$ estimates of $\hat{\Delta}_P^{(m)}$ that we retrieve from Algorithm 2 explained in Section 3.1 and Appendix E. As a default, $M = 5,000$ posterior draws (after $10,000$ burn-in draws) are chosen for SBART+SPL. The data-generating process of this simulation study follows the high-dimensional simulation setup in Nethery et al. (2019). There are 10 true confounding variables $X_{1i}, ..., X_{10i}$ that follow either Bernoulli or Normal distributions and depend on the treatment status D_i. In addition to the true confounding variables mentioned above, a set of additional covariates of size $ncov \in \{10, 25, 50\}$ is included in the dataset and drawn independently from a standard Normal distribution irrespective of the treatment status D_i. Appendix F.1 formalizes the data-generating process for the potential outcomes and Figure 5

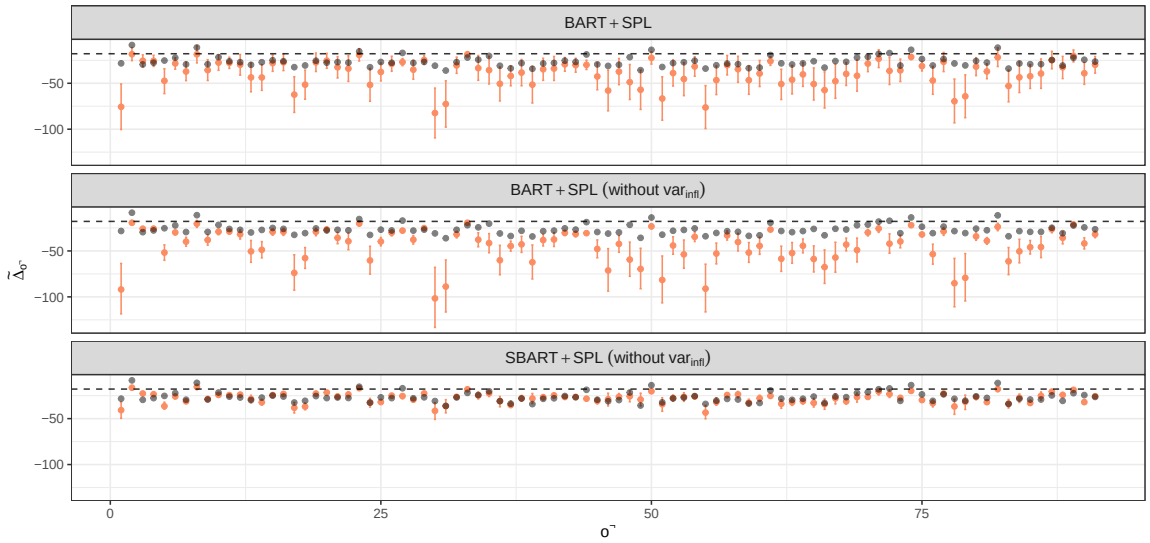

Figure 1: Estimated individual treatment effects $\tilde{\Delta}_{o^\neg}$ with 95% credible intervals for three methods: BART+SPL, BART+SPL (without var_{infl}), and SBART+SPL (without var_{infl} by default). Coloured points and error bars denote posterior point estimates and credible interval estimates for each unit $o^\neg \in \{1, ..., n_{O^\neg}\}$. Black points indicate the true individual treatment effects. The dashed black line marks the true PATE. Results are based on a simulated dataset with $n = 500$ and $ncov = 10$ irrelevant covariates (see Appendices F.1 and G).

in Appendix F.1 illustrates the corresponding propensity score estimates for a selected simulation dataset. The regions of overlap and non-overlap are defined using the approach of Nethery et al. (2019); see Appendix A. The figure shows that there is a substantial region of non-overlap, such that BART+SPL and SBART+SPL have to extrapolate into these regions.

Table 1 shows the comparison of SBART+SPL to BART+SPL and the Accelerated Bayesian Causal Forest (XBCF) (Krantsevich et al., 2023) in terms of precision and coverage for additional covariates $ncov \in \{10, 25, 50\}$ and sample sizes $n \in \{500, 2000\}$ over 120 simulation runs. SBART+SPL improves upon BART+SPL and XBCF in absolute bias and credible interval coverage for both sample sizes and across all numbers of additional covariates. The performance gap in precision between SBART+SPL and BART+SPL is especially pronounced at $n = 500$, where BART+SPL's root-mean-squared error (RMSE) and absolute bias are more than double that of SBART+SPL. For SBART+SPL and BART+SPL, precision and coverage improve with an increase in sample size. SBART+SPL's coverage values are the closest to the nominal level of 95% across all settings, particularly at $n = 2000$. The width of the credible intervals for SBART+SPL reduces only slightly with increasing sample size and reflects more honest uncertainty rather than false precision. While BART+SPL systematically undercovers the true PATE even at $n = 2000$, attributable to overly narrow credible intervals, SBART+SPL exhibits undercoverage at $n = 2000$ only for $ncov = 50$. This results in a more conservative but well-calibrated uncertainty quantification for SBART+SPL in larger samples and moderate sizes of irrelevant covariates.

XBCF decomposes the outcome variable into a prognostic and a treatment effect component in its estimation strategy and severely undercovers credible intervals for the PATE across all configurations in Table 1. This failure likely reflects a mismatch between population-level inference on τ_P and XBCF's indirect potential outcomes modeling strategy, which mainly targets $\tau(\boldsymbol{X}_i)$ with a separate BART prior. Instead, SBART+SPL and BART+SPL directly impute potential outcomes and thereby propagate uncertainty more coherently to the aggregate estimand. In Appendix F.2, we provide additional simulation evidence based on the data-generating process of Krantsevich et al. (2023) and Wang et al. (2024). This second simulation study differs from the setting in Appendix F.1 and aligns more closely with XBCF's estimation approach: unit-level treatment effects are either homogeneous or heterogeneous based on two directly specified covariates only. Moreover, treatment assignment is governed by an explicit non-linear propensity score function rather than implicit covariate distributional shifts as in Appendix F.1. Across both homogeneous and heterogeneous settings with varying numbers of irrelevant covariates, SBART+SPL consistently achieves lower RMSE and absolute bias than BART+SPL while maintaining comparable or improved credible interval coverage, and performs competitively relative to XBCF with only moderately wider intervals.

Table 1: Comparison of PATE estimates in terms of RMSE, absolute bias, and 95% credible interval coverage and width for varying numbers of covariates $ncov \in \{10, 25, 50\}$ and sample sizes $n \in \{500, 2000\}$ across 120 simulation runs based on Appendix F.1.

		$n = 500$				$n = 2000$			
$ncov$	Method	RMSE	|Bias|	Cov.	Width	RMSE	|Bias|	Cov.	Width
10	XBCF	0.759	0.741	0.31	1.324	0.768	0.741	0.00	0.711
	BART+SPL	1.614	1.307	0.51	2.385	0.441	0.365	0.60	0.783
	SBART+SPL	0.702	0.584	0.82	1.842	0.344	0.282	0.98	1.493
25	XBCF	0.710	0.690	0.39	1.335	0.779	0.749	0.01	0.730
	BART+SPL	1.457	1.175	0.47	2.060	0.485	0.374	0.60	0.736
	SBART+SPL	0.631	0.509	0.80	1.681	0.379	0.323	0.95	1.378
50	XBCF	0.659	0.639	0.60	1.337	0.784	0.752	0.00	0.703
	BART+SPL	1.392	1.147	0.43	1.899	0.475	0.382	0.53	0.692
	SBART+SPL	0.696	0.568	0.78	1.747	0.426	0.382	0.90	1.253

Besides BART+SPL and XBCF, Table 3 in Appendix F.1 and Table 4 in Appendix F.2 compare SBART+SPL to three alternative benchmark methods (Chipman et al., 2010; Gutman and Rubin, 2015; Linero, 2022) similarly to the baseline models in Nethery et al. (2019) and a Bayesian linear regression (BLR) model. For each of the three benchmark methods, we implement an untrimmed and a trimmed version such that 10 methods are considered in total. Appendix F presents more information about all methods.

5. Empirical Application: NG Compressor Stations and Mortality Rates in the U.S.

We revisit the empirical application in Nethery et al. (2019) who analyze the effect of natural gas (NG) compressor stations on cancer mortality rates. More detailed information about data collection

can be found in their publication and the original description of Mokdad et al. (2017). The research question is whether the existence of NG compressor stations in United States (U.S.) counties affects the county-level mortality rates for thyroid cancer and leukemia. The study focuses on the mid-western region within the U.S. to exclude confounding alternative pollutions that are argued to be more prevalent in the coastal area. Moreover, exposure to pollution by NG compressor stations might be more relevant in the mid-west due to longer industrial history. To further unconfound the estimation, many pre-treatment variables ($\mathcal{P} = 22$) are considered and represent demographic, socio-economic and behavioral characteristics of each county. Summaries of the four outcome variables of mortality rates and the continuous explanatory variables are presented in Table 7 and Table 8 in Appendix H. In total, $n = 978$ counties with full confounder information are considered. The treatment group consists of 291 counties, where a county is considered to be exposed to treatment when at least one NG compressor station exists in that county. On the contrary, 687 counties are considered to be unexposed. The empirical analysis rests on the assumption that the minimum latency periods of thyroid cancer (2.5 years) and leukemia (0.4 years) are covered by the analysis. Nethery et al. (2019) verified that most of the NG compressor stations have their highest operating dates before or in 2012. That is, the mortality rates of thyroid cancer and leukemia observed in 2014 are argued to cover their minimum latency periods with respect to the exposure to emissions from NG compressor stations. The histogram of propensity score estimates in Figure 7 in Appendix H indicates that the region of non-overlap cannot be neglected as roughly 13% of the observations fall into that region when defined by the approach of Nethery et al. (2019) described in Appendix A. In Table 2, we focus on the (1) log-transformed 2014 leukemia mortality rate and (2) log-transformed 2014 thyroid cancer mortality rate as outcome variables. SBART+SPL yields PATE estimates very close to zero for both leukemia and thyroid cancer mortality, with relatively tight credible intervals that suggest no strong evidence of a treatment effect. Compared to BART+SPL, SBART+SPL considerably reduces interval width. Relative to XBCF, SBART+SPL produces comparable point estimates with intervals that are wider than XBCF's credible intervals.

Table 2: PATE estimates and 95% credible intervals for outcomes (1) leukemia log–2014 mortality, and (2) thyroid log–2014 mortality.

	(1) leukemia				(2) thyroid cancer			
Method	Effect	Lower	Upper	Width	Effect	Lower	Upper	Width
XBCF	0.001	-0.004	0.007	0.011	0.001	-0.002	0.005	0.007
BART+SPL	-0.004	-0.026	0.018	0.044	-0.007	-0.030	0.015	0.045
SBART+SPL	0.001	-0.007	0.009	0.016	0.001	-0.008	0.010	0.017

Table 9 and Table 10 in Appendix H show PATE estimates and the corresponding credible intervals for all methods described in Appendix F. Table 9 extends the results in Table 2 and considers outcome variables (1) log-transformed 2014 leukemia mortality rate and (2) log-transformed 2014 thyroid cancer mortality rate. All methods in Table 9 report credible intervals that include the null and mainly weak but positive PATE estimates of the exposure to NG compressor stations for the 2014 mortality rates of leukemia and thyroid cancer. Table 10 considers the percent point change from 1980 to 2014 in the (3) leukemia mortality rate and (4) thyroid cancer mortality rate. For these long-run changes in mortality reported in Table 10, SBART+SPL delivers positive effect estimates

but substantially narrower credible intervals than BART+SPL. Compared to XBCF, SBART+SPL produces slightly smaller PATE estimates and intervals that include zero, suggesting more conservative inference. Overall, effects are less clearly identified in Table 10. Although all methods yield positive point estimates, only a subset of methods excludes zero from their credible intervals. These findings justify a more detailed investigation of these health effects with finer spatial data. Moreover, other data-related issues like the potentially close locational connection of NG compressor stations and natural gas drilling stations as well as the more insightful (but not available) information about a cancer diagnosis, instead of mortality rates, might affect and influence the analysis presented here and any policy-related measures derived from it (Nethery et al., 2019).

6. Conclusion

Violations of the positivity assumption in the potential outcome framework affect the estimation of the population average treatment effect. While remedies like trimming and weighting often change the initial target estimand, extrapolation approaches might be useful if a researcher wants to conduct inference tailored to the original observational data at hand. This paper proposes SBART+SPL, a method that builds on the framework of BART+SPL by Nethery et al. (2019). By using the sparsity-inducing splitting rule prior and smooth decision rules of SoftBART (Linero and Yang, 2018; Linero, 2018), SBART+SPL adapts more appropriately to an increasing number of covariates. Instead of using BART for the imputation stage, the SoftBART algorithm introduced by Linero and Yang (2018) computes the individual treatment effects in the region of overlap. Moreover, the spline model in the extrapolation phase does not rely on an unidentifiable variance inflation parameter, as used in the BART+SPL algorithm, to properly account for larger uncertainty in regions of non-overlap.

