Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Shrinkage Bayesian Causal Forests for Heterogeneous Treatment Effects Estimation

Alberto Caron, Gianluca Baio & Ioanna Manolopoulou

To cite this article: Alberto Caron, Gianluca Baio & Ioanna Manolopoulou (2022): Shrinkage Bayesian Causal Forests for Heterogeneous Treatment Effects Estimation, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2067549

To link to this article: https://doi.org/10.1080/10618600.2022.2067549

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/10618600.2022.2067549

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Shrinkage Bayesian Causal Forests for Heterogeneous Treatment Effects Estimation

Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou

Department of Statistical Science, University College London, London, UK

ABSTRACT
This article develops a sparsity-inducing version of Bayesian Causal Forests, a recently proposed nonparametric causal regression model that employs Bayesian Additive Regression Trees and is specifically designed to estimate heterogeneous treatment effects using observational data. The sparsity-inducing component we introduce is motivated by empirical studies where not all the available covariates are relevant, leading to different degrees of sparsity underlying the surfaces of interest in the estimation of individual treatment effects. The extended version presented in this work, which we name Shrinkage Bayesian Causal Forest, is equipped with an additional pair of priors allowing the model to adjust the weight of each covariate through the corresponding number of splits in the tree ensemble. These priors improve the model's adaptability to sparse data generating processes and allow to perform fully Bayesian feature shrinkage in a framework for treatment effects estimation, and thus to uncover the moderating factors driving heterogeneity. In addition, the method allows prior knowledge about the relevant confounding covariates and the relative magnitude of their impact on the outcome to be incorporated in the model. We illustrate the performance of our method in simulated studies, in comparison to Bayesian Causal Forest and other state-of-the-art models, to demonstrate how it scales up with an increasing number of covariates and how it handles strongly confounded scenarios. Finally, we also provide an example of application using real-world data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2021
Accepted April 2022

KEYWORDS
Bayesian nonparametrics; Causal inference; Heterogeneous treatment effects; Observational studies; Machine learning; Tree ensembles

1. Introduction

Inferring the treatment effect at an individual level in a population of interest lies at the heart of disciplines such as precision medicine or personalized advertisement, where decision making in terms of treatment administration is based on individual characteristics. The ever-increasing amount of observational data available offers a unique opportunity for drawing inferences at the resolution of each individual. However, since Individual Treatment Effects (ITEs) are never directly observable in the real world, standard supervised learning techniques cannot be directly applied. Moreover, the process of treatment allocation in large, observational datasets is usually unknown and can obscure the effect of the actual treatment through confounding (Dawid 2000; Pearl 2009a; Imbens and Rubin 2015).

The application of statistical learning tools (Hastie, Tibshirani, and Friedman 2001) for causal inference has led to significant improvements in the estimation of heterogeneous treatment effects. These improvements stem from the predictive power of advanced nonparametric regression models, that, after being appropriately adapted to the causal inference setting, can leverage large observational datasets and capture nonlinear relationships. Caron, Manolopoulou, and Baio (2020) provide a review of the most recent and popular methods, together with a comparison of their performance.

Two of the early contributions that paved the way toward the use of tree-based statistical learning tools on large datasets for causal analysis purposes are Foster, Taylor, and Ruberg (2011) and Hill (2011), who advocate the use of tree ensemble methods for the estimation of ITE. The former focuses on randomized experiments and makes use of Random Forests (Breiman 2001; Lu et al. 2018); the latter instead addresses the problem from an observational study perspective, and employs Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 1998, 2010). Another early contribution that focuses instead on Average Treatment Effect (ATE) estimation in observational studies is Traskin and Small (2011), where the authors propose classification trees for identifying the study population with sufficient overlap. A more recent and popular tree-based method for ITE estimation is Causal Forests (CF) (Athey and Imbens 2016; Wager and Athey 2018), a causal implementation of Random Forests. Hahn, Murray, and Carvalho (2020) instead build on the work of Chipman, George, and McCulloch (2010) and Hill (2011) to formulate a new BART framework for causal analysis, under the name of Bayesian Causal Forests (BCF), specifically designed to address strong confounding and separate between prognostic and moderating effects of the covariates when estimating ITE. The prognostic effect (or prognostic score) is defined as the impact of the covariates on the outcome in the absence of treatment, while the moderating effect is the impact of the covariates on the response to treatment.

A different stream of contributions that do not focus on any specific regression model is that of *Meta-Learners*. Meta-Learners are meta-algorithms that design a procedure to estimate ITE via any suitable off-the-shelf supervised regression

CONTACT Alberto Caron *alberto.caron.19@ucl.ac.uk* Torrington Pl, London WC1E 7HB, UK.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

model (e.g., random forests, neural networks, etc.). Recent popular work on Meta-Learners include Künzel et al. (2017), where the authors develop a framework to deal with unbalanced treatment groups (named X-learner), and Nie and Wager (2020), where parameterization in Robinson (1988) is exploited (hence, the name R-learner) to design a direct loss function on the treatment effect surface for parameter tuning. Among other notable works on ITE estimation, Alaa and van der Schaar (2017, 2018) adopt a multi-task learning approach using Gaussian Processes, while Johansson, Shalit, and Sontag (2016), Shalit, Johansson, and Sontag (2017), and Yao et al. (2018) employ deep neural networks to learn balanced representations that aim at minimizing a distributional distance between treatment groups. Moreover, contributions such as those of Powers et al. (2018); Zhao, Small, and Ertefaie (2018); Zimmert and Lechner (2019), and Fan et al. (2020) focus specifically on high-dimensional settings, where a large number of covariates is available. In particular, Zhao, Small, and Ertefaie (2018) use Robinson (1988) decomposition to estimate nuisance parameters with machine learning methods and isolate the treatment effect function, which is then fit via LASSO regression, for a more interpretable output on effect modifiers. Zimmert and Lechner (2019) and Fan et al. (2020) instead propose and derive properties of a cross-validated two stage estimator where nuisance parameters are fitted with ML methods in the first stage, and a nonparametric local kernel smoother is instead applied to fit treatment effect.

Regression-based methods for treatment effect estimation typically leverage large samples in observational studies. However, large observational data often also feature a large number of pretreatment covariates, many of which may not affect the response variable in question nor act as a modifier of the treatment effect. Hence, the task of estimating ITE, whose complexity inevitably depends on the smoothness and sparsity of the outcome surface (Alaa and van der Schaar 2018), necessitates regularization. At the same time, prior subject-matter knowledge on the relative importance of the covariates may be available, and can improve estimates if embedded in the model. In light of these considerations, none of the aforementioned approaches mentioned allow to jointly: (i) account for heterogeneous smoothness and sparsity across covariates; (ii) tease apart prognostic and moderating covariates through targeted feature shrinkage; (iii) incorporate prior knowledge on the relevant covariates and their relative impact on the outcome. Carefully designed regularization can lead to improved ITE estimates and inferences on prognostic and moderating factors, since including a large number of covariates in a fully-saturated model to adjust for confounding may lead to misspecification. In this work, we propose an extension of the Bayesian Causal Forest framework (Starling et al. 2019; Hahn, Murray, and Carvalho 2020), consisting in additional Dirichlet priors placed on the trees splitting probabilities (Linero 2018), that implement fully Bayesian feature shrinkage on the prognostic and moderating covariates, and allow the incorporation of prior knowledge on their relative importance. BART (and consequently BCF) was originally designed to adapt to smoothness but not to sparsity.[1] Our extended version of the model can be easily fitted with a slight modification of the existing MCMC algorithm and provably results in improved performance thanks to its better adaptability to sparse DGPs, for negligible extra computational cost.

The rest of the article is organized as follows. Section 2 introduces the problem of estimating treatment effects using the Neyman–Rubin causal model framework, and formulates the necessary assumptions to recover treatment effect estimates under confounded observational data. Section 3 offers an overview on Bayesian Additive Regression Trees and their popular causal version, Bayesian Causal Forest. Section 4 introduces our shrinkage-inducing extension, under the name of "Shrinkage Bayesian Causal Forest." Section 5 presents results from simulated studies carried out to compare Shrinkage Bayesian Causal Forest performance with other state-of-the-art models. Section 6 provides an example of analysis using data from the Infant Health and Development Program aimed at investigating the effects of early educational support on cognitive abilities in low birth weight infants. Section 7 concludes with a discussion.