The simulation study in Section 4 illustrates the violation of the positivity assumption under an increasing number of covariates based on two different data-generating processes (Nethery et al., 2019; Krantsevich et al., 2023; Wang et al., 2024). SBART+SPL suggests improvements in precision and coverage for PATE estimation compared to BART+SPL and benchmark methods. Although, SBART+SPL shows still some undercoverage, particularly at $n = 500$, performance improves considerably as sample size n increases for the main simulation study in Section 4 and Appendix F.1. Moreover, SBART+SPL remains competitive in performance under the second data-generating process in Appendix F.2. Section 5 demonstrates the applicability of SBART+SPL by re-estimating the population average treatment effect of NG compressor stations on leukemia and thyroid cancer mortality rates at U.S. county level. In Table 9, SBART+SPL aligns with competing methods in finding no evidence of a positive PATE, while delivering substantially tighter credible intervals than BART+SPL. Results for long-run mortality changes in Table 10 are inconclusive. Although several methods produce positive point estimates with credible intervals bounded away from zero, others fail to rule out a null effect, suggesting that definitive conclusions require further analysis with larger or richer data sources.

Acknowledgments

The author is grateful to Christoph Hanck for valuable comments and suggestions that significantly improved the paper. The author also thanks three anonymous referees whose constructive feedback and insightful discussions contributed substantially to the final version. Earlier versions of this work

were presented at the European Causal Inference Meeting (EuroCIM) 2024 in Copenhagen and the Workshop on Causal Inference and Machine Learning 2024 in Groningen. The author thanks the participants at both events for their helpful remarks and stimulating discussions.

References

Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424, May 2011. ISSN 0027-3171. doi: 10.1080/00273171.2011.568786.

Paul-Christian Bürkner. Brms: An R package for bayesian multilevel models using stan. *J. Stat. Softw.*, 80(1), 2017. ISSN 1548-7660. doi: 10.18637/jss.v080.i01.

Hugh Chipman, Edward George, and Robert Mcculloch. Bayesian Ensemble Learning. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948, 1998. ISSN 0162-1459. doi: 10.2307/2669832. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, March 2010. ISSN 1932-6157, 1941-7330. doi: 10.1214/09-AOAS285. Publisher: Institute of Mathematical Statistics.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, March 2009. ISSN 0006-3444. doi: 10.1093/biomet/asn055.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, 34(1):43–68, 2019. ISSN 0883-4237. Publisher: Institute of Mathematical Statistics.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, April 2021. ISSN 0304-4076. doi: 10.1016/j.jeconom.2019.10.014.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, October 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013203451. Publisher: Institute of Mathematical Statistics.

R. Gutman and Donald B. Rubin. Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Statistics in Medicine*, 32(11):1795–1814, 2013. ISSN 1097-0258. doi: 10.1002/sim.5627. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5627.

Roee Gutman and Donald B. Rubin. Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in medicine*, 34(26):3381–3398, November 2015. ISSN 0277-6715. doi: 10.1002/sim.6532.

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):965–1056, September 2020. ISSN 1936-0975, 1931-6690. doi: 10.1214/19-BA1195. Publisher: International Society for Bayesian Analysis.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 1st edition, 2020. doi: https://doi.org/10.1201/9781315374932.

Miguel A. Hernán and James M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586, July 2006. ISSN 0143-005X. doi: 10.1136/jech.2004.029496.

Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 7(3), September 2013. ISSN 1932-6157. doi: 10.1214/13-AOAS630. arXiv: 1311.7244.

Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, March 2020. ISSN 2326-8298, 2326-831X. doi: 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, January 2011. ISSN 1061-8600. doi: 10.1198/jcgs.2010.08162. Publisher: ASA Website _eprint: https://doi.org/10.1198/jcgs.2010.08162.

Liangyuan Hu, Bian Liu, Jiayi Ji, and Yan Li. Tree-Based Machine Learning to Identify and Understand Major Determinants for Stroke at the Neighborhood Level. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 9(22):e016745, November 2020. ISSN 2047-9980. doi: 10.1161/JAHA.120.016745.

Nikolay Krantsevich, Jingyu He, and P. Richard Hahn. Stochastic tree ensembles for estimating heterogeneous effects. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6120–6131. PMLR, 25–27 Apr 2023.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400, January 2018. ISSN 0162-1459. doi: 10.1080/01621459.2016.1260466. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2016.1260466.

Fan Li, Laine E Thomas, and Fan Li. Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*, 188(1):250–257, January 2019. ISSN 0002-9262. doi: 10.1093/aje/kwy201.

Fan Li, Peng Ding, and Fabrizia Mealli. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, May 2023. doi: 10.1098/rsta.2022.0153. Publisher: The Royal Society.

Antonio R. Linero. Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, 113(522):626–636, April 2018. ISSN 0162-1459. doi: 10.1080/01621459.2016.1264957. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1264957.

Antonio R. Linero. SoftBart: Soft Bayesian Additive Regression Trees, October 2022. arXiv:2210.16375 [stat].

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, November 2018. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12293.

Ali H. Mokdad, Laura Dwyer-Lindgren, Christina Fitzmaurice, Rebecca W. Stubbs, Amelia Bertozzi-Villa, Chloe Morozoff, Raghid Charara, Christine Allen, Mohsen Naghavi, and Christopher J. L. Murray. Trends and Patterns of Disparities in Cancer Mortality Among US Counties, 1980-2014. *JAMA*, 317(4):388–406, January 2017. ISSN 0098-7484. doi: 10.1001/jama.2016.20324.

Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, June 2003. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1056562461. Publisher: Institute of Mathematical Statistics.

Rachel C. Nethery, Fabrizia Mealli, and Francesca Dominici. ESTIMATING POPULATION AVERAGE CAUSAL EFFECTS IN THE PRESENCE OF NON-OVERLAP: THE EFFECT OF NATURAL GAS COMPRESSOR STATION EXPOSURE ON CANCER MORTALITY. *The Annals of Applied Statistics*, 13(2):1242–1267, June 2019. ISSN 1932-6157. doi: 10.1214/18-AOAS1231.

Matthew T. Pratola. Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian Analysis*, 11(3):885–911, September 2016. ISSN 1936-0975, 1931-6690. doi: 10.1214/16-BA999. Publisher: International Society for Bayesian Analysis.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

Veronika Ročková and Stéphanie Van Der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4), August 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1879.

Donald B. Rubin. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134, January 1981. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345338. Publisher: Institute of Mathematical Statistics.

Donald B. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. ISSN 0162-1459. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Enakshi Saha. Theory of Posterior Concentration for Generalized Bayesian Additive Regression Trees, April 2023. arXiv:2304.12505 [math].

Til Stürmer, Kenneth J. Rothman, Jerry Avorn, and Robert J. Glynn. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. *American Journal of Epidemiology*, 172(7):843–854, October 2010. ISSN 1476-6256. doi: 10.1093/aje/kwq198.

G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990. ISBN 978-0-89871-244-5.

Chi Wang, Francesca Dominici, Giovanni Parmigiani, and Corwin Matthew Zigler. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3):654–665, September 2015. ISSN 1541-0420. doi: 10.1111/biom.12315.

Meijia Wang, Jingyu He, and P. Richard Hahn. Local gaussian process extrapolation for bart models with applications to causal inference. *Journal of Computational and Graphical Statistics*, 33(2): 724–735, 2024. doi: 10.1080/10618600.2023.2240384.

T. Wendling, K. Jung, A. Callahan, A. Schuler, N. H. Shah, and B. Gallego. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37(23):3309–3324, 2018. ISSN 1097-0258. doi: 10.1002/sim.7820. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7820.

Daniel Westreich and Stephen R. Cole. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677, March 2010. ISSN 0002-9262. doi: 10.1093/aje/kwp436.

Yuhong Wu, Tjelmeland , Håkon, , and Mike West. Bayesian CART: Prior Specification and Posterior Simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66, March 2007. ISSN 1061-8600. doi: 10.1198/106186007X180426. Publisher: ASA Website _eprint: https://doi.org/10.1198/106186007X180426.

Angela Yaqian Zhu, Nandita Mitra, and Jason Roy. Addressing positivity violations in causal effect estimation using Gaussian process priors. *Statistics in Medicine*, 42(1):33–51, January 2023. ISSN 1097-0258. doi: 10.1002/sim.9600.

Yaqian Zhu, Rebecca A. Hubbard, Jessica Chubak, Jason Roy, and Nandita Mitra. Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches. *Pharmacoepidemiology and drug safety*, 30(11):1471–1485, November 2021. ISSN 1053-8569. doi: 10.1002/pds.5338.

Appendix A. Specifying Regions of Overlap and Non-overlap

This paper follows the definition of the region of overlap O and region of non-overlap $O^{\neg}$ of Nethery et al. (2019). For notational convenience, we drop $\boldsymbol{X}_i$ here when considering estimated propensity scores. Let $\widehat{p.sc}_{\text{cand}}$ be a candidate propensity score value $\widehat{p.sc}_{\text{cand}} \in P.SC = \left[\widehat{p.sc}_{(1)}, \widehat{p.sc}_{(n)}\right] \subset$

$(0,1)$ with $\widehat{p.sc}_{(j)}$ being the j-th order statistic, i.e., $\widehat{p.sc}_{(1)}$ is the minimum of the estimated propensity scores for all observations in the full sample. Furthermore, let n_d be the number of individuals in each treatment group $d \in \{0,1\}$ and $\widehat{p.sc}^{d}_{(i)}$ be the i-th propensity score order statistic in treatment group d. The candidate $\widehat{p.sc}_{\text{cand}}$ has a reasonable degree of overlap if one can construct a set,

$$\left\{\widehat{p.sc}_{\text{cand}}, \widehat{p.sc}^{d}_{(i)}, ..., \widehat{p.sc}^{d}_{(i+b_O)}\right\}, \tag{9}$$

with the following two conditions for the treatment group and the control group, separately:

- The set includes $\widehat{p.sc}_{\text{cand}}$ itself, as well as more than b_O estimated propensity scores.
- The set has a range smaller than a_O.

Therefore, we define the region of overlap O as the set of estimated propensity scores that fulfill those conditions and, hence, its units belong to the overlap region subsample with $o = 1, ..., n_O$. $O^{\neg}$ is the complement of this set, and its units belong to the non-overlap subsample with $o^{\neg} = 1, ..., n_{O^{\neg}}$.

$$O = \left\{ \begin{array}{c} \widehat{p.sc}_{\text{cand}} \in P.SC : \text{range}\left(\left\{\widehat{p.sc}_{\text{cand}}, \widehat{p.sc}^{d}_{(i)}, ..., \widehat{p.sc}^{d}_{(i+b_O)}\right\}\right) < a_O, \\ \text{for some } i = 1, ..., n_d - b_O, \\ \text{for } d \in \{0,1\}. \end{array} \right\}. \tag{10}$$

Any user of this overlap definition is left with choosing appropriate values for a_O and b_O. As outlined in Nethery et al. (2019), the tuning parameters have recommended default choices with $a_O = 0.1$ and $b_O = 7$, respectively. This implies that an estimated propensity score belongs to the region of overlap if, for each treatment group, one can construct a set with more than 7 estimated propensity scores around this point such that this set has a range smaller than 0.1. In general, lower values of a_O together with higher values of b_O define the region of non-overlap more conservatively. Estimated propensity scores belonging to the set are allowed to vary only mildly by opting for a low a_O, and one needs relatively many of these estimated propensity scores that are close to $\widehat{p.sc}_{\text{cand}}$ due to the large b_O. Likewise, high values for a_O and low values for b_O define the region of non-overlap less conservatively.

Appendix B. Bayesian Additive Regression Trees

We follow the notation of Linero and Yang (2018) to describe how BART can be extended to Soft-BART. Let us consider the general semiparametric Gaussian regression problem in (5) to exemplify how BART is used for predicting outcomes Y_i given covariates $\boldsymbol{X}_i$. Let us define the non-parametric regression function with $f_0(\boldsymbol{X}_i = \boldsymbol{x})$ being some realization of the function

$$f(\boldsymbol{X}_i = \boldsymbol{x}) = \sum_{j=1}^{J} g\left(\boldsymbol{x}; \mathcal{T}_j, \mathcal{M}_j\right) = \sum_{j=1}^{J}\sum_{l=1}^{L_j} \mu_{jl}\phi\left(\boldsymbol{x}; \mathcal{T}_j, l\right), \boldsymbol{x} \in \mathbb{R}^{\mathcal{P}}, \tag{11}$$

with $\mathcal{T}_j$ the branching process of tree $j \in \{1, ..., J\}$ while $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$ represents the leaf node parameters of tree j with leaf $l \in \{1, ..., L_j\}$. For each tree j and given data x, the function

$g(\boldsymbol{x}; \mathcal{T}_j, \mathcal{M}_j)$ relates a branching process to the leaf node parameters and can be reformulated by multiplying each leaf node parameter μ_{jl} with the respective splitting rule $\phi(\boldsymbol{x}; \mathcal{T}_j, l)$. BART uses the sum-of-trees specification in (11) to learn flexible functions $g(\boldsymbol{x}; \mathcal{T}_j, \mathcal{M}_j)$ that are able to predict Y_i based on $\boldsymbol{X}_i$ and prior distributions on $\mathcal{T}_j, \mathcal{M}_j$ and error variance σ^2 from (5).

One can view the general BART algorithm of Chipman et al. (1998, 2006, 2010) as an aggregation of weak-learning decision trees as described in the decision tree boosting setup, but in a Bayesian fashion. The BART algorithm proceeds similar to Gradient Boosting Machines (GBM) in that it only uses the observed data and performs tree construction dependently (Friedman, 2001). Different from GBM, BART does not fit an entirely new tree to the residuals of the previous tree. Instead, BART has a built-in perturbation process that changes the tree structure of the previous tree by either adding or pruning branches or modifying each of the leaf node predictions. This prevents getting stuck in local minima through increased exploration of the model space. To avoid overfitting, GBM uses tuning parameters by limiting the maximum depth of each tree, resulting in weak learning trees and scaling down the individual contribution of each new tree. In contrast, BART reduces overfitting in a rather data-driven way by using a prior distribution for each tree to regularize its size and fit. The independent tree priors prefer the construction of rather small trees and lead to leaf parameter values approaching zero (Hill et al., 2020). Besides its flexible nonparametric function estimation approach, BART possesses further advantages that are useful for drawing inference in empirical studies with observational data. First, the BART algorithm shows good performance in setups with high noise that are prevalent in the causal inference literature of economics or medicine. The usage in several data challenges and empirical applications demonstrates competitive and superior performance compared to other methods (Dorie et al., 2019; Hu et al., 2020; Wendling et al., 2018). Second, the prior distributions used in the BART algorithm favor data patterns with low-order interactions over high-order interactions. Although high-order interactions play a central role in pattern recognition for, i.e., languages or images, they are argued to be of limited relevance in the traditional statistical analysis of the social sciences. In these circumstances of low-order interactions, BART approaches optimal posterior concentration rates (Ročková and Van Der Pas, 2020; Saha, 2023).

The standard BART algorithm by Chipman et al. (2010) describes an ensemble of decision trees with $(\mathcal{T}_j, \mathcal{M}_j) \overset{i.i.d.}{\sim} \pi_{\mathcal{T}}(\mathcal{T}_j)\, \pi_{\mathcal{M}}(\mathcal{M}_j|\mathcal{T}_j)$ with $(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ being parameter priors of the decision trees. Referring back to the semiparametric Gaussian problem in (5), BART takes the decision trees j to be independent such that one can write

$$\pi\left((\mathcal{T}_1, \mathcal{M}_1), ..., (\mathcal{T}_J, \mathcal{M}_J), \sigma|\theta\right) = \left[\prod_{j=1}^{J} \pi_{\mathcal{T}}(\mathcal{T}_j|\theta)\, \pi_{\mathcal{M}}(\mathcal{M}_j|\mathcal{T}_j)\right] \pi_\sigma(\sigma) \tag{12}$$

with $\pi_{\mathcal{M}}(\mathcal{M}_j|\mathcal{T}_j) = \prod_{l=1}^{L_j} \pi_{\mathcal{M}}(\mu_{jl}|\mathcal{T}_j)$ and θ being the vector of hyperparameters specified in Algorithm 1. Consequently, one needs to define prior distributions for $(\pi_{\mathcal{T}}, \pi_{\mathcal{M}}, \pi_\sigma)$. The prior on $\pi_{\mathcal{T}}$ is twofold, focusing on (1) the tree shape and (2) the splitting rules connected to each branch node. The prior on $\pi_{\mathcal{M}}$ describes the distribution of leaf node parameters $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$. The prior on π_σ is a distribution for the noise variance.

B.1. Prior Specification

Following Chipman et al. (2010), the prior on the tree shape of tree $\mathcal{T}_j$ can be outlined as a branching process. Each tree begins with a root node of depth $dep = 0$. Then, it is iteratively decided how deep one is growing the tree. The root node is converted into a branch node b with two child nodes with probability $Pr\left(b \text{ is branch node}\right) = \frac{\gamma}{(1+dep)^{\beta}}$ and is converted into a leaf node with its complementary probability. This procedure repeats until all nodes at a given level of depth level are leaf nodes. Consequently, one has to specify the hyperparameters (γ, β). This paper uses the default values of Chipman et al. (2010) by $\gamma = 0.95$ and $\beta = 2$.

The prior on the splitting rules connected to each branch node b is defined by sampling a predictor p_b and a cutpoint C_b. Based on those sampled values, one can compare x_{p_b} and C_b to decide how to channel $\boldsymbol{x}$ down the tree. The standard BART algorithm samples p_b and C_b from separate uniform distributions. The hard branch splitting rules of the standard BART algorithm can be formalized by specifying $\phi\left(\boldsymbol{x}; \mathcal{T}_j, l\right)$ used in (11) by

$$\phi\left(\boldsymbol{x}; \mathcal{T}_j, l\right) = \prod_{b \in \mathcal{A}(l)} \mathbb{1}\left(x_{p_b} \leq C_b\right)^{R_b} \mathbb{1}\left(x_{p_b} > C_b\right)^{1-R_b}. \tag{13}$$

Let us denote $\mathcal{A}(l)$ as nodes that are ancestral to leaf node l. Furthermore, $R_b \in \{0, 1\}$, such that $R_b = 0$ if the path from the root to l goes right at b and $R_b = 1$ if left.

For the prior on the leaf node parameters $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$, one can exploit conjugacy of the normal distribution $\mathcal{N}(\mu_\mu, \sigma_\mu^2)$ to marginalize out μ_{jl} as in Chipman et al. (2010). For the standard BART algorithm, the outcome variable Y is transformed to be placed into the interval $[-0.5, 0.5]$ and one samples $\mu_{jl} \sim \mathcal{N}(\mu_\mu = 0, \sigma_\mu^2 = 0.5/k\sqrt{J})$. Consequently, one has to specify two further hyperparameters (k, J) with $k = 2$ and $J = 200$ being default values for the standard BART algorithm.

Again, one can exploit conjugacy for the prior on the noise variance σ^2. For $\pi_\sigma(\sigma)$ one can use the inverse chi-square distribution such that $\sigma^2 \sim \upsilon\lambda/\chi_\upsilon^2$ with default choices of $(\upsilon = 3, q_\lambda = 0.9)$ to avoid neither an overly conservative nor an excessively aggressive prior. Here, λ is chosen such that the the probability of σ being lower than the initial $\hat{\sigma}$ is q_λ by determining $Pr\left(\sigma < \hat{\sigma}\right) = q_\lambda$. The naive $\hat{\sigma}$ is either estimated by being the sample standard deviation of Y or the residual standard deviation from a simple linear regression of Y on $\boldsymbol{X}$. This ensures that the prior is weakly informative and favors values of σ that are similar to its naive estimate, but still allows for variation. Oftentimes, the random draw is taken from the related Inverse-Gamma distribution with $\sigma^2 \sim \text{IG}\left(\frac{\upsilon}{2}, \frac{\upsilon\lambda}{2}\right)$.

B.2. Posterior Inference

Algorithm 1 describes the pseudo-code of one iteration of the general Bayesian backfitting procedure to update $(\mathcal{T}_j, \mathcal{M}_j)$ for BART (Chipman et al., 2010; Hill et al., 2020). Bayesian backfitting adopts a blocked Metropolis-Hastings strategy to explore the posterior distribution (Chipman et al., 1998; Wu et al., 2007). In essence, the sampling proceeds in a structured manner: the tree $\mathcal{T}_j$ is first drawn from its marginal posterior distribution, followed by drawing the leafs $\mathcal{M}_j$ from their full conditional distribution. More complicated models use a Metropolis-within-Gibbs Markov Chain Monte Carlo model to update other parameters by exploiting further Gibbs or Metropolis-Hastings steps as presented in Algorithm 2 later on for SBART+SPL.

In step 1, Algorithm 1 proposes a new tree structure $\mathcal{T}_j^\star$ based on some proposal distribution based on the current tree structure $\mathcal{T}_j$ by $q(\mathcal{T}_j^\star; \mathcal{T}_j)$. Mostly, BART implementations rely on random perturbations of the current tree structure by one of the following moves: *Grow/Birth* turns a considered leaf node into a branching node and splits into two new leaf nodes, *Prune/Death* collapses two neighboring leaf nodes back to one leaf node, *Change* modifies the decision rule associated with a non-terminal node, and, *Swap* exchanges the decision rules of two non-terminal nodes (Hill et al., 2020; Linero and Yang, 2018). Wu et al. (2007) and Pratola (2016) provide an extended discussion of alternative proposal distributions within BART.

The second step sets the Metropolis ratio, a_{MR} in Equation (15), by using the proposal distribution from step 1 and computing the integrated likelihood function

$$\mathcal{L}(\mathcal{T}_j; \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \theta) = \int \left(\prod_{i=1}^{n} p(Y_i \mid \mathcal{T}_j, \mathcal{M}_j, \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \theta) \right) p(\mathcal{M}_j \mid \mathcal{T}_j, \theta)\, d\mathcal{M}_j, \tag{14}$$

where $\mathcal{T}_{(j)} \equiv \{\mathcal{T}_v : 1 \leq v \leq J,\ v \neq j\}$ is the set of all tree structures except $\mathcal{T}_j$ while the set $\mathcal{M}_{(j)}$ is defined similarly for leaf parameters (Hill et al., 2020). Moreover, the vector θ collects all necessary parameters considered in the branching process in Appendix B.1 and might be extended to include further parameters in more involved models (i.e. bandwidth parameter τ_j^{bw} in Appendix C.1 for SoftBART).

Step 3 accepts the tree proposal $\mathcal{T}_j^\star$ with probability $\min(1, a_{MR})$. If the tree proposal is rejected, $\mathcal{T}_j$ stays at its current status. Finally, leaf parameters $\mathcal{M}_j$ are drawn from their full conditional distribution in step 4.

Algorithm 1 One iteration of Bayesian backfitting for BART to update $(\mathcal{T}_j, \mathcal{M}_j)$

Input :

- Initial parameters $\left\{\mathcal{T}_j^{(0)}, \mathcal{M}_j^{(0)}\right\}$
- Hyperparameters $\theta = (\gamma, \beta, k, J)$ as specified in Appendix B.1
- Observed data $\boldsymbol{Y}^{obs}, \boldsymbol{X}$

Output: Updated values for $(\mathcal{T}_j, \mathcal{M}_j)$

for $j \leftarrow 1$ **to** J **do**

1. Propose new tree structure $\mathcal{T}_j^\star$ from proposal distribution $q(\mathcal{T}_j^\star; \mathcal{T}_j)$
2. Set Metropolis ratio, a_{MR}, to

$$a_{MR} \leftarrow \frac{\mathcal{L}(\mathcal{T}_j^\star; \mathcal{M}_j, \theta) p(\mathcal{T}_j^\star)}{\mathcal{L}(\mathcal{T}_j; \mathcal{M}_j, \theta) p(\mathcal{T}_j)} \frac{q(\mathcal{T}_j; \mathcal{T}_j^\star)}{q(\mathcal{T}_j^\star; \mathcal{T}_j)} \tag{15}$$

3. Set $\mathcal{T}_j \leftarrow \mathcal{T}_j^\star$ with probability $\min(1, a_{MR})$
4. Sample $\mathcal{M}_j \sim p(\mathcal{M}_j | \mathcal{T}_j, \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \theta, \boldsymbol{Y}^{obs}, \boldsymbol{X})$

end

Appendix C. From BART to SoftBART

A drawback of the BART algorithm in Appendix B is its reliance on step-wise constant functions that lead to non-smooth predictions. The SoftBART algorithm of Linero and Yang (2018) resolves this shortcoming by introducing sparsity and smoothness into the conventional BART algorithm

of Chipman et al. (1998, 2010). Linero and Yang (2018) and Ročková and Van Der Pas (2020) show that the convergence rate of BART can be improved via smoothness. Moreover, they derive suitable posterior concentration rates for SoftBART when the number of relevant, true predictors is much smaller than the number of considered covariates $\mathcal{P}$ and the true data-generating function is driven by low-order interactions. Appendix C.1 generalizes BART to SoftBART by introducing sparsity-inducing splitting probabilities and allowing for smoothness-inducing decision rules.