2. Problem Framework

In this section we outline the problem of deriving an estimator for ITE using observational data, using the formalism of the Neyman–Rubin potential outcomes framework (Rubin 1978; Imbens and Rubin 2015).[2] We consider a setup where the outcome variable is continuous and the treatment assignment is binary (of the type exposure versus nonexposure), but most of the notions in this section can be generalized to noncontinuous responses and more than two treatment arms. For each individual $i \in \{1, \dots, N\}$, the two potential outcomes are defined as $Y_i^{(Z_i)}$, where $Z_i \in \{0, 1\}$ is the binary treatment assignment, with $Z_i = 1$ indicating exposure to the treatment, while $Z_i = 0$ nonexposure. We consider continuous type of outcomes such that $\left(Y_i^{(0)}, Y_i^{(1)}\right) \in \mathbb{R}^2$. Given the potential outcomes and the binary treatment assignment, ITE is defined, for each individual i, as the difference $Y_i^{(1)} - Y_i^{(0)}$. The fundamental problem of causal inference is that, for each i, we get to observe only one of the two potential outcomes $\left(Y_i^{(0)}, Y_i^{(1)}\right) \in \mathbb{R}^2$, specifically the one corresponding to the realization of Z_i, that is, $Y_i = Z_i Y_i^{(1)} + (1 - Z_i) Y_i^{(0)}$, so that ITE is never observable.

Given a dataset $\{\boldsymbol{X}_i, Z_i, Y_i\}$ of sample size N, where $\boldsymbol{X}_i \in \mathcal{X}$ are P pretreatment covariates, the ITE is the (unobserved) difference $Y_i^{(1)} - Y_i^{(0)}$. In practice, the goal is often to estimate the *Conditional Average Treatment Effects* (CATE), defined as

$$\tau(\boldsymbol{x}_i) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} | \boldsymbol{X}_i = \boldsymbol{x}_i\right]. \qquad (1)$$

CATE is the conditional mean of the ITE for the given value of the covariates, averaging across individual-level noise, so it is the best estimator for ITE in terms of mean squared error.

[1] Regularization in BART is introduced via shallow trees structures, to avoid overfitting (similarly to Gradient Boosting). Linero and Yang (2018) proposed a way to further enhance smoothness adaptation in BART through a probabilistic version of the trees, where inputs follow a probabilistic, rather than deterministic, path to the terminal nodes.

[2] Note that identification of causal effects can be achieved also with other causal frameworks, such as *do*-calculus in Structural Causal Models (Pearl 2009a,b, 2018), or decision-theoretic approach (Dawid 2000, 2015), and the contribution of this work, which concerns solely estimation, still apply.

In order to estimate $\tau(\boldsymbol{x}_i)$ through the observational quantities $\{\boldsymbol{X}_i, Z_i, Y_i\}$, we rely on a common set of assumptions to achieve identification. First of all, as already implied by the notation introduced above, we are assuming that *Stable Unit Treatment Value Assumption* (SUTVA) holds, ensuring that one unit's outcome is not affected by other units' assignment to treatment (*no interference*). The second assumption is unconfoundedness, which can be expressed through the conditional independence $(Y_i^{(0)}, Y_i^{(1)}) \perp\!\!\!\perp Z_i|\boldsymbol{X}_i$, and it rules out the presence of unobserved common causes of Z and Y (i.e., no unobserved confounders). The third and final assumption is common support, which means that all units have a probability of falling into either treatment groups which is strictly between 0 and 1. More formally, after defining the propensity score (Rosenbaum and Rubin 1983) as the probability of unit i being selected into treatment given $\boldsymbol{x}_i$,

$$\pi(\boldsymbol{x}_i) = \mathbb{P}(Z_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) \ , \tag{2}$$

common support implies $\pi(\boldsymbol{x}_i) \in (0,1) \forall i \in \{1, \dots, N\}$, so that there is no deterministic assignment to one of the groups given features $\boldsymbol{X}_i = \boldsymbol{x}_i$. Note that unconfoundedness and common support are automatically satisfied in the case of fully randomized experiments. While the degree of overlap between the two treatment groups can be typically examined in the data, SUTVA and unconfoundedness are untestable assumptions, and their plausibility must be justified based on domain knowledge.

In this work, we focus on nonparametric regression-based approaches to CATE estimation. Imbens (2004) offers a comprehensive overview on different methodologies (regression-based, matching-based, etc.) to derive different causal estimands of interest, such as sample and population ATE and CATE, Average Treatment effect on the Treated (ATT), Conditional Average Treatment effect on the Treated (CATT) etc. A nonparametric regression approach entails modeling the response surface as an unknown function of the covariates and treatment assignment indicator, and an error term. As typically done in the vast majority of the contributions on regression-based CATE estimation, we assume that the error term is additive and normally distributed with zero mean, such that Y_i is modeled as

$$Y_i = f(\boldsymbol{X}_i, Z_i) + \varepsilon_i \ , \qquad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

and where $f(\cdot)$ is of unknown form, and learnt from the data. A broad variety of methods to retrieve a CATE estimator from Equation (3) have been developed in the literature (see, Caron, Manolopoulou, and Baio 2020 for a review). Among these, we will follow in particular the one presented in Hahn, Murray, and Carvalho (2020), that we introduce and discuss in the next section.

3. BART for Causal Inference

Bayesian Additive Regression Trees (BART) are a nonparametric regression model that estimates the conditional expectation of a response variable Y_i via a "sum-of-trees." Considering the regression framework in (3), one can use BART to flexibly represent $f(\cdot)$ as

$$f(\boldsymbol{X}, Z) = \sum_{j=1}^{m} g_j\Big(\big[\boldsymbol{X} \;\; Z\big], \; \big(T_j, M_j\big)\Big) \ , \tag{4}$$

where m is the total number of trees in the model; the pair (T_j, M_j) defines the structure of the jth tree, namely T_j embeds the collection of binary split rules while $M_j = \{\psi_1, \dots, \psi_b\}$ the collection of b terminal nodes in that tree; $g_j(\cdot)$ is a tree-specific function mapping the predictors $[\boldsymbol{X} \;\; Z]$ to the set of terminal nodes M_j, following the set of binary split rules expressed by T_j. The conditional mean function $f(\boldsymbol{x}, z) = \mathbb{E}\big[Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i, Z_i = z_i\big]$ fit is computed by summing up all the terminal nodes ψ_{ij} assigned to the predictors $[\boldsymbol{X} \;\; Z]$ by the tree functions $g_j(\cdot)$, that is, $\sum_{j=1}^{m} g_j(\cdot)$. We refer the reader to Chipman, George, and McCulloch (1998, 2010) and Ročková and Saha (2019) for more details about BART priors and inference.

3.1. Bayesian Causal Forests

As briefly mentioned in Section 2, we will follow the representation proposed by Hahn, Murray, and Carvalho (2020), that avoids imposing direct regularization on $f(\cdot)$ in Equation (3). Hahn et al. (2018) and Hahn, Murray, and Carvalho (2020) in fact show that regularization on $f(\cdot)$ can generate unintended bias in the final estimation of CATE, and propose a simple reparameterization of (3) that uses a two-stage regression approach that dates back to the early contributions of Heckman (1979) and Robinson (1988). The two-stage representation reads:

$$Z_i \sim \text{Bernoulli}\big(\pi(\tilde{\boldsymbol{X}}_i)\big) \ , \ \pi(\tilde{\boldsymbol{x}}_i) = \mathbb{P}(Z_i = 1|\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_i) \ , \tag{5}$$

$$Y_i = \mu\Big(\big[\boldsymbol{X}_i \;\; \pi(\tilde{\boldsymbol{X}}_i)\big]\Big) + \tau(\boldsymbol{W}_i)Z_i + \varepsilon_i \ . \tag{6}$$

The first stage (5) deals with propensity score estimation, for which any probabilistic classifier is suitable (e.g., logistic regression, Probit BART, neural nets, etc.). In the simulated experiments shown in later sections, we will specifically employ either default probit BART or a one-hidden-layer neural network. In general, it is advisable not to rely on aggressive regularization in the estimation of $\pi(\cdot)$, as this could accidentally result into one or more main confounders being over-shrunk and/or left out of the model. The second stage (6) estimates the prognostic score $\mu(\cdot)$, defined as the effect of the covariates $\boldsymbol{X}_i \in \mathcal{X}$ on the outcome Y_i in the absence of treatment $\mu(\boldsymbol{x}_i) = \mathbb{E}\big[Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i, Z_i = 0\big]$, and CATE $\tau(\cdot)$. Note that we use slightly different notation for the covariates in $\mu(\cdot)$, $\tau(\cdot)$ and $\pi(\cdot)$. This is to highlight the fact that the set of available covariates $\boldsymbol{X}_i \in \mathcal{X}$ might consist of four different types (Herren and Hahn 2020): (i) *confounders*, that is, direct and indirect common causes of Z and Y; (ii) *prognostic covariates*, that is, predictors of $\mu(\cdot)$ only; (iii) *moderators*, that is, predictors of $\tau(\cdot)$ only; (iv) *propensity covariates*, entering only $\pi(\cdot)$ equation. Any covariate that does not fall into one of these categories is an irrelevant/nuisance predictor.

The two-stage procedure described above belongs to a class of models known as "modularized," as opposed to joint-models, that attempt to embed uncertainty around propensity scores in a single stage, which nonetheless can lead to poor estimates due to feedback issues in the approximation of the full posterior (Zigler et al. 2013; Zigler and Dominici 2014). See Jacob et al. (2017) for a thorough discussion on the issue of modularized versus joint models.

A Bayesian Causal Forest model (Hahn, Murray, and Carvalho 2020) is based on the reparameterization of the second stage regression (6). The advantage of this reparameterization from a Bayesian standpoint lies in the fact that separate priors, offering targeted regularization, can be placed on the prognostic score $\mu(\cdot)$ and on CATE $\tau(\cdot)$ directly. This approach mitigates unintended bias attributable to what the authors call *Regularization Induced Confounding* (RIC). The intuition behind RIC is that CATE posterior is strongly influenced by the regularization effects of the prior on $f(\cdot)$ in Equation (3), such that posterior estimates of CATE are bound to be biased, even more so in presence of strong confounding, such as when treatment selection is suspected to be "targeted," that is, when individuals are selected into treatment based on the prediction of an adverse potential outcome if left untreated. In order to alleviate confounding from targeted selection, the authors suggest to employ propensity score estimates obtained from the first stage $\hat{\pi}$ as an additional covariate in the estimation of $\mu(\cdot)$.

In practice, a BCF model assigns a default BART prior to $\mu(\cdot)$, while a prior with stronger regularization is chosen for $\tau(\cdot)$, as moderating patterns are believed to be simpler. The BART prior on $\tau(\cdot)$, compared to the default specification, consists in the use of a smaller number of trees in the ensemble (50 trees instead of 200), and a different combination of hyperparameters that govern the depth of each tree. In particular, in the context of BART priors, the probability that a node at depth $d \in \{0, 1, 2, \ldots\}$ in a tree is nonterminal is given by $\nu(1+\beta)^{-d}$, where (ν, β) are the hyperparameters to set (Chipman, George, and McCulloch 2010). The default specification $(\nu, \beta) = (0.95, 2)$ already has a shrinkage effect that accommodates small trees. The BCF prior on $\tau(\cdot)$ instead sets $(\nu, \beta) = (0.25, 3)$, with the purpose of assigning higher probability mass to even smaller trees. This combination of hyperparameters in the CATE prior allows to detect weak heterogeneous patterns, and provides robustness in case of homogeneous treatment effects.

For the reasons illustrated above, BCF tends to outperform BART and other tree-based methods for CATE estimation, such as Causal Forests (Wager and Athey 2018). As we will illustrate in the following sections, our work extends the BCF framework by introducing explicit shrinkage of irrelevant predictors, which results into higher computational efficiency, and accommodates different levels of smoothness across covariates, while, at the same time, returning interpretable measures of feature importance in the estimation of $\mu(\cdot)$ and $\tau(\cdot)$, separately.

4. Shrinkage Bayesian Causal Forests

BART, and consequently BCF, are known to handle sparsity quite well, thanks to the fact that splitting variables are chosen uniformly at random. However, they do not actively implement heterogeneous sparsity, nor feature shrinkage, which inevitably implies assigning equal level of heterogeneity to every covariate in the model. We will briefly illustrate in this section the concept of feature shrinkage in the context of tree ensemble models such as BART. Let us define first $\boldsymbol{s} = (s_1, \ldots, s_P)$ as the vector of splitting probabilities of each predictor $j \in \{1, \ldots, P\}$, where each s_j represents the probability for the jth predictor of being chosen as a splitting variable in one of the decision nodes of a tree. The default version of BART places a uniform distribution over the splitting variables, meaning that each predictor has equal chance of being picked as a splitting variable: $s_j = P^{-1} \quad \forall j \in \{1, \ldots, P\}$. As a consequence, predictors are virtually given equal prior importance in the model. A sparsity-inducing solution in this framework implies having a vector $\boldsymbol{s}$ of "stick-breaking" posterior splitting probabilities where ideally the entries corresponding to irrelevant predictors are near-zero, while the ones corresponding to relevant predictors are significantly higher than P^{-1}. Posterior splitting probabilities in this context can be intuitively viewed as a measure of variables importance (Breiman 2001). A complementary, decision-theoretic interpretation of sparsity-inducing solutions in this setup is given by the posterior probabilities that a predictor j appears in a decision node at least once in the ensemble. The two interpretations above (variables importance and probability of inclusion) are interchangeable and qualitatively lead to the same conclusions. In the next section we review how a simple extension of BART proposed by Linero (2018) can accommodate sparse solutions as described above, and how this modified version of BART can be put to use in the context of Bayesian Causal Forests.

4.1. Dirichlet Additive Regression Trees

Dirichlet Additive Regression Trees (Linero 2018), or DART, constitute an effective and practical way of inducing sparsity in BART. The proposed modification consists in placing an additional Dirichlet prior on the vector of splitting probabilities $\boldsymbol{s}$, which triggers a consequent posterior update in the backfitting MCMC algorithm. The Dirichlet prior on $\boldsymbol{s}$ reads

$$(s_1, \ldots, s_P) \sim \text{Dirichlet}\left(\frac{\alpha}{P}, \ldots, \frac{\alpha}{P}\right) , \tag{7}$$

where α is the hyperparameter governing the a priori preference for sparsity. Lower values of α correspond to sparser solutions, that is, fewer predictors included in the model. The hyperparameter α is in turn assigned a prior distribution, in order to deal with unknown degree of sparsity. This prior is chosen to be a Beta distribution, placed over a standardized version of the α parameter, of the following form

$$\frac{\alpha}{\alpha + \rho} \sim \text{Beta}(a, b) , \tag{8}$$

where the default parameter values are $(a, b, \rho) = (0.5, 1, P)$. The combination of values $a = 0.5$ and $b = 1$ assigns higher probability to low values of α, thus, giving preference to sparse solutions (the combination $(a, b) = (1, 1)$ would instead revert back to default BART splitting probabilities, that is, uniform distribution over the splitting variables). The prior is assigned to the standardized version of α in Equation (8) instead of α directly, as this allows to easily govern preference for sparsity through the parameter ρ. If one suspects that the level of sparsity is, although unknown, rather high, setting a smaller value of ρ facilitates even sparser solutions.

The modified version of DARTs MCMC implies an extra step to update $\boldsymbol{s}$, according to the conjugate posterior

$$s_1, \ldots, s_P | (u_1, \ldots, u_P) \sim \text{Dirichlet}\left(\frac{\alpha}{P} + u_1, \ldots, \frac{\alpha}{P} + u_P\right) , \tag{9}$$

where the update depends on u_j, defined as the number of attempted splits on the jth predictor in the current MCMC iteration. The phrase "attempted splits" refers to the fact that BART MCMC algorithm generates trees through a branching process undergoing a Metropolis–Hastings step, so that a proposed tree in the process might be rejected, but the chosen splitting variables are counted anyway in $\boldsymbol{u} = (u_1, \dots, u_P)$ (Chipman, George, and McCulloch 1998, 2010; Linero and Yang 2018).

The rationale behind the update in Equation (9) follows the natural Dirichlet-Multinomial conjugacy. The more frequently a variable is chosen for a splitting rule in the trees of the ensemble in a given MCMC iteration (or equivalently the higher is u_j), the higher the weight given to that variable by the updated $\boldsymbol{s}|(u_1, \dots, u_P)$ in the next MCMC iteration. Hence, the higher s_j, the higher the chance for the jth predictor of being drawn as splitting variable from the multinomial distribution described by $\text{Multinom}\big(1,\ \boldsymbol{s}|\boldsymbol{u}\big)$. This extra Gibbs step comes at negligible computational cost when compared to default BART typical running time.