C.1. Soft Bayesian Additive Regression Trees

In Linero (2018), the Dirichlet Additive Regression Trees (DART) algorithm extends the BART algorithm with regard to the prior on the splitting rules of branch node b by adding a sparsity-inducing prior. At each branch node, the tree randomly selects a predictor index $p_b \in \{1, ..., \mathcal{P}\}$ and a threshold C_b, then splits the data by comparing x_{p_b} to C_b. Repeating this across nodes b and trees j yields a flexible, data-adaptive partition of the predictor space. DART samples $p_b \sim \text{Categorical}(\boldsymbol{s})$ with probability vector $\boldsymbol{s} = [s_1, ..., s_{\mathcal{P}}]'$. We specify an additional prior on $\boldsymbol{s}$ that induces sparsity into the splitting probabilities. That is, we sample

$$\boldsymbol{s} \sim \text{Dirichlet}\left(\alpha/\mathcal{P}^{\xi}, ..., \alpha/\mathcal{P}^{\xi}\right) \tag{16}$$

with $\frac{\alpha}{\alpha+\mathcal{P}} \sim \text{Beta}\left(\alpha_{\boldsymbol{s}} = 0.5, \beta_{\boldsymbol{s}} = 1\right)$ and $\xi = 1$ such that α governs the assumed sparsity with $(\alpha_{\boldsymbol{s}}, \beta_{\boldsymbol{s}})$ balancing between sparse and non-sparse setups. The threshold values are sampled by $C_b \sim \text{Uniform}(a_u, b_u)$ with (a_u, b_u) chosen such that the cell $\mathbb{R}^{\mathcal{P}}$ is split along the p_b-th coordinate.

The SoftBART algorithm builds upon the DART algorithm. It extends the procedure by implementing smoothness-inducing decision rules (Linero and Yang, 2018; Linero, 2022). The algorithm replaces the hard decision rules $\mathbb{1}\left(x_{p_b} \leq C_b\right)$ and $\mathbb{1}\left(x_{p_b} > C_b\right)$ as in Equation (13) with the smooth decision rule $\psi\left(\frac{x_{p_b} - C_b}{\tau_j^{bw}}\right)$ leading to

$$\phi\left(\boldsymbol{x}; \mathcal{T}_j, l\right) = \prod_{b \in \mathcal{A}(l)} \psi\left(\frac{x_{p_b} - C_b}{\tau_j^{bw}}\right)^{R_b} \left[1 - \psi\left(\frac{x_{p_b} - C_b}{\tau_j^{bw}}\right)\right]^{1-R_b}. \tag{17}$$

The SoftBART algorithm assigns each tree $\mathcal{T}_j$ a separate, exponentially-distributed bandwidth parameter τ_j^{bw}[1] and a logistic link function ψ. With regard to the prior on the leaf node parameters $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$, the SoftBART algorithm places an additional hyperprior on σ_μ^2 by sampling $\sigma_\mu \sim \text{Cauchy}^{+}(\hat{\sigma}_\mu)$ with $\hat{\sigma}_\mu = 0.5/k\sqrt{J}$ and choosing by default fewer trees with $J = 20$.

Differences between SoftBART and BART are illustrated in two stylized examples in Appendix D.1 and D.2 which are closely related to illustrations presented in Hahn et al. (2020) and Linero and Yang (2018). The first example points out the difference between hard and soft decision rules in a simple prediction task of a smooth sine curve as well as a step function. SoftBART indicates lower RMSE compared to BART in both cases as the usage of BART leads to non-smooth jumps around the true data-generating function. In the second example in Appendix D.2, SoftBART improves uncertainty quantification in the region of non-overlap in a simple conditional average treatment effect estimation setup. BART does not account for an increase in uncertainty and predicts a constant

1. $\tau_j^{bw} \sim \text{Exp}\left(\text{scale} = 0.1\right)$. Note that if $\tau_j^{bw} \to 0$ yields a standard decision tree.

treatment effect in the region of non-overlap. Instead, SoftBART inflates the variance around the conditional average treatment effects more properly when faced with weak overlap. These SoftBART properties of improved predictive power and enhanced uncertainty quantification are leveraged within the SBART+SPL algorithm described in Section 3.1 to alleviate the shortcomings of BART+SPL concerning precision and coverage for population average treatment effect estimation in the presence of sparse data.

Appendix D. Properties of SoftBART

We demonstrate two properties of SoftBART that motivate its usage within SBART+SPL in Section 3: proper approximation of smooth and step functions (Appendix D.1) and appropriate uncertainty inflation in regions of non-overlap (Appendix D.2).

D.1. Hard and Soft Decision Rules

The following stylized prediction example based on Linero and Yang (2018) visualizes the contrast between hard decision rules, as in Equation (13), and soft decision rules, as in Equation (17). Let the univariate regressor be a sequence of values ranging from 0 to 1, incremented by 0.01, such that $X_i = (0, 0.01, 0.02, ..., 1)$. The continuous outcome variable is generated with

$$Y_i(X_i = x) = \sin(2\pi x) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 0.1^2). \tag{18}$$

In-sample predictions of the outcome variable using BART and SoftBART are given in Figure 2. Whereas the BART predictions wiggle non-smoothly around the true sine curve, the SoftBART predictions result in a similarly shaped curve compared to the true sine curve. This results in a lower RMSE value for the SoftBART algorithm (0.046) compared to the BART algorithm (0.067).

The improvement in RMSE is not an artifact of the underlying smooth sine curve that generates our outcome. Let us change our outcome model to a step function by:

$$Y_i(X_i = x) = 2 - 4 \cdot \mathbb{1}(x > 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 0.1^2). \tag{19}$$

In-sample predictions of the outcome variable using BART and SoftBART are given in Figure 3. The RMSE for SoftBART (1.360) is slightly lower compared to the RMSE of BART (1.366), although the outcome model mimics a step function with a hard jump from 2 to -2 around $x = 0.5$.

D.2. Uncertainty Quantification

In addition to improved precision, the SoftBART algorithm seems to be able to improve uncertainty quantification compared to BART when estimating treatment effects. This section exemplifies this improvement by recollecting a simple example of uncertainty quantification around conditional average treatment effects similar to the discussion part in Hahn et al. (2020). Let us generate $n = 200$ observations with 100 units for each treatment group $D_i \in \{0, 1\}$. The univariate regressor $X_i \sim \text{Ga}(\mu_{Ga}, sd = 8)$ has $\mu_{Ga} = 35$ for the treated units and $\mu_{Ga} = 60$ for the control units. Based on the regressor, a linear outcome model is used to generate the respective outcome values by

$$Y_i(X_i = x, D_i = d) = 10 + 5 \cdot d + 0.3 \cdot x + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 1). \tag{20}$$

Figure 4 displays in the upper panel observed and predicted outcome values for each treatment group separately based on either SoftBART or BART. The lower panel shows estimates of the conditional average treatment effects (CATE), $\hat{\tau}(X_i)$ and the corresponding credible intervals for SoftBART and BART. The observed values overlap for both treatment groups with values lying approximately in the interval $X_i = x \in (40, 60)$ but we have non-overlap for values outside of this interval. Consequently, estimation methods should quantify this uncertainty in the non-overlapping region with increasing credible intervals around $\hat{\tau}(X_i)$. SoftBART accounts for this increase in uncertainty while BART does not transmit the increased uncertainty into its credible intervals resulting in overconfident predictions and treatment effect estimates for x-values outside the interval $(40, 60)$.

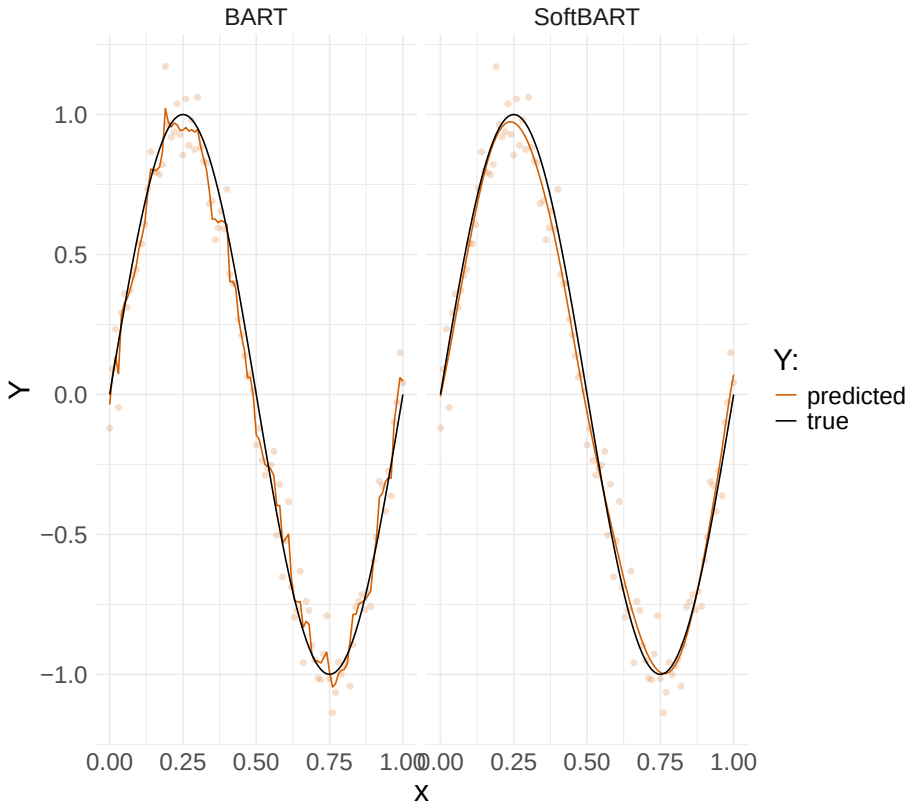

Figure 2: In-sample predictions for BART and SoftBART based on the example in Linero and Yang (2018) and Appendix D.1. Outcomes are generated by the sine curve in (18).

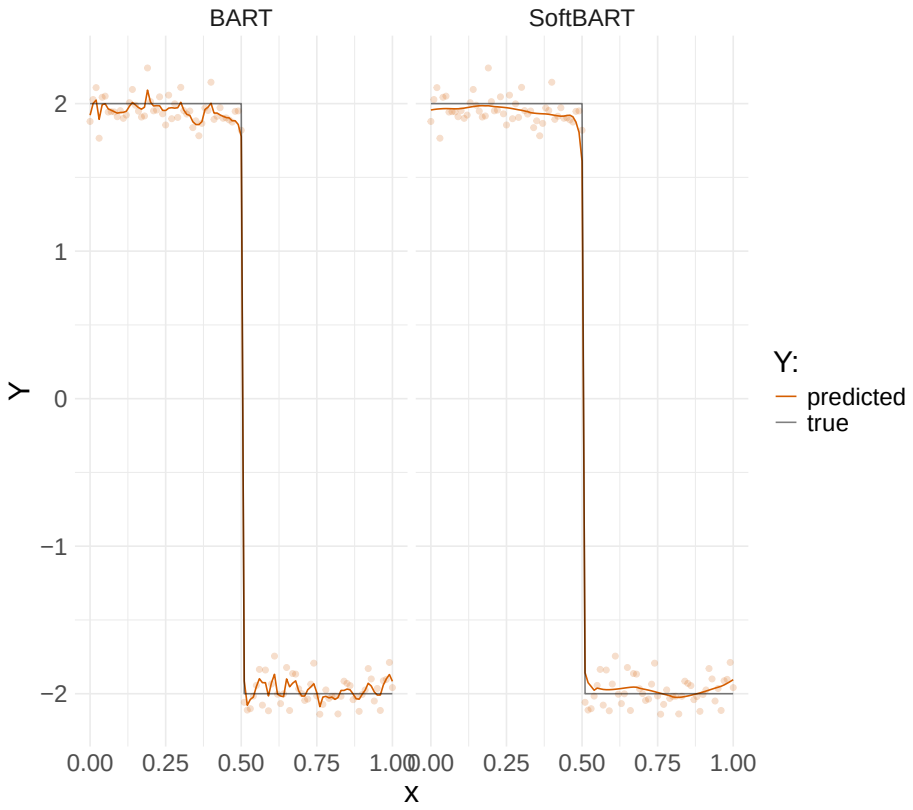

Figure 3: In-sample predictions for BART and SoftBART based on the example in Linero and Yang (2018) and Appendix D.1. Outcomes are generated by the step function in Equation (19).

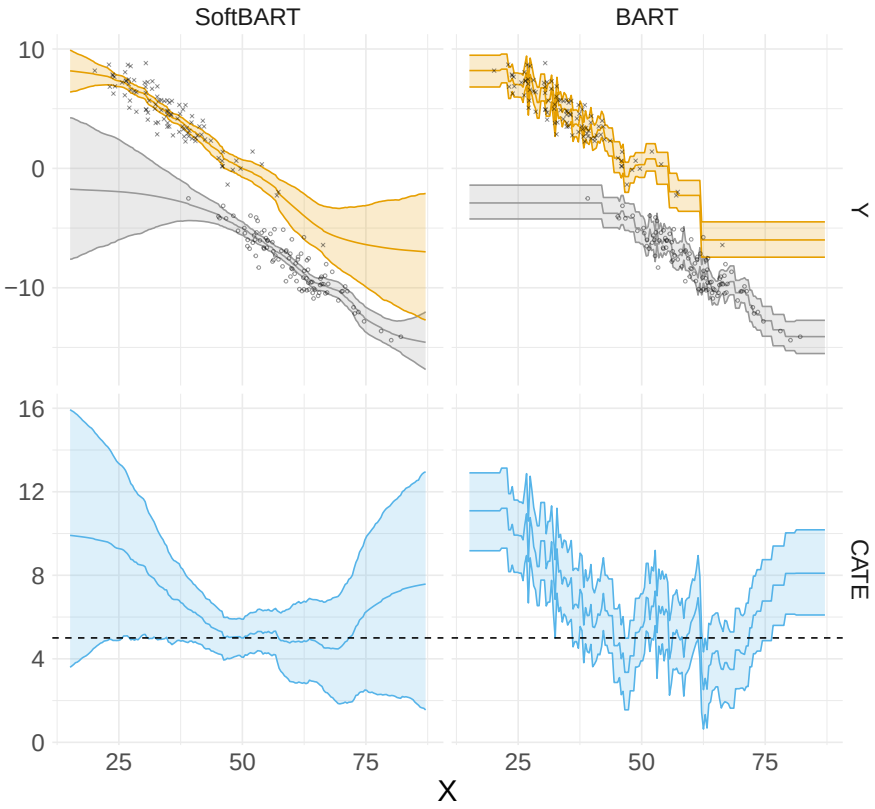

Figure 4: Comparison of uncertainty quantification for BART and SoftBART based on the example in Hahn et al. (2020) and Appendix D.2. Outcomes are generated following Equation (20).

Appendix E. SBART+SPL algorithm

This paper implements the SBART+SPL algorithm by embedding the SoftBART algorithm into the Bayesian backfitting algorithm of BART+SPL using the `MakeForest()` function and its corresponding `Rcpp_Forest` data structure as outlined in Linero (2022) for the `R` programming language (R Core Team, 2021). The SoftBART package by Linero (2022) introduces a flexible implementation of BART models that can adapt to sparsity and smoothness as outlined in Appendix C.1. Appendix E.1 sketches the pseudo-code of SBART+SPL. Appendix E.2 explains the imputation of missing outcomes in the region of overlap (steps 4 and 5 of the algorithm) more in-depth while Appendix E.3 discusses the extrapolation of individual treatment effects into the non-overlap region.

E.1. Pseudo-code of SBART-SPL

Algorithm 2 SBART+SPL

Input :

- Initial parameters $\left\{\mathcal{T}_j^{(0)}, \mathcal{M}_j^{(0)}, {\sigma^2_{Imp}}^{(0)}, {\sigma^2_{\mu}}^{(0)}, \boldsymbol{s}^{(0)}, \alpha^{(0)}, \boldsymbol{\beta}_{Smo}^{(0)}, {\sigma^2_{Smo}}^{(0)}\right\}$
- Hyperparameters $\theta = (\gamma, \beta, \alpha_s, \beta_s, k, J, \upsilon, q_\lambda,)$ as specified in Appendix B.1 and Appendix C.1
- Observed data $\boldsymbol{Y}^{obs}$, $\boldsymbol{X}$ and region of overlap/non-overlap $O, O^{\neg}$

Output: $\tilde{\Delta}_P$ as m draws of the population average treatment effect from the posterior density

for $m \leftarrow 1$ **to** M **do**

1. **for** $j \leftarrow 1$ **to** J **do**
   - 1.1 Draw $\mathcal{T}_j^{(m)}$ from $p\left(\mathcal{T}_j | \tau_j^{bw\,(m)}, \boldsymbol{V}_{Oj}^{(m-1)}, {\sigma^2_{Imp}}^{(m-1)}\right)$ using the Metropolis Hastings algorithm described by Chipman et al. (1998)
   - 1.2 Draw $\tau_j^{bw\,(m)}$ from $p\left(\tau_j^{bw} | \mathcal{T}_j^{(m)}, \boldsymbol{V}_{Oj}^{(m-1)}, {\sigma^2_{Imp}}^{(m-1)}\right)$ using the Metropolis Hastings algorithm described by Chipman et al. (1998)
   - 1.3 Draw $\mathcal{M}_j^{(m)}$ from $p\left(\mathcal{M}_j | \boldsymbol{V}_{Oj}^{(m-1)}, \sigma^{2\,(m-1)}_{Imp}, \mathcal{T}_j^{(m)}\right)$ through a random sample from the Normal distribution

   end
2. Draw $\boldsymbol{s}^{(m)}$ through a random draw from a Dirichlet distribution
3. Draw noise variance $\sigma^{2\,(m)}_{Imp}$ and leaf variance parameters ${\sigma^2_{\mu}}^{(m)}$ from Inverse-Gamma distributions and the sparsity-control parameter $a^{(m)}$ using slice sampling (Neal, 2003)
4. Draw $\boldsymbol{Y}_O^{mis\,(m)}$ from $p\left(\boldsymbol{Y}_O^{mis} | \boldsymbol{Y}_O^{obs}, \mathcal{T}_1^{(m)}, ..., \mathcal{T}_J^{(m)}, \mathcal{M}_1^{(m)}, ..., \mathcal{M}_J^{(m)}, {\sigma^2_{Imp}}^{(m)}\right)$ through a random sample from a Normal distribution
5. Form $\tilde{\boldsymbol{\Delta}}_O^{(m)}$ as a linear combination of $\boldsymbol{Y}_O^{mis\,(m)}, \boldsymbol{Y}_O^{obs}$
6. Draw $\boldsymbol{\beta}_{Smo}^{(m)}$ from $p\left(\boldsymbol{\beta}_{Smo} | \tilde{\boldsymbol{\Delta}}_O^{(m)}, {\sigma^2_{Smo}}^{(m-1)}\right)$
7. Draw ${\sigma^2_{Smo}}^{(m)}$ from $p\left(\sigma^2_{Smo} | \tilde{\boldsymbol{\Delta}}_O^{(m)}, \boldsymbol{\beta}_{Smo}^{(m)}\right)$ through a random sample from an Inverse-Gamma distribution
8. Draw $\tilde{\boldsymbol{\Delta}}_{O^{\neg}}^{(m)}$ from $p\left(\tilde{\boldsymbol{\Delta}}_{O^{\neg}} | \tilde{\boldsymbol{\Delta}}_O^{(m)}, \boldsymbol{\beta}_{Smo}^{(m)}, {\sigma^2_{Smo}}^{(m)}\right)$ through a random sample from a Normal distribution
9. Draw $\hat{\Delta}_P^{(m)}$ by executing B iterations of the Bayesian bootstrap on $\left\{\tilde{\boldsymbol{\Delta}}_O^{(m)}, \tilde{\boldsymbol{\Delta}}_{O^{\neg}}^{(m)}\right\}$ and randomly selecting one of the B bootstrap sample averages

end

E.2. Imputation in the Region of Overlap

Similarly to the notation for $\boldsymbol{Y}_O^{obs}$ in Subsection 3.1.1, the corresponding vector of missing outcomes in the region of overlap is $\boldsymbol{Y}_O^{mis} = [Y_1^{mis}, ..., Y_{n_O}^{mis}]'$. As discussed in Section 2 related to SUTVA,

only one potential outcome is realized for each individual i. For individual o in the overlap region, the non-realized potential outcome is captured with $\boldsymbol{Y}_O^{mis}$ as a missing data point and needs to be imputed. The vector $\tilde{\boldsymbol{Y}}_O^{mis}$ collects the draws of the imputed values of $\boldsymbol{Y}_O^{mis}$ from its posterior predictive distribution. More precisely, a specific Y_o^{mis} is imputed with $\tilde{Y}_o^{mis}$ in step 4 by applying SoftBART to

$$Y_o^{obs} = \sum_j^J g\left(D_o, \widehat{p.sc}(\boldsymbol{X}_o), \boldsymbol{X}_o; \mathcal{T}_j, \mathcal{M}_j\right) + \epsilon_o, \epsilon_o \sim \mathcal{N}(0, \sigma^2_{Imp}), \tag{21}$$

similar to the usage in Equations (11) and (5). Here, the estimated propensity score $\widehat{p.sc}(\boldsymbol{X}_o)$ serves as another predictor while D_o lets us differentiate between treatment and control group. Let us collect parameters for the imputation phase in $\theta^{Imp} = \{\sigma^2_{Imp}, \mathcal{T}_1, ..., \mathcal{T}_J, \mathcal{M}_1, ..., \mathcal{M}_J\}$. This gives us a posterior distribution of $p\left(\theta^{Imp}|\boldsymbol{Y}_O^{obs}\right)$ which can be used to obtain draws from the posterior predictive distribution

$$p\left(\boldsymbol{Y}_O^{mis}|\boldsymbol{Y}_O^{obs}\right) = \int p\left(\boldsymbol{Y}_O^{mis}|\boldsymbol{Y}_O^{obs}, \theta^{Imp}\right) p\left(\theta^{Imp}|\boldsymbol{Y}_O^{obs}\right) \mathrm{d}\theta^{Imp}. \tag{22}$$

These draws are defined as imputed values of missing potential outcomes in the region of overlap, $\tilde{\boldsymbol{Y}}_O^{mis}$. Subsequently, the vector of estimated individual treatment effects for the overlap region O, $\tilde{\boldsymbol{\Delta}}_O = [\tilde{\Delta}_1, ..., \tilde{\Delta}_{n_O}]'$, is computed in step 5 of Algorithm 2 such that the individual treatment effect of a specific observation $o \in \{1, ..., n_O\}$ in the region of overlap is obtained as

$$\tilde{\Delta}_o = \begin{cases} Y_o^{obs} - \tilde{Y}_o^{mis}, \text{ if } D_o = 1 \\ \tilde{Y}_o^{mis} - Y_o^{obs}, \text{ if } D_o = 0 \end{cases}. \tag{23}$$

E.3. Extrapolation for the Non-overlap Region

Let $\boldsymbol{Y}_O^{mis}$ and $\boldsymbol{Y}_{O^\neg}^{mis}$ be the vectors of missing potential outcomes of individuals in the regions of overlap and non-overlap, respectively. Two smoothing models use restricted cubic splines $rcs(z)$ with basis z to extrapolate individual treatment effects into the region of non-overlap based on the linear model in (7) with $\boldsymbol{W}_o$ as

$$\boldsymbol{W}_o = \begin{cases} \left[rcs\left(\widehat{p.sc}(\boldsymbol{X}_o)\right), rcs\left(Y_o^\star(1)\right), \boldsymbol{X}_o\right]', \text{ if } D_o = 1 \\ \left[rcs\left(\widehat{p.sc}(\boldsymbol{X}_o)\right), rcs\left(Y_o^\star(0)\right), \boldsymbol{X}_o\right]', \text{ if } D_o = 0 \end{cases}. \tag{24}$$

With Equation (24), two smoothing models are defined: One for treated individuals in the non-overlap region ($D_{o^\neg} = 1$) and one for non-treated individuals in the non-overlap region ($D_{o^\neg} = 0$). For the former case, one uses

$$Y_o^\star(1) = \begin{cases} Y_o^{obs}, \text{ if } D_o = 1, \\ \tilde{Y}_o^{mis}, \text{ if } D_o = 0 \end{cases}, \tag{25}$$

with $\tilde{Y}_o^{mis}$ as in Equation (21). For the latter case, one similarly uses

$$Y_o^\star(0) = \begin{cases} Y_o^{obs}, \text{ if } D_o = 0, \\ \tilde{Y}_o^{mis}, \text{ if } D_o = 1 \end{cases}. \tag{26}$$