4.2. Shrinkage BCF Priors

Similarly to Linero (2018), symmetric Dirichlet priors can be straightforwardly embedded in the Bayesian Causal Forest framework to induce sparsity in the estimation of prognostic and moderating effects. Bearing in mind that, as described in the previous section, BCF prior consists in two different sets of independent BART priors, respectively, placed on the prognostic score $\mu(\cdot)$ and CATE $\tau(\cdot)$, our proposed extension implies adding an additional Dirichlet prior over the splitting probabilities to these BART priors. Throughout the rest of the work we will consider the case where $\boldsymbol{W}_i = \boldsymbol{X}_i$, that is, where the same set of covariates is used for the estimation of $\mu(\cdot)$ and $\tau(\cdot)$ (see Equation (6) for reference), but the ideas easily extend to scenarios where a different set of covariates is designed, based on domain knowledge, to be used for $\mu(\cdot)$ and $\tau(\cdot)$[3]. The additional priors are, respectively,

$$\begin{aligned} \boldsymbol{s}_\mu &\sim \text{Dirichlet}\left(\tfrac{\alpha_\mu}{P+1}, \dots, \tfrac{\alpha_\mu}{P+1}\right), &\quad \tfrac{\alpha_\mu}{\alpha_\mu+\rho_\mu} &\sim \text{Beta}(a,b) \\ \boldsymbol{s}_\tau &\sim \text{Dirichlet}\left(\tfrac{\alpha_\tau}{P}, \dots, \tfrac{\alpha_\tau}{P}\right), &\quad \tfrac{\alpha_\tau}{\alpha_\tau+\rho_\tau} &\sim \text{Beta}(a,b), \end{aligned} \tag{10}$$

where the Beta's parameters are chosen to be $(a, b) = (0.5, 1)$ as default. The hyperparameter ρ is set equal to $(P + 1)$ in the case of the prognostic score ($\rho_\mu = P + 1$) since, when estimating $\mu(\boldsymbol{x}_i)$, we make use of P covariates plus an estimate of the propensity score $\widehat{\pi}(\boldsymbol{x}_i)$ as an additional covariate. In the case of $\tau(\boldsymbol{x}_i)$, we set it equal to $\rho_\tau = \frac{P}{2}$ to give preference to even more targeted shrinkage, as the CATE is typically believed to display simple heterogeneity patterns and a higher degree of sparsity compared to the prognostic score.

Algorithm 1: Bayesian Backfitting MCMC in Shrinkage BCF

```
Input: Data $(X, Z, Y)$
Output: MCMC samples of
        $\{\mu^{(b)}(\cdot), \tau^{(b)}(\cdot), (\boldsymbol{s}_\mu|\boldsymbol{u}_\mu)^{(b)}, (\boldsymbol{s}_\tau|\boldsymbol{u}_\tau)^{(b)}, \sigma^{(b)}\}_{b=1}^{B}$
for $b = 1, \dots, B$ do
    Result: Sample $\mu^{(b)}(\boldsymbol{x}), (\boldsymbol{s}_\mu|\boldsymbol{u}_\mu)^{(b)}$
    for $j = 1, \dots, m_\mu$ do
        Sample tree structure
        $T_j\mu \sim p(T_j|R_j, \sigma) \propto p(T_j)p(R_j|T_j, \sigma)$
        Sample terminal nodes $M_j\mu \sim p(M_j|T_j, R_j, \sigma)$
        (conjugate normal)
    end
    Sample $(\boldsymbol{s}_\mu|\boldsymbol{u}_\mu) \sim$
    $\mathcal{D}\big(\alpha_\mu/(P+1) + u_{1\mu}, \dots, \alpha_\mu/(P+1) + u_{(P+1)\mu}\big)$

    Result: Sample $\tau^{(b)}(\boldsymbol{x}), (\boldsymbol{s}_\tau|\boldsymbol{u}_\tau)^{(b)}$
    for $j = 1, \dots, m_\tau$ do
        Sample tree structure
        $T_j\tau \sim p(T_j|R_j, \sigma) \propto p(T_j)p(R_j|T_j, \sigma)$
        Sample terminal nodes $M_j\tau \sim p(M_j|T_j, R_j, \sigma)$
        (conjugate normal)
    end
    Sample $(\boldsymbol{s}_\tau|\boldsymbol{u}_\tau) \sim \mathcal{D}\big(\alpha_\tau/P + u_{1\tau}, \dots, \alpha_\tau/P + u_{P\tau}\big)$

    Result: Sample $\sigma^{(b)}$
    Sample $\sigma \sim p\big(\sigma|\widehat{\mu}(\boldsymbol{x}_i), \widehat{\tau}(\boldsymbol{x}_i), Y\big)$
end
```

We refer to this setup as Shrinkage Bayesian Causal Forest (Shrinkage BCF). Naturally, the two Dirichlet priors trigger two separate extra steps in the Gibbs sampler, implementing draws from the conjugate posteriors:

$$\begin{aligned} \boldsymbol{s}_\mu \mid \boldsymbol{u}_\mu &\sim \text{Dirichlet}\big(\alpha_\mu/(P+1) + u_{1\mu}, \dots, \alpha_\mu/(P+1) + u_{(P+1)\mu}\big) \\ \boldsymbol{s}_\tau \mid \boldsymbol{u}_\tau &\sim \text{Dirichlet}\big(\alpha_\tau/P + u_{1\tau}, \dots, \alpha_\tau/P + u_{P\tau}\big). \end{aligned} \tag{11}$$

Shrinkage BCFs setup allows first of all to adjust to different degrees of sparsity in $\mu(\cdot)$ and $\tau(\cdot)$, and thus to induce different levels of smoothness across the covariates. Second, it naturally outputs feature importance measures on both the prognostic score and CATE separately, given that separate draws of the posterior splitting probabilities are returned. The raw extra computational time, per MCMC iteration, is slightly greater, albeit negligible, compared to default BCF; however, Shrinkage BCF demonstrates higher computational efficiency thanks to the fact that it avoids splitting on irrelevant covariates. Thus, it necessitate far fewer MCMC iterations to converge, and improves performance under sparse DGPs. A sketch of pseudo-code illustrating the backfitting MCMC algorithm in Shrinkage BCF can be found in Algorithm 1.

The Dirichlet priors in Shrinkage BCF can be also adjusted to convey prior information about the relevant covariates and their relative impact on the outcome. This can be achieved by introducing a set of scalar prior weights $\boldsymbol{k} = \{k_1, \dots, k_P\} \in \mathbb{R}_+^P$,

[3] In certain cases, the set of pretreatment covariates might benefit from an initial screening by the researcher in the design of the study, and later undergo feature shrinkage in Shrinkage BCF, with the possibility of incorporating further a priori knowledge through the prior distributions, as described later in this section. As we will show in Section 4.3, in fact, Shrinkage BCF not only adjusts to sparse data generating processes (DGPs) per se, but allocates splitting probabilities in a more efficient way among the covariates, compared to uniformly at random splits, increasing computational efficiency.

such that

$$\begin{aligned} \boldsymbol{s}_\mu &\sim \text{Dirichlet}\left(k_{1\mu}\frac{\alpha_\mu}{P+1}, \ldots, k_{(P+1)\mu}\frac{\alpha_\mu}{P+1}\right) , \\ \boldsymbol{s}_\tau &\sim \text{Dirichlet}\left(k_{1\tau}\frac{\alpha_\tau}{P}, \ldots, k_{P\tau}\frac{\alpha_\mu}{P}\right) . \end{aligned} \tag{12}$$

The weights can take on different values for each covariate and can be set separately for prognostic score and CATE. If the jth covariate is believed to be significant in predicting $\mu(\cdot)$, then its corresponding prior weight $k_{j\mu}$ can be set higher than the others, in order to generate draws from a Dirichlet distribution that allocate higher splitting probability to that covariate. In the simulated experiment of Section 5.2 we will introduce a version of Shrinkage BCF with informative priors assigning higher a priori weight to the propensity score in $\mu\big(\boldsymbol{x}_i, \pi(\boldsymbol{x}_i)\big)$, to investigate whether this helps tackling strong confounding.

4.3. Targeted Sparsity and Covariate Heterogeneity

As a result of a fully Bayesian approach to feature shrinkage, Shrinkage BCF returns nonuniform posterior splitting probabilities that assign higher weight to more predictive covariates. This automatically translates into more splits along covariates with higher predictive power, compared to default BCF. To investigate whether this more strategic allocation of splitting probabilities in Shrinkage BCF leads to better performance, we test it against a default version of BCF including all the covariates and a version of BCF that already employs the subset of relevant covariates only. Think of the latter as a sort of “oracle” BCF that knows a priori the subset of relevant covariates, but may not assign different weights to them in terms of relative importance in the estimation of $\mu(\cdot)$ and $\tau(\cdot)$, respectively. To this end, we run a simple simulated example with $P = 10$ correlated covariates, of which only 5 are relevant, meaning that they exert some effect on the prognostic score or on CATE. We compare default BCF, “oracle” BCF using only the 5 relevant covariates and Shrinkage BCF using all the covariates (5 relevant and 5 nuisance). We generate the $P = 10$ covariates from a multivariate Gaussian $(X_1, \ldots, X_{10}) \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, where the entries of the covariance matrix are such that $\Sigma_{jk} = 0.6^{|j-k|} + 0.1\mathbb{I}(j \neq k)$, indicating positive correlation between predictors. Sample size is set equal to $N = 1000$. We then generate treatment assignment as $Z_i \sim \text{Bern}\big(\pi(\boldsymbol{x}_i)\big)$, where the propensity score is

$$\pi(\boldsymbol{x}_i) = \mathbb{P}(Z_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \Phi\big(-0.4 + 0.3X_{i,1} + 0.2X_{i,2}\big) , \tag{13}$$

and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The prognostic score, CATE and response Y_i are, respectively, generated as

$$\begin{aligned} \mu(\boldsymbol{X}_i) &= 3 + X_{i,1} + 0.8\sin(X_{i,2}) + 0.7X_{i,3}X_{i,4} - X_{i,5} , \\ \tau(\boldsymbol{X}_i) &= 2 + 0.8X_{i,1} - 0.3X_{i,12}^2 , \\ Y_i &= \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)Z_i + \varepsilon_i , \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, 1) . \end{aligned} \tag{14}$$

Table 1. Sample average bias, $\sqrt{\text{PEHE}}$ and 95% coverage for default BCF, ”oracle” BCF which uses only the five relevant predictors (Oracle BCF-5) and shrinkage BCF (SH-BCF).

Model	Bias	$\sqrt{\text{PEHE}}$	95% Coverage
BCF	0.037 ± 0.008	0.447 ± 0.006	**0.92 ± 0.01**
Oracle BCF-5	0.034 ± 0.008	0.440 ± 0.006	0.91 ± 0.01
SH-BCF	**0.031 ± 0.007**	**0.380 ± 0.006**	0.88 ± 0.01

NOTE: Bold text represents better performance.

In this experiment only the first five predictors are relevant. Table 1 shows performances of the default BCF, “oracle” BCF run on just the 5 relevant predictors (oracle BCF-5) and Shrinkage BCF (SH-BCF), averaged over $H = 500$ Monte Carlo simulations. Performance of the methods is measured through: bias, defined as $\mathbb{E}\big[(\hat{\tau}_i - \tau_i)|\boldsymbol{X}_i = \boldsymbol{x}_i\big]$; the quadratic loss function

$$\mathbb{E}\big[(\hat{\tau}_i - \tau_i)^2|\boldsymbol{X}_i = \boldsymbol{x}_i\big] , \tag{15}$$

where $\hat{\tau}_i$ is the model-specific CATE estimate, while τ_i is the ground-truth CATE; and finally 95% frequentist coverage, defined as $\mathbb{P}\big(\hat{\tau}(\boldsymbol{x_i})_{\text{low}} \leq \tau(\boldsymbol{x_i}) \leq \hat{\tau}(\boldsymbol{x_i})_{\text{upp}}\big)$, where $\hat{\tau}(\boldsymbol{x_i})_{\{\text{low,high}\}}$ are the upper and lower bounds of 95% credible interval around $\hat{\tau}(\boldsymbol{x_i})$, returned by the MCMC. The loss function in (15) is also known as the *Precision in Estimating Heterogeneous Treatment Effects* (PEHE) from Hill (2011). Bias, PEHE and coverage estimates are estimated by computing, for each of the $H = 500$ Monte Carlo simulations, their sample equivalents

$$\begin{aligned} \widehat{\text{Bias}}_\tau &= \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\tau}(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i)\Big) \\ \widehat{\text{PEHE}}_\tau &= \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\tau}(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i)\Big)^2 \\ \widehat{\text{Coverage}}_\tau &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big(\hat{\tau}(\boldsymbol{x_i})_{\text{low}} \leq \tau(\boldsymbol{x_i}) \leq \hat{\tau}(\boldsymbol{x_i})_{\text{upp}}\Big) , \end{aligned}$$

and then averaging these over all the simulations. More precisely, Table 1 reports bias, $\sqrt{\text{PEHE}}$ and coverage estimates together with 95% Monte Carlo confidence intervals.

Shrinkage BCF shows better performance than default BCF as well as the “oracle” BCF version in terms of bias and $\sqrt{\text{PEHE}}$, while reports just marginally lower coverage, indicating that the method allocates “stick-breaking” splitting probabilities in an efficient way and necessitates fewer MCMC iterations for convergence. The intuition as to why Shrinkage BCF performs better than “oracle” BCF, is that its priors allow not only to split more along relevant covariates instead of irrelevant ones (which explains the advantage over BCF), but also to split more frequently along covariates that are more predictive of the outcome, resulting in higher computational efficiency. To illustrate this concept, suppose we have the following trivial linear DGP with two covariates on the same scale, $Y = 2X_1 + X_2$. Both covariates are relevant for predicting Y, but X_1 has a relatively higher impact in magnitude. DART, and thus Shrinkage BCF, allocate more splits along the more predictive dimension X_1, while BART produces a similar level of splits along both X_1 and X_2 and hence requires a larger number of MCMC iterations and provides noisier estimates.

4.4. Targeted Regularization in Confounded Studies

The parameterization in BCF, and thus in Shrinkage BCF as well, is designed to effectively disentangle prognostic and moderating effects of the covariates and to induce different levels of sparsity when estimating these effects, in contrast to other methods for CATE estimation. The purpose of this section is to

Table 2. Posterior splitting probabilities from S-learner DART, T-learner DART and Shrinkage BCF over the five available covariates.

Method		Variable					
		X_1	X_2	X_3	X_4	X_5	Z
S-DART	$f(\cdot)$	**0.12**	**0.43**	0.00	0.00	0.00	**0.45**
T-DART	$f_0(\cdot)$	**0.29**	**0.70**	0.01	0.00	0.00	–
	$f_1(\cdot)$	**0.09**	**0.90**	0.00	0.01	0.00	–
SH-BCF	$\mu(\cdot)$	**0.98**	0.01	0.00	0.00	0.01	–
	$\tau(\cdot)$	0.00	**0.96**	0.00	0.03	0.01	–

NOTE: Values in bold denote which covariates receive significant chunks of splitting probability in fitting the corresponding functions, that characterize each model.

briefly illustrate with a simple example how naively introducing sparsity through a model that does not explicitly guard against RIC can have a detrimental effect on CATE estimates. To this end, we simulate, for $N = 1000$ observations, $P = 5$ correlated covariates as $(X_1, \dots, X_5) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where the entries of the covariance matrix are $\Sigma_{jk} = 0.6^{|j-k|} + 0.1\mathbb{I}(j \neq k)$. The treatment allocation, prognostic score, CATE and response Y_i are then, respectively, generated as follows:

$$\begin{aligned}
Z_i &\sim \text{Bernoulli}\big(\pi(\boldsymbol{x}_i)\big) ,\\
\pi(\boldsymbol{x}_i) &= \Phi\big(-0.5 + 0.4X_{i,1}\big) ,\\
\mu(\boldsymbol{X}_i) &= 3 + X_{i,1} ,\\
\tau(\boldsymbol{X}_i) &= 0.5 + 0.5X_{i,2}^2 ,\\
Y_i &= \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)Z_i + \varepsilon_i , \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0,1) .
\end{aligned}$$

Notice that in this simple setup the prognostic effect is determined by the first covariate $X_{i,1}$, while the moderating effect by the second covariate $X_{i,2}$. We run CATE estimation via three different methods that make use of DART priors. The first is a "Single-Learner" (S-learner) that employs DART (S-DART) to fit a single surface $f(\cdot)$ and computes CATE estimates as $\hat{\tau}(\boldsymbol{x}_i) = \hat{f}(\boldsymbol{x}_i, Z_i = 1) - \hat{f}(\boldsymbol{x}_i, Z_i = 0)$. The second is a "Two-Learner" (T-learner) that employs DART (T-DART) to fit two separate surfaces, $f_1(\cdot)$ and $f_0(\cdot)$, for the two treatment groups and derives CATE estimates as $\hat{\tau}(\boldsymbol{x}_i) = \hat{f}_1(\boldsymbol{x}_i) - \hat{f}_0(\boldsymbol{x}_i)$. The last method is our Shrinkage BCF (SH-BCF). Each of these methods is able to account for sparsity when estimating CATE. However, the interpretation of covariate importance is very different across them, due to the way the CATE estimator is derived. In particular, as indicated by the posterior splitting probabilities of each method in Table 2, S-DART fits a single surface $f(\cdot)$, where Z is treated as an extra covariate, so it ends up assigning most of the splitting probability to Z and then in turn to other relevant covariates. T-DART performs "group-specific" feature shrinkage, in that it fits separate surfaces for each of the treatment groups. Although both S-DART and T-DART turn out to select the relevant covariates for the final estimation of CATE, they are unable, by construction, to distinguish between prognostic and moderating ones. Shrinkage BCF instead, thanks to its parameterization, is capable of doing so, disentangling the two effects.