Equation (7) constitutes a linear regression model such that one can retrieve updates of $\theta^{Smo} = \left\{\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}\right\}$ by drawing from Normal and conjugate Inverse-Gamma distributions for each of the two smoothing models, separately. With the updates of these parameters in θ^{Smo}, one can draw from the posterior predictive distribution for the non-overlap region by

$$p\left(\tilde{\Delta}_{o^\neg}|\tilde{\boldsymbol{\Delta}}_O\right) = \int p\left(\tilde{\Delta}_{o^\neg}|\boldsymbol{W}^\star_{o^\neg}, \theta^{Smo}\right) p\left(\theta^{Smo}|\tilde{\boldsymbol{\Delta}}_O\right) \mathrm{d}\theta^{Smo}\,, \tag{27}$$

where the likelihood model reads

$$p\left(\tilde{\Delta}_{o^\neg}|\boldsymbol{W}^\star_{o^\neg}, \theta^{Smo}\right) = \mathcal{N}(\boldsymbol{W}^{\star\prime}_{o^\neg}\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}). \tag{28}$$

This draw is a sample from the normal distribution where $\boldsymbol{W}^\star_{o^\neg}$ stores information from the non-overlap region by

$$\boldsymbol{W}^\star_{o^\neg} = \begin{cases} \left[rcs\left(\widehat{p.sc}(\boldsymbol{X}_{o^\neg})\right)\,, rcs\left(Y^\star_{o^\neg}(1)\right)\,, \boldsymbol{X}_{o^\neg}\right]', \text{ if } D_{o^\neg} = 1 \\ \left[rcs\left(\widehat{p.sc}(\boldsymbol{X}_{o^\neg})\right)\,, rcs\left(Y^\star_{o^\neg}(0)\right)\,, \boldsymbol{X}_{o^\neg}\right]', \text{ if } D_{o^\neg} = 0 \end{cases}, \tag{29}$$

and $(\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo})$ generated from either of the two smoothing models described above. Hence, an extrapolated individual treatment effect for the non-overlap region, $\tilde{\Delta}_{o^\neg}$, is sampled from:

$$p\left(\tilde{\Delta}_{o^\neg} \mid \boldsymbol{W}^\star_{o^\neg}, \theta_{Smo}\right) = \mathcal{N}\left(\boldsymbol{W}^{\star\prime}_{o^\neg}\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}\right), \qquad (\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}) \sim p\left(\theta_{Smo} \mid \tilde{\boldsymbol{\Delta}}_O\right). \tag{30}$$

E.4. Bayesian Bootstrap

We use the Bayesian bootstrap in step 9 of Algorithm 2 to embrace variability in the choice of confounders and effect modifiers (Rubin, 1981; Wang et al., 2015; Nethery et al., 2019). For each $\hat{\Delta}^{(m)}_P$ in Algorithm 2, the sampled set of individual treatment effects is bootstrapped $B = 250$ times, and the B averages of these bootstrap draws are draws from the posterior distribution of the PATE. One random sample of these B averages is then termed $\hat{\Delta}^{(m)}_P$.

Appendix F. Additional Information for the Simulation Study

This section provides additional information for the main simulation study in Section 4 based on Nethery et al. (2019) (see Appendix F.1) and a second simulation study based on Krantsevich et al. (2023) and Wang et al. (2024) (see Appendix F.2). Code and data to replicate the results are accessible on GitHub via `https://github.com/LennMass/sbartspl`. We use the absolute bias and root-mean-squared error (RMSE), as well as credible interval coverage and width, to evaluate the precision and efficiency of the respective estimation methods across all simulations. If we use a trimmed version of a method, this method is abbreviated **T-method** and its untrimmed version is abbreviated **U-method**. Note that for the trimmed methods (T-GR, T-BART, T-SoftBART) the target estimand changes from the PATE to an ATE for the trimmed population that omits the observations in the region of non-overlap.

- **U-GR, T-GR**: The GR approach implements the method of multiple imputation with two subclassification splines (MITSS) proposed by Gutman and Rubin (2015).

- **U-BART, T-BART**: Both methods fit a single BART model with covariates, exposure variable and estimated propensity scores to the outcome variable. Afterwards, potential outcomes are predicted based on the posterior predictive distributions and used to compute the treatment effect (Hill, 2011; Nethery et al., 2019).
- **U-SoftBART, T-SoftBART**: The procedure is similar to the previous BART approach but uses SoftBART (Linero, 2022) instead of BART for modeling.
- **BLR**: We estimate the PATE using a Bayesian linear regression (BLR) model fitted with `brms` (Bürkner, 2017), regressing the observed outcome on treatment, continuous covariates, and categorical covariates under weakly informative priors. Posterior draws of individual potential outcomes under treatment and control are generated via Hamiltonian Monte Carlo, and the PATE posterior is computed as the sample-averaged contrast.
- **XBCF**: The Accelerated Bayesian Causal Forest (XBCF) is a Bayesian tree-based model for estimating heterogeneous treatment effects that decomposes outcomes into a prognostic component and a treatment effect component, each modeled via BART priors (Hahn et al., 2020; Krantsevich et al., 2023). The method incorporates estimated propensity scores into the prognostic function and uses a split rule that prevents treatment trees from splitting when only treated or control units remain, making overlap explicit in the tree structure. This design captures treatment effect heterogeneity well in regions of overlap while highlighting non-overlap regions. The code for the recommended extension in Wang et al. (2024), XBCF with Gaussian Process (GP) priors within leaf nodes (XBCF-GP), is not available to the author and is therefore not used for comparison. However, the differences in performance for PATE estimation between XBCF and XBCF-GP presented in the simulation study of Wang et al. (2024) are only minor.
- **BART+SPL**: The original two-stage Bayesian modeling approach proposed by Nethery et al. (2019).
- **SBART+SPL**: The extension of BART+SPL proposed in this paper and mainly described in Section 3.1 and Appendix E.

The simulation study follows the notation of Zhu et al. (2023) such that

$$\hat{\Delta}_P^{(sim)} = \frac{1}{M} \sum_{m=1}^{M} \hat{\Delta}_P^{(m)}, \tag{31}$$

where $\hat{\Delta}_P^{(sim)}$ is an estimate of the population average treatment effect for the specific dataset $sim \in \{1, ..., Sim\}$ with Sim being the total number of simulated datasets. We evaluate the absolute bias and RMSE over all simulations by

$$|\text{Bias}| = \frac{1}{Sim} \sum_{sim=1}^{Sim} \left| \hat{\Delta}_P^{(sim)} - ace_{true}^{(sim)} \right|, \tag{32}$$

$$\text{RMSE} = \sqrt{\frac{1}{Sim} \sum_{sim=1}^{Sim} \left(\hat{\Delta}_P^{(sim)} - ace_{true}^{(sim)} \right)^2}, \tag{33}$$

with the true PATE for each simulation dataset as $ace_{true}^{(sim)} = \frac{1}{n}\sum_{i=1}^{n}\tau_i$ based on the true potential outcomes for each observation i with $\tau_i = Y_i(1) - Y_i(0)$. Besides precision, coverage rates evaluate the efficiency of the respective estimation methods. For each simulation dataset, the indicator function checks if the true PATE lies in $CI_{95}^{(sim)}$, defined as the 95% credible interval based on the posterior draws $\hat{\Delta}_P^{(m)}$. Moreover, the average width of the 95% credible interval indicates the tightness of the intervals.

$$\text{Coverage}_{95\%-CI} = \frac{1}{Sim}\sum_{sim=1}^{Sim}\mathbb{1}\left(ace_{true}^{(sim)} \in CI_{95}^{(sim)}\right), \tag{34}$$

$$\text{Width}_{95\%-CI} = \frac{1}{Sim}\sum_{sim=1}^{Sim}\text{length}\left(CI_{95}^{(sim)}\right). \tag{35}$$

F.1. Data-generating Process based on Nethery et al. (2019)

The main simulation study in Section 4 generates the true confounder variables by

$$X_{1i}|D_i = 1, ..., X_{5i}|D_i = 1 \sim \text{Bernoulli}(0.45), \tag{36}$$
$$X_{1i}|D_i = 0, ..., X_{5i}|D_i = 0 \sim \text{Bernoulli}(0.4), \tag{37}$$
$$X_{6i}|D_i = 1, ..., X_{10i}|D_i = 1 \sim \mathcal{N}(2, 4), \tag{38}$$
$$X_{6i}|D_i = 0, ..., X_{10i}|D_i = 0 \sim \mathcal{N}(1.3, 1). \tag{39}$$

Based on the true confounder variables from above, we produce the two potential outcomes for each individual $i = 1, ..., n$ with $n \in \{500, 2000\}$:

$$\begin{aligned} Y_i(0) &= 0.5\left(X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i}\right) + \\ &\quad \frac{15}{(1 + \exp\{-8X_{6i} + 1\})} + X_{7i} + X_{8i} + X_{9i} + X_{10i} - 5, \\ Y_i(1) &= X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i} - \\ &\quad 0.5\left(X_{6i} + X_{7i} + X_{8i} + X_{9i} + X_{10i}\right). \end{aligned} \tag{40}$$

Figure 5 illustrates a distribution of propensity score estimates by treatment status for $ncov = 10$ and $n = 500$. The graphic indicates that the region of non-overlap is non-negligible.

Table 3 shows the complete version of Table 1 in Section 4 with all benchmark models described in Appendix F. The comparison of results reveals that SBART+SPL delivers a compelling balance between point estimation accuracy and uncertainty quantification. On the accuracy side, BLR attains the lowest RMSE and bias across all settings due to the mostly linear potential outcomes in (40), but its coverage deteriorates at larger sample sizes, falling to 0.71–0.84 at $n = 2000$ and thus departing from the nominal 95% level. The trimmed variants (T-GR, T-BART, T-SoftBART) are uniformly dominated: their RMSE values are two to three times those of SBART+SPL, and their coverage collapses to 0.00–0.08 in nearly every scenario, indicating severe miscalibration. Among the untrimmed models, U-SoftBART consistently overcovers the nominal coverage rate (except for $n = 2000$ and $ncov = 50$) and shows higher RMSE at larger sample sizes (i.e., 0.429 vs. 0.379 at $ncov = 25$, $n = 2000$). U-BART and U-GR, meanwhile, combine moderate-to-high bias with poor coverage (0.05–0.47), offering no dimension along which they dominate SBART+SPL. In contrast,

SBART+SPL maintains coverage between 0.78 and 0.98 across all configurations while exhibiting RMSE and bias that are competitive with or superior to every method except BLR. Furthermore, SBART+SPL shows stable or improving performance as the number of covariates increases from 10 to 25, suggesting that the SoftBART component within SBART+SPL accommodates growing covariate dimensionality.

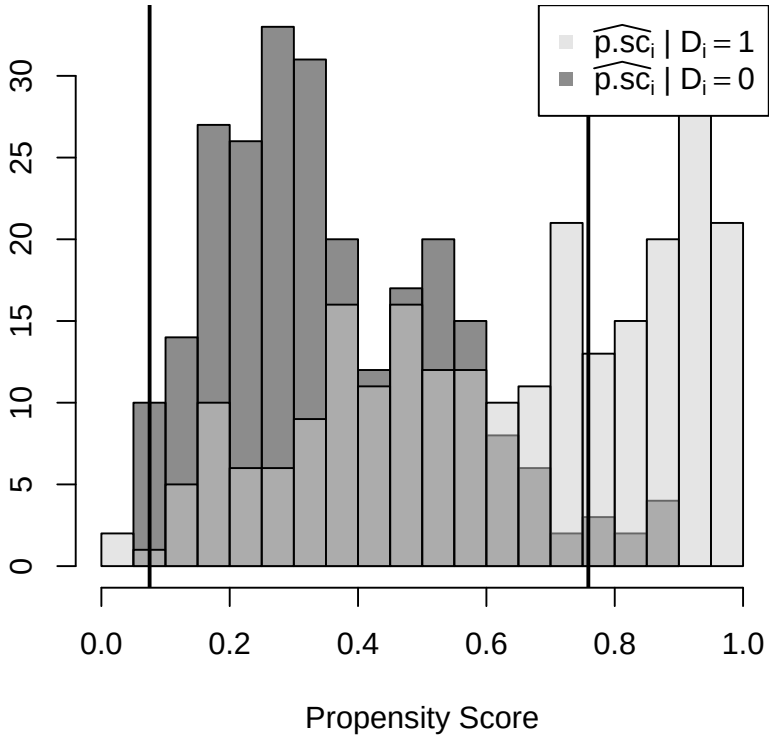

Figure 5: Example of a distribution of propensity score estimates by treatment status for $ncov = 10, n = 500$ based on the data-generating process in Appendix F.1. Regions of overlap and non-overlap are indicated by the vertical straight lines. The region of overlap lies in-between these two vertical straight lines.