In Section 5, we will show that Shrinkage BCF outperforms default BCF and other state-of-the-art methods in estimating CATE under two more challenging simulated exercises. Furthermore, in the supplementary materials we present results from few additional simulated experiments.

5. Simulated Experiments

In this section, we report results from two simulated studies carried out to demonstrate the performance of Shrinkage BCF and its informative prior version under sparse DGPs. The first simulated study is intended to evaluate Shrinkage BCF performance compared to other popular state-of-the-art methods for CATE estimation, and to show how it scales up with an increasing number of nuisance covariates. In addition, we will also illustrate how the method returns interpretable feature importance measures, as posterior splitting probabilities on $\mu(\cdot)$ and $\tau(\cdot)$. The second simulated setup instead mimics a strongly confounded study, and is designed to show how versions of Shrinkage BCF deal with targeted selection scenarios. In the supplementary materials, we present further results from four additional simulated exercises, designed to: (i) study what happens with perfectly known propensity scores in confounded settings; (ii) investigate computational advantage of DART priors; (iii) test Shrinkage BCFs reliability under increasingly larger P; (iv) consider different types of sparse DGPs. The R code implementing Shrinkage BCF is available at: *https://github.com/albicaron/SparseBCF*.

5.1. Comparison to Other Methods

The first setup consists of two parallel simulated studies, where only the total number of predictors ($P = 25$ and $P = 50$) is changed. The purpose underlying this setup is to illustrate how Shrinkage BCF relative performance scales up when nuisance predictors are added and the level of sparsity increases.

For both simulated exercises, sample size is set equal to $N = 1000$. In order to introduce correlation between the covariates, they are generated as correlated uniforms from a Gaussian Copula $C_{\Theta}^{\text{Gauss}}(u) = \Phi_{\Theta}\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_P)\big)$, where Θ is a covariance matrix such that $\Theta_{jk} = 0.3^{|j-k|} + 0.1\mathbb{I}(j \neq k)$. A 40% fraction of the covariates is generated as continuous, drawn from a standard normal distribution $\mathcal{N}(0,1)$, while the remaining 60% as binary, drawn from a binomial $\text{Bin}(N, 0.3)$. Propensity score is generated as

$$\begin{aligned}
\pi(\boldsymbol{x}_i) &= \mathbb{P}(Z_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)\\
&= \Phi\left(-0.5 + 0.2X_{i,1} + 0.1X_{i,2} + 0.4X_{i,21} + \frac{\eta_i}{10}\right), \qquad (16)
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal, and η_i is a noise component drawn from a uniform $\mathcal{U}(0,1)$. The binary treatment indicator is drawn as $Z_i \sim \text{Bernoulli}\big(\pi(\boldsymbol{x}_i)\big)$. Prognostic score and CATE functions are simulated as follows:

$$\begin{aligned}
\mu(\boldsymbol{x}_i) &= 3 + 1.5\sin(\pi X_{i,1}) + 0.5(X_{i,2} - 0.5)^2 + 1.5(2 - |X_{i,3}|)\\
&\quad + 1.5X_{i,4}(X_{i,21} + 1)\\
\tau(\boldsymbol{x}_i) &= 0.1 + |X_{i,1} - 1|(X_{i,21} + 2) .
\end{aligned} \qquad (17)$$

Notice that only five predictors among $P \in \{25, 50\}$, namely $\{X_1, X_2, X_3, X_4, X_{21}\}$, are relevant to the estimation of the prognostic score and CATE. Eventually, the response variable Y_i is generated as usual:

$$Y_i = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)Z_i + \varepsilon_i , \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) . \qquad (18)$$

Table 3. List of models tested on the simulated experiment in Section 5.1.

Family	Label	Description
Linear models	S-OLS	Linear regression as S-learner
	T-OLS	Linear regression as T-learner
	R-LASSO	LASSO regression as R-learner
Naive nonparametrics	*k*NN	*k*-Nearest neighbors as T-learner
Tree-based methods	S-BART	BART as S-learner
	T-BART	BART as T-learner
	CF	Causal forest
	S-DART	DART as S-learner
	T-DART	DART as T-learner
	BCF	Bayesian causal forest
	SH-BCF	Shrinkage Bayesian causal forest
Gaussian processes	CMGP	Causal multi-task Gaussian process
	NSGP	Nonstationary Gaussian process

The error term standard deviation is set equal to $\sigma = \frac{\hat{\sigma}_\mu}{2}$, where $\hat{\sigma}_\mu$ is the sample standard deviation of the simulated prognostic score $\mu(\boldsymbol{x}_i)$ in (17).

Performance of each method is evaluated through $\sqrt{\text{PEHE}}$ estimates, averaged over $H = 1000$ replications, reported together with 95% Monte Carlo confidence intervals. Data are randomly split in 70% train set, used to train the models, and 30% test set to evaluate the model on unseen data; $\sqrt{\text{PEHE}}$ estimates are reported both for train and test data.

The models evaluated on the simulated data are summarized in Table 3. We make use of the Meta-Learners terminology described in Künzel et al. (2017) and Caron, Manolopoulou, and Baio (2020). The first set of models includes a S-learner and a T-learner least squares regressions (S-OLS and T-OLS), and a R-learner (Nie and Wager 2020) LASSO regression (R-LASSO). The second set consists just in a naive *k*-nearest neighbors (*k*NN) as a T-learner. The third set includes the following popular tree ensembles methods: Causal Forest (CF) (Wager and Athey 2018); a S-learner and a T-learner versions of BART (S-BART and T-BART) (Hill 2011; Green and Kern 2012; Sivaganesan, Müller, and Huang 2017) and DART (S-DART and T-DART); Bayesian Causal Forest (BCF) (Hahn, Murray, and Carvalho 2020); and finally our method, Shrinkage Bayesian Causal Forest (SH-BCF). The last set includes two causal multitask versions of Gaussian Processes, with stationary (CMGP) and nonstationary (NSGP) kernels, respectively, both implementing sparsity-inducing Automatic Relevance Determination over the covariates (Alaa and van der Schaar 2017, 2018).

Performance of each method, for the two simulated scenarios with $P = 25$ and $P = 50$ covariates, respectively, is shown in Table 4. Results demonstrate the high adaptability and scalability of Shrinkage BCF, as the method displays the lowest estimated error in both simulated scenarios, and its performance is not undermined when extra nuisance covariates are added, while the other methods generally deteriorate.

Figure 1 shows how Shrinkage BCF correctly picks the relevant covariates behind both prognostic and moderating effects, in contrast to default BCF which assigns equal probability of being chosen as a splitting variable to each predictor. Notice also that results do not essentially vary between the $P = 25$ and the $P = 50$ scenarios (respectively, first and second row graphs in Figure 1), as Shrinkage BCF virtually selects the same relevant predictors.

Table 4. Train and test set $\sqrt{\text{PEHE}}$ estimates, together with 95% confidence interval, in the case of $P = 25$ covariates and $P = 50$ covariates scenarios.

	$P = 25$		$P = 50$	
	Train	Test	Train	Test
S-OLS	1.91 ± 0.00	1.91 ± 0.01	1.91 ± 0.00	1.91 ± 0.01
T-OLS	1.41 ± 0.01	1.47 ± 0.01	1.68 ± 0.01	1.78 ± 0.01
R-LASSO	1.17 ± 0.01	1.19 ± 0.01	1.20 ± 0.01	1.22 ± 0.01
*k*NN	1.62 ± 0.01	1.66 ± 0.01	1.72 ± 0.01	1.76 ± 0.01
S-BART	0.77 ± 0.01	0.79 ± 0.01	0.85 ± 0.01	0.86 ± 0.01
T-BART	1.11 ± 0.01	1.11 ± 0.01	1.28 ± 0.01	1.29 ± 0.01
CF	1.05 ± 0.01	1.05 ± 0.01	1.23 ± 0.01	1.23 ± 0.01
S-DART	0.59 ± 0.01	0.60 ± 0.01	0.59 ± 0.01	0.60 ± 0.01
T-DART	0.88 ± 0.01	0.89 ± 0.01	0.90 ± 0.01	0.90 ± 0.01
BCF	0.79 ± 0.01	0.82 ± 0.01	0.86 ± 0.01	0.88 ± 0.01
SH-BCF	**0.54 ± 0.01**	**0.56 ± 0.01**	**0.55 ± 0.01**	**0.55 ± 0.01**
CMGP	0.59 ± 0.01	0.61 ± 0.01	0.85 ± 0.03	0.77 ± 0.02
NSGP	0.60 ± 0.01	0.62 ± 0.01	0.74 ± 0.03	0.75 ± 0.03

Bold indicates best performance.

5.2. Strongly Confounded Simulated Study

This section presents results from a second simulated study, aimed at showing how Shrinkage BCF addresses scenarios characterized by strong confounding. In particular, the setup is designed around the concept of targeted selection, a common type of selection bias in observational studies, expressively tackled by the BCF framework, that implies a direct relationship between $\mu(\cdot)$ and $\pi(\cdot)$. We run the simulated experiment in the usual way, by first estimating the unknown propensity score; then we also rerun the same experiment assuming that propensity score is known (results in the supplementary materials), to gain insights by netting out effects due to propensity model misspecification.

We simulate $N = 500$ observations from $P = 15$ correlated covariates (the first five continuous and the remaining 10 binary), generated as correlated uniforms from the Gaussian Copula $C_\Theta^{\text{Gauss}}(u) = \Phi_\Theta\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_P)\big)$, where the covariance matrix is such that $\Theta_{jk} = 0.6^{|j-k|} + 0.1\mathbb{I}(j \neq k)$. The relevant quantities are simulated as follows:

$$
\begin{aligned}
\mu(\boldsymbol{x}_i) &= 5\Big(2 + 0.5\sin(\pi X_{i,1}) - 0.25X_{i,2}^2 + 0.75X_{i,3}X_{i,9}\Big)\,,\\
\tau(\boldsymbol{x}_i) &= 1 + 2|X_{i,4}| + 1X_{i,10}\,,\\
\pi(\boldsymbol{x}_i) &= 0.9\,\Lambda\,(1.2 + 0.2\mu(\boldsymbol{x}_i))\,,\\
Z_i &\sim \text{Bernoulli}\big(\pi(\boldsymbol{x}_i)\big)\,,\\
Y_i &= \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)Z_i + \varepsilon_i\,, \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0,\sigma^2)\,,
\end{aligned}
\tag{19}
$$

where $\Lambda(\cdot)$ is the logistic cumulative distribution function. The error's standard deviation is set equal to half the sample standard deviation of the generated $\tau(\cdot)$, $\sigma^2 = \frac{\hat{\sigma}_\tau}{2}$. Targeted selection is introduced by generating the propensity score $\pi(\boldsymbol{x}_i)$ as a function of the prognostic score $\mu(\boldsymbol{x}_i)$ (Hahn, Murray, and Carvalho 2020). The BCF models tested on this simulated setup are: (i) Default BCF; (ii) agnostic prior Shrinkage BCF; (iii) agnostic prior Shrinkage BCF, without propensity score estimate as an additional covariate; (iv) Shrinkage BCF with informative prior on $\mu(\cdot)$ only, where prior weight given to propensity score is $k_{PS} = 50$; (v) Shrinkage BCF with the same prior as (iv), but $k_{PS} = 100$. We test a variety of BCF versions to examine how they tackle confounding deriving from targeted selection. In particular, with (iv) and (v), we investigate whether nudging

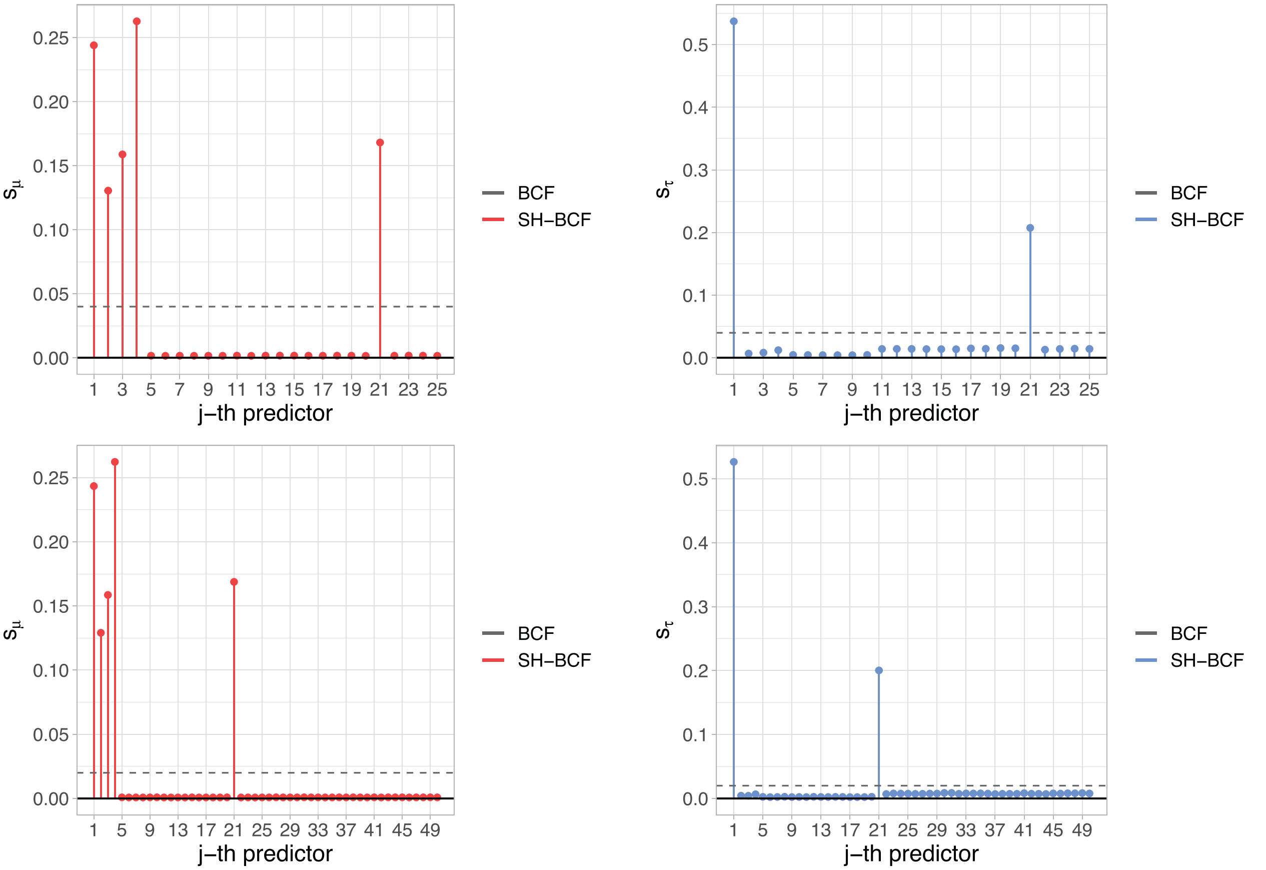

Figure 1. Shrinkage BCF posterior splitting probabilities for each single covariates, indexed on the x-axis, for $\mu(\cdot)$ (on the left) and $\tau(\cdot)$ (on the right), in the scenarios with $P = 25$ predictors (first row) and $P = 50$ predictors (second row). Spikes indicate higher probability assigned by Shrinkage BCF to the relevant predictors. The horizontal dashed lines denote default BCF uniform splitting probabilities.

more splits on the propensity score covariate induces better handling of confounding and better CATE estimates. With (ii) and (iii) we study whether it is sensible to have propensity score as an extra covariate, once we have accounted for sparsity, in settings such as the one described in (19), where propensity $\pi(\cdot)$ and prognostic score $\mu(\cdot)$ are functions of the same set of covariates—more specifically $\pi(\cdot)$ is a function of $\mu(\cdot)$.