Table 3: Model comparison of PATE estimates in terms of RMSE, absolute bias, and 95% credible interval coverage and width for varying numbers of covariates $ncov \in \{10, 25, 50\}$ and sample sizes $n \in \{500, 2000\}$ across 120 simulation runs based on the data-generating process in Appendix F.1.

		$n = 500$				$n = 2000$			
$ncov$	Method	RMSE	|Bias|	Cov.	Width	RMSE	|Bias|	Cov.	Width
10	U-GR	0.710	0.596	0.47	0.971	0.464	0.415	0.18	0.477
	U-BART	0.741	0.719	0.30	1.229	0.411	0.389	0.30	0.640
	U-SoftBART	0.505	0.490	1.00	1.646	0.337	0.324	1.00	1.660
	T-GR	1.285	1.221	0.02	0.695	0.769	0.744	0.00	0.357
	T-BART	1.862	1.831	0.00	1.233	1.146	1.131	0.00	0.622
	T-SoftBART	1.865	1.836	0.00	1.381	1.214	1.198	0.00	0.877
	XBCF	0.759	0.741	0.31	1.324	0.768	0.741	0.00	0.711
	BLR	0.397	0.333	0.97	1.478	0.272	0.251	0.83	0.718
	BART+SPL	1.614	1.307	0.51	2.385	0.441	0.365	0.60	0.783
	SBART+SPL	0.702	0.584	0.82	1.842	0.344	0.282	0.98	1.493
25	U-GR	0.682	0.568	0.45	0.946	0.461	0.405	0.20	0.478
	U-BART	0.891	0.879	0.05	1.258	0.489	0.477	0.10	0.686
	U-SoftBART	0.569	0.560	0.99	1.594	0.429	0.417	0.98	1.349
	T-GR	1.245	1.169	0.04	0.751	0.757	0.724	0.00	0.362
	T-BART	1.879	1.837	0.00	1.250	1.218	1.204	0.00	0.626
	T-SoftBART	1.836	1.791	0.00	1.428	1.244	1.231	0.00	0.865
	XBCF	0.710	0.690	0.39	1.335	0.779	0.749	0.01	0.730
	BLR	0.360	0.309	0.99	1.520	0.280	0.258	0.84	0.722
	BART+SPL	1.457	1.175	0.47	2.060	0.485	0.374	0.60	0.736
	SBART+SPL	0.631	0.509	0.80	1.681	0.379	0.323	0.95	1.378
50	U-GR	0.722	0.624	0.39	0.973	0.542	0.501	0.10	0.487
	U-BART	1.071	1.052	0.02	1.297	0.597	0.589	0.01	0.786
	U-SoftBART	0.586	0.576	0.99	1.630	0.534	0.524	0.73	1.284
	T-GR	1.249	1.162	0.08	0.806	0.789	0.769	0.00	0.376
	T-BART	1.921	1.882	0.00	1.290	1.256	1.245	0.00	0.650
	T-SoftBART	1.786	1.741	0.01	1.480	1.262	1.252	0.00	0.878
	XBCF	0.659	0.639	0.60	1.337	0.784	0.752	0.00	0.703
	BLR	0.403	0.342	0.98	1.569	0.324	0.300	0.71	0.737
	BART+SPL	1.392	1.147	0.43	1.899	0.475	0.382	0.53	0.692
	SBART+SPL	0.696	0.568	0.78	1.747	0.426	0.382	0.90	1.253

F.2. Data-generating Process based on Wang et al. (2024)

We follow Wang et al. (2024) and use the data-generating process presented in Hahn et al. (2020), Krantsevich et al. (2023), and Wang et al. (2024) to analyze the PATE based on either homogeneous or heterogeneous conditional treatment effects while increasing the number of irrelevant covariates. We generate two potential outcomes that rely on five covariates $X_{1i}, ..., X_{5i}$ where the first three covariates are drawn independently from a standard normal distribution. X_{4i} represents an unordered categorical variable with levels $(1, 2, 3)$, while X_{5i} is a binary variable. The prognostic component,

$\mu_{\text{prog}}(\boldsymbol{X_i})$, and the treatment effect function, $\tau(\boldsymbol{X_i})$, that determine the potential outcomes are given below:

$$Y_i(0) = \mu_{\text{prog}}(\boldsymbol{X_i}), \tag{41}$$

$$Y_i(1) = \mu_{\text{prog}}(\boldsymbol{X_i}) + \tau(\boldsymbol{X_i}), \tag{42}$$

$$\mu_{\text{prog}}(\boldsymbol{X_i}) = -6 + g_{\text{prog}}(X_{4i}) + 6|X_{3i} - 1|, \tag{43}$$

$$\tau(\boldsymbol{X_i}) = \begin{cases} 3, & \text{if} \quad \text{homogeneous}, \\ 1 + 2X_{2i}X_{5i}, & \text{if} \quad \text{heterogeneous}. \end{cases} \tag{44}$$

$$g_{\text{prog}}(X_{4i}) = \begin{cases} 2, & \text{if} \quad X_{4i} = 1, \\ -1, & \text{if} \quad X_{4i} = 2, \\ 4, & \text{if} \quad X_{4i} = 3. \end{cases} \tag{45}$$

Treatment assignment is generated using a nonlinear propensity score model designed to induce complex covariate overlap. Specifically, the propensity score is defined as

$$p.sc(\boldsymbol{X}_i) = \min\left\{\max\left\{\text{logit}^{-1}(\eta(\boldsymbol{X}_i),\, 0.01\right\},\, 0.99\right\}, \tag{46}$$

where the linear predictor $\eta(\boldsymbol{X_i})$ is given by

$$\eta(\boldsymbol{X_i}) = 2\left(X_{1i}^2 - 0.5\right)^2 - 1.5\left(X_{2i}^2 - 0.2\right)^2 + 0.8\sin(\pi X_{3i}) + \Phi(X_{4i} - X_{5i})\,. \tag{47}$$

In Equation (46), $\text{logit}^{-1}(u) = (1 + \exp\{-u\})^{-1}$ denotes the logistic function and $\Phi(\cdot)$ in Equation (47) is the standard normal cumulative distribution function. The truncation to the interval $[0.01, 0.99]$ is applied to avoid excessively extreme propensity scores that equal 0 or 1. Besides the five mentioned covariates that generate potential outcomes and propensity scores, we consider $ncov \in \{10, 20\}$ irrelevant covariates with $ncov = ncov_{cont} + ncov_{cat}$ and $ncov_{cont} \in \{5, 10\}$ the number of continuous variables and $ncov_{cat} \in \{5, 10\}$ the number of categorical variables. All continuous covariates follow a standard normal distribution, while half of the categorical variables are binary and the other half are unordered variables with levels (1,2,3). We generate 20 simulation datasets that have regions of non-overlap between 10% and 30%.

Notably, the simulation design in this section differs from the simulation study in Appendix F.1 in several dimensions. Instead of relying on five continuous and five binary covariates, the true individual treatment effects in this study are either homogeneous or depend only on one continuous and one categorical variable. Moreover, treatment effect heterogeneity arises implicitly from differences in covariates in Appendix F.1 while effect heterogeneity in $\tau(\boldsymbol{X}_i)$ is either explicitly modeled (heterogeneous case) or absent (homogeneous case). The simulation study in Appendix F.1 tests the ability to handle weak overlap induced by distributional shifts in covariates based on treatment status. This leads to implicit treatment assignment, and regions of overlap and non-overlap arise from covariate distribution shifts. In comparison to that, Equation (46) in this study explicitly determines the treatment assignment using a non-linear propensity score function.

Table 4 displays the estimates of $\hat{\Delta}_P$ for SBART+SPL and all competing methods described in Appendix F. Across both homogeneous and heterogeneous settings, Table 4 shows that SBART+SPL consistently achieves lower RMSE and absolute bias than BART+SPL, with comparable or improved 95% credible interval coverage, indicating more accurate and stable PATE estimation under

weak overlap. Compared to XBCF, SBART+SPL is slightly less competitive in RMSE in some configurations and delivers similar coverage with moderately wider intervals, reflecting reliable uncertainty quantification.

Compared to Table 3 in Appendix F.1, performance differences across methods are narrower in Table 4. Within this compressed range, SBART+SPL remains consistently competitive. Under homogeneous treatment effects, U-SoftBART achieves the lowest RMSE in both covariate settings. The key differentiator, however, is the transition to underlying heterogeneous treatment effects: BLR degrades substantially in terms of precision, and T-GR deteriorates sharply at higher dimensions (RMSE 0.544 at $ncov = 20$), whereas SBART+SPL's RMSE increases only modestly (from 0.145 to 0.187 and from 0.159 to 0.203), maintaining overcoverage at 1.00 in both heterogeneous scenarios. U-SoftBART and XBCF are similarly robust to this transition.

Table 4: Model comparison of PATE estimates in terms of RMSE, absolute bias, and 95% credible interval coverage and width for sample size $n = 500$ and varying numbers of irrelevant covariates $ncov \in \{10, 20\}$ across 20 simulation runs based on the data-generating process in Appendix F.2. Note that for some simulation runs at $ncov = 20$, PATE estimates for T-GR could not be computed leading to coverage rates of 0.88 and 0.94, respectively.

		$\tau(\boldsymbol{X}_i)$: homogeneous				$\tau(\boldsymbol{X}_i)$: heterogeneous			
$ncov$	Method	RMSE	|Bias|	Cov.	Width	RMSE	|Bias|	Cov.	Width
10	U-GR	0.210	0.154	0.85	0.604	0.230	0.168	0.85	0.745
	U-BART	0.179	0.138	0.95	0.574	0.188	0.159	1.00	0.780
	U-SoftBART	0.128	0.098	0.95	0.452	0.166	0.130	1.00	0.793
	T-GR	0.199	0.155	0.95	0.586	0.354	0.258	0.85	0.851
	T-BART	0.175	0.133	0.90	0.603	0.174	0.143	0.90	0.822
	T-SoftBART	0.146	0.111	0.95	0.475	0.180	0.131	0.95	0.896
	XBCF	0.152	0.120	1.00	0.539	0.187	0.154	0.95	0.764
	BLR	0.147	0.119	1.00	0.659	0.299	0.232	0.90	0.923
	BART+SPL	0.188	0.147	1.00	0.779	0.210	0.176	0.95	0.974
	SBART+SPL	0.145	0.110	1.00	0.660	0.187	0.145	1.00	0.843
20	U-GR	0.199	0.180	0.90	0.793	0.280	0.256	0.90	0.966
	U-BART	0.176	0.147	0.95	0.724	0.190	0.152	1.00	0.956
	U-SoftBART	0.128	0.082	0.95	0.566	0.207	0.171	1.00	0.910
	T-GR	0.233	0.179	0.88	0.919	0.544	0.352	0.94	1.282
	T-BART	0.202	0.162	0.95	0.749	0.258	0.211	0.90	1.007
	T-SoftBART	0.141	0.110	1.00	0.624	0.224	0.170	1.00	1.025
	XBCF	0.117	0.097	1.00	0.605	0.204	0.158	1.00	0.888
	BLR	0.141	0.116	1.00	0.773	0.354	0.245	0.90	1.096
	BART+SPL	0.181	0.153	1.00	1.094	0.223	0.176	1.00	1.360
	SBART+SPL	0.159	0.110	0.90	0.689	0.203	0.164	1.00	1.118

Appendix G. Analysis of the Variance Inflation Parameter

We illustrate the influence of the variance inflation parameter in BART+SPL in comparison to SBART+SPL. In Nethery et al. (2019), the recommended default specification for the variance inflation parameter in BART+SPL uses $\text{var}_{\text{infl}} = 10 \cdot \text{dist}_{\widehat{p.sc}(\boldsymbol{X}_o)} \cdot \text{range}(\tilde{\boldsymbol{\Delta}}_O)$. Here, $\text{dist}_{\widehat{p.sc}(\boldsymbol{X}_o)}$ is the distance of an observation's estimated propensity score to the nearest estimated propensity score in the overlap region, while $\text{range}(\tilde{\boldsymbol{\Delta}}_O)$ is the range of the estimated individual effects in the overlap region. For SBART+SPL, we do not use this unidentifiable variance inflation hyperparameter to properly account for the increase in variance in the region of non-overlap, which is equivalent to using $\text{var}_{\text{infl}} = 0$ in (8) by default. As a third model, we compare BART+SPL and SBART+SPL with BART+SPL using $\text{var}_{\text{infl}} = 0$. Table 5 shows a comparison of the three models for PATE estimation using $n = 500$ and $ncov = 10$ for the data-generating process in Appendix F.1 across 20 simulation runs. The results show that SBART+SPL outperforms both versions of BART+SPL in terms of RMSE and absolute bias, reducing error by more than half. SBART+SPL also achieves higher 95% credible interval coverage compared to BART+SPL, indicating more reliable uncertainty quantification. Additionally, SBART+SPL produces slightly narrower intervals than BART+SPL, suggesting improved precision without sacrificing coverage. Moreover, the default variance inflation parameter of BART+SPL improves uncertainty quantification only slightly by sacrificing precision compared to BART+SPL without variance inflation ($\text{var}_{\text{infl}} = 0$).