We first compare the usual performance metrics (bias, $\sqrt{\text{PEHE}}$, 95% coverage), averaged over $H = 500$ replications, which are gathered in Table 5, together with the average posterior splitting probability assigned to propensity score $(s_\pi|u_\pi)$ by each model, where applicable. As for the posterior splitting probability $(s_\pi|u_\pi)$, we notice that in (ii) this is nearly zero, thus, not really different than not having $\pi(\cdot)$ at all, as in (iii). This means that estimates of $\pi(\cdot)$ do not virtually contribute a lot to the fit. Also, in (i) and (iv), the probability is more or less the same, meaning that, in this example, setting $k_{PS} = 50$ implies assigning similar $(s_\pi|u_\pi)$ as default BCF, but allowing sparsity across the other covariates. In addition to the information in Table 5, for a better visual inspection, we plot the posterior fit of the $\pi(\cdot)$ and $\mu(\cdot)$ relationship for each specification of BCF[4].

Table 5. Bias, $\sqrt{\text{PEHE}}$, 95% coverage and posterior splitting probability on $\hat{\pi}(x_i)$ — $(s_\pi|u_\pi)$ — for: (i) default BCF; (ii) Shrinkage BCF; (iii) Shrinkage BCF without $\hat{\pi}(x_i)$; (iv) informative prior BCF with $k_{PS} = 50$; (v) informative prior BCF with $k_{PS} = 100$.

Model	Bias	$\sqrt{\text{PEHE}}$	95% coverage	$(s_\pi\|u_\pi)$
(i) BCF	-0.06 ± 0.01	0.49 ± 0.01	0.94 ± 0.00	9.09%
(ii) SH-BCF	-0.05 ± 0.01	0.38 ± 0.01	0.96 ± 0.00	0.29%
(iii) SH-BCF (no PS)	-0.05 ± 0.01	0.38 ± 0.01	0.96 ± 0.00	–
(iv) I-BCF ($k_{PS} = 50$)	-0.05 ± 0.01	0.39 ± 0.01	0.96 ± 0.00	9.76%
9v) I-BCF ($k_{PS} = 100$)	-0.05 ± 0.01	0.40 ± 0.01	0.96 ± 0.01	17.48%

The results corroborate those of the previous sections, as all the Shrinkage BCF versions (ii)–(v) outperform default BCF (i), thanks to their ability to adapt to sparsity (Table 5). In order to net out effects that are due to propensity model misspecification, we rerun the same example in (19) for $H = 250$, this time assuming that PS is known, thus, plugging in the true values in $\mu(\boldsymbol{x}_i, \pi)$. Results can be found in the supplementary materials.

The picture emerging from this exercise is the following. Methods (ii)–(v) all have comparable performances in the realistic scenario where PS is to be estimated (see Table 5); moreover, Figure 2 show that, in this case, they all effectively capture the relationship between $\pi(\cdot)$ and $\mu(\cdot)$. Hence, adjusting prior weights to nudge more splits on the estimated PS—methods (iv) and (v)—does not seem to improve performance. In the more abstract scenario where PS is assumed to be known (whose results are gathered in supplementary materials), and thus the

[4] We avoid plotting the fit for (iii) Shrinkage BCF without $\pi(\cdot)$, since it yields very similar results to (ii) Shrinkage BCF with $\pi(\cdot)$—In Table 5, (ii) allocates nearly 0% splits to $\pi(\cdot)$, as in (iii).

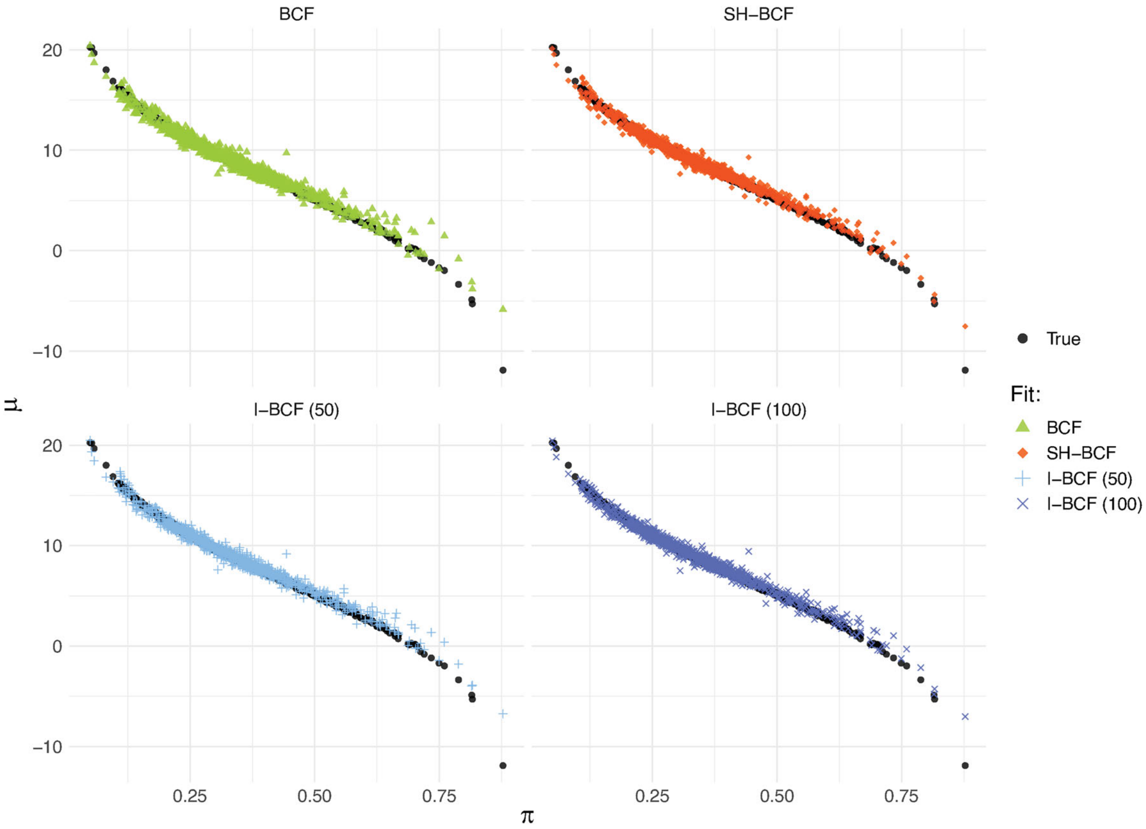

Figure 2. Posterior fit of $\pi(\cdot)$ and $\mu(\cdot)$ relationship, for default BCF, Shrinkage BCF (with $\pi(\cdot)$) and the two versions of informative prior BCF ($k_{PS} = 50$ and $k_{PS} = 100$). All the specifications effectively capture the underlying relationship.

relationship between $\pi(\cdot)$ and $\mu(\cdot)$ can be directly estimated, versions (i) and iii) perform poorly. The first because it does not induce sparsity, while iii) does not include $\pi(\cdot)$ as extra covariate. Versions (ii), (iv), and (v) instead perform comparatively better as they virtually assign all the splitting probability to $\pi(\cdot)$, leaving the other covariates out of the model. This is unsurprising in a setup where $\pi(\cdot)$ is known, as its relationship with $\mu(\cdot)$ is straightforwardly captured. Even under this abstract scenario, specifications (iv) and (v), which assign higher weight to $\pi(\cdot)$, do not show improvements on performance, as also the agnostic prior version (ii) effectively allocates the entire splitting probability to the $\pi(\cdot)$ covariate.

Results from the example where PS is perfectly known are in line with the findings of Hahn, Murray, and Carvalho (2020) and shed light on why adding $\pi(\cdot)$ as an extra covariate is always helpful in tackling targeted selection. Naturally, the success of this practice in addressing strong confounding heavily depends on the quality of the approximation of $\pi(\cdot)$, that is, the quality of the propensity model that estimates $\hat{\pi}(\cdot)$.

6. Case Study: The Effects of Early Intervention on Cognitive Abilities in Low Birth Weight Infants

In this section, we illustrate the use of Shrinkage BCF by revisiting the study in Brooks-Gunn et al. (1992), which analyzes data from the Infant Health and Development Program (IHDP), found also in the more recent contribution of Hill (2011). The IHDP was a randomized controlled trial aimed at investigating the efficacy of educational and family support services, with pediatric follow-ups, in improving cognitive skills of low birth weight preterm infants, who are known to have developmental problems regarding visual-motor and receptive language skills (McCormick