In Table 6 and Figure 6, we focus on one randomly drawn simulation dataset to further investigate the behavior of the variance inflation parameter at the individual treatment effect level. The results in Table 6 indicate that in the overlap region, all three methods perform similarly, with comparably low RMSE and absolute bias, reflecting accurate individual treatment effect estimation where data support is strong. In the non-overlap region, SBART+SPL outperforms both BART+SPL variants, decisively reducing RMSE and bias. While variance inflation in BART+SPL improves coverage at the cost of extremely large credible intervals, SBART+SPL yields more precise individual treatment effect estimates with the shortest average interval length in the non-overlap region, which translates into an increase in precision and coverage for PATE estimation presented in Table 5. Figure 6 graphically supports the findings in Table 6 for the regions of overlap and non-overlap.

Table 5: Comparison of PATE estimates in terms of RMSE, absolute bias, and 95% credible interval coverage and width for the data-generating process in Appendix F.1 with sample size $n = 500$ and $ncov = 10$ across 20 simulation runs.

Method	RMSE	|Bias|	Cov.	Width
BART+SPL	1.8253	1.3100	0.55	2.1368
BART+SPL (without var_{infl})	1.7186	1.2436	0.50	2.1297
SBART+SPL (without var_{infl})	0.7621	0.6465	0.80	1.9091

Table 6: Model comparison of individual treatment effect estimates for one simulation dataset in terms of RMSE, absolute bias, and 95% credible interval coverage and width for the data-generating process in Appendix F.1 with sample size $n = 500$ and $ncov = 10$.

Region	Method	RMSE	|Bias|	Cov.	Width
Overlap	BART+SPL	0.874	0.565	0.88	2.136
	BART+SPL (without var_{infl})	0.797	0.517	0.92	2.131
	SBART+SPL (without var_{infl})	0.822	0.555	0.99	4.660
Non-Overlap	BART+SPL	13.269	9.891	0.69	22.335
	BART+SPL (without var_{infl})	14.564	10.875	0.11	7.897
	SBART+SPL (without var_{infl})	4.581	3.716	0.51	7.827

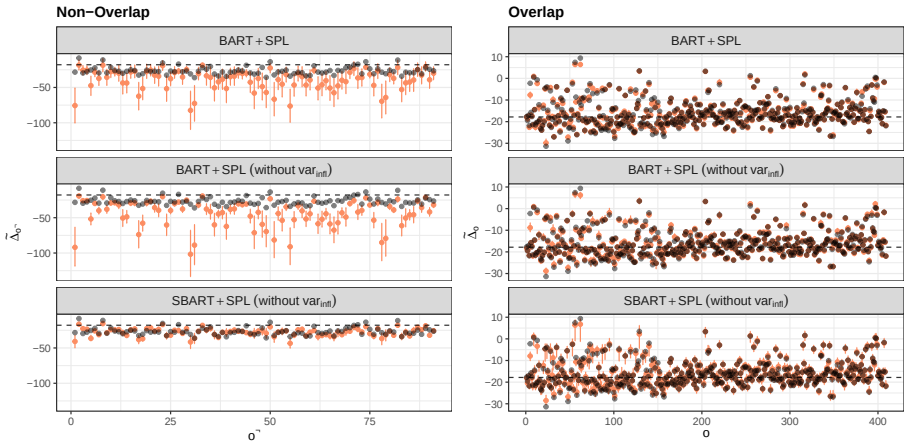

Figure 6: Estimated individual treatment effects with 95% credible intervals for the non-overlap region (left panel, $\tilde{\Delta}_{o^\neg}$), and for the overlap region (right panel, $\tilde{\Delta}_o$), for three methods: BART+SPL, BART+SPL (without var_{infl}), and SBART+SPL (without var_{infl} by default). Coloured points and error bars denote posterior point estimates and credible interval estimates for each unit. Black points indicate the true individual treatment effects. The dashed black line marks the true PATE. Results are based on a simulated dataset with $n = 500$ and $ncov = 10$ irrelevant covariates from Appendix F.1.

Appendix H. Additional Information for the Empirical Application

This section presents descriptive statistics for the empirical application in Section 5. Table 7 shows descriptive statistics for the four outcome variables, Table 8 gives summary statistics for the considered continuous covariate variables, and Figure 7 illustrates the distribution of propensity score estimates by treatment status. Table 9 and Table 10 present the complete comparison of PATE estimates for all competitor models explained in Appendix F. Code and data to replicate the results are accessible on GitHub via https://github.com/LennMass/sbartspl.

Table 7: County-level ($n = 978$) descriptive statistics of the four outcome variables. 2014 mortality rates are measured as deaths per $100,000$ population, the changes in mortality rates are measured as percentage points.

Variable	Mean	St. Dev.	Min	Max
leukemia mortality rate				
2014 (log-transformed)	9.56 (2.25)	0.10 (0.11)	4.17 (1.43)	16.55 (2.81)
change from 1980 to 2014	-2.02	8.98	-42.94	40.14
thyroid cancer mortality rate				
2014 (log-transformed)	0.56 (-0.60)	0.06 (0.09)	0.30 (-0.89)	0.930 (-0.07)
change from 1980 to 2014	1.71	8.03	-20.11	29.06

Table 8: County-level ($n = 978$) descriptive statistics of continuous covariate variables.

Variable	Mean	St. Dev.	Min	Max
population per prim. care physician	801.39	1,330.09	60	13,664
less than 65 year olds uninsured (%)	17.74	5.53	4.72	38.85
diabetic (%)	83.02	8.14	25.46	100.00
current smokers (%)	20.51	6.17	6.60	49.20
limited access to healthy foods (%)	10.18	8.19	0.00	62.94
obese (%)	28.43	5.19	10.40	43.50
food environment index	7.31	1.32	0.61	9.84
population per mi^2	99.38	332.66	0.52	5,144.64
male (%)	50.10	1.89	45.02	67.60
less than age 55 (%)	69.61	6.17	45.07	86.12
white (%)	84.88	16.39	10.73	99.45
avg. household size	2.50	0.25	1.95	4.05
with bachelor's degree or higher (%)	20.01	7.80	6.23	64.01
unemployed (%)	7.00	3.69	0.61	28.98
median household income (US-$)	46,296.10	10,436.37	21,399	105,989
Gini index of inequality	0.44	0.03	0.34	0.56
owner-occupied housing units (%)	71.85	7.44	40.70	89.99
median rent as proportion of income (%)	27.43	4.43	10.00	44.70
avg. commute time to work (minutes)	21.28	5.30	10	42

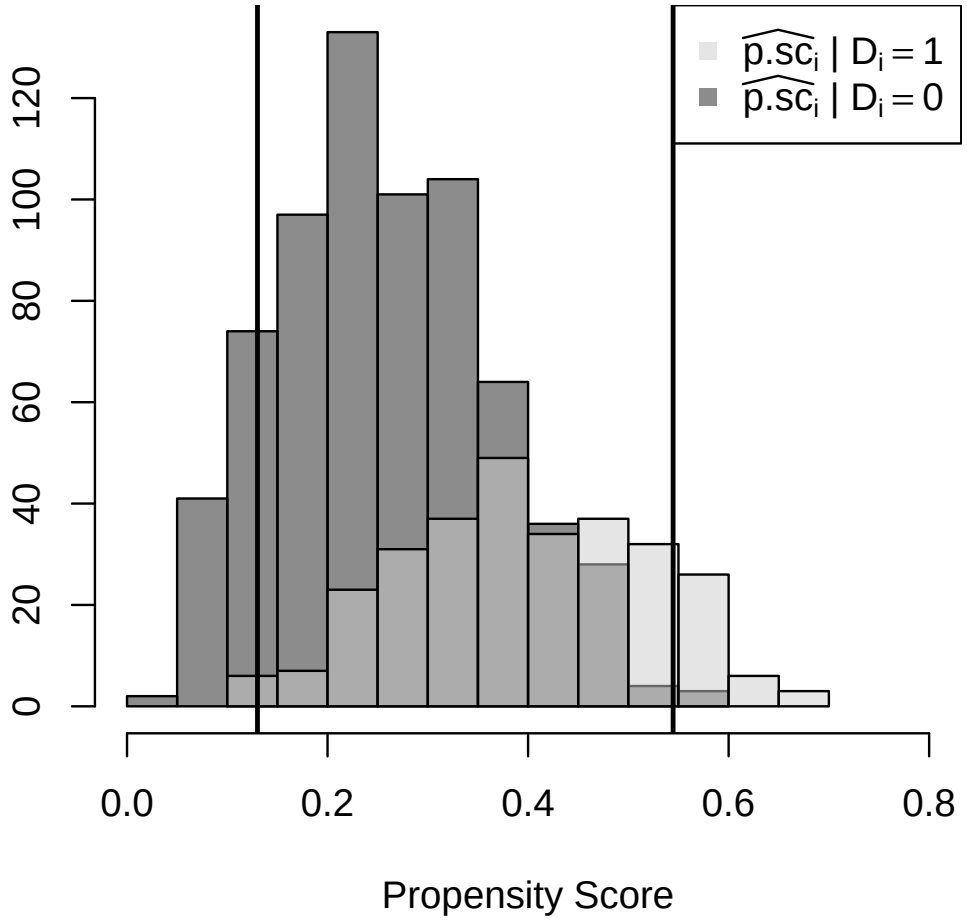

Figure 7: Distribution of propensity score estimates by treatment status for the empirical application in Section 5 based on Nethery et al. (2019). Regions of overlap and non-overlap are indicated by the vertical straight lines.

Table 9: PATE estimates and 95% credible intervals for outcomes (1) leukemia log–2014 mortality, and (2) thyroid log–2014 mortality. Section 5 gives more detailed information about the empirical application.

	(1) leukemia				(2) thyroid cancer			
Method	Effect	Lower	Upper	Width	Effect	Lower	Upper	Width
U-GR	0.010	-0.002	0.023	0.025	-0.015	-0.026	-0.003	0.022
U-BART	0.003	-0.008	0.015	0.023	0.002	-0.008	0.012	0.020
U-SoftBART	0.000	-0.008	0.008	0.016	0.000	-0.009	0.008	0.017
T-GR	0.011	-0.001	0.023	0.024	-0.014	-0.023	-0.003	0.020
T-BART	0.006	-0.006	0.016	0.022	0.002	-0.010	0.013	0.023
T-SoftBART	0.001	-0.008	0.010	0.017	0.001	-0.007	0.010	0.018
XBCF	0.001	-0.004	0.007	0.011	0.001	-0.002	0.005	0.007
BLR	0.008	-0.004	0.020	0.024	0.005	-0.007	0.017	0.024
BART+SPL	-0.004	-0.026	0.018	0.044	-0.007	-0.030	0.015	0.045
SBART+SPL	0.001	-0.007	0.009	0.016	0.001	-0.008	0.010	0.017

Table 10: PATE estimates and 95% credible intervals for outcomes (3) leukemia change 1980–2014 and (4) thyroid cancer change 1980–2014. Section 5 gives more detailed information about the empirical application.

	(3) leukemia change				(4) thyroid cancer change			
Method	Effect	Lower	Upper	Width	Effect	Lower	Upper	Width
U-GR	1.773	0.488	3.054	2.566	0.621	-0.521	1.809	2.331
U-BART	0.810	-0.250	1.884	2.135	0.995	-0.144	2.115	2.259
U-SoftBART	0.564	-0.494	1.717	2.211	0.636	-0.438	2.005	2.443
T-GR	1.965	0.895	3.169	2.274	0.836	-0.248	1.849	2.097
T-BART	0.960	-0.224	2.059	2.283	1.068	-0.083	2.258	2.341
T-SoftBART	0.722	-0.382	1.944	2.325	0.889	-0.257	2.135	2.391
XBCF	0.852	-0.194	1.938	2.132	1.115	0.267	2.092	1.825
BLR	1.085	-0.093	2.263	2.356	1.233	0.022	2.501	2.480
BART+SPL	0.017	-1.832	1.841	3.673	0.916	-0.898	2.665	3.563
SBART+SPL	0.637	-0.444	1.842	2.287	0.667	-0.350	2.012	2.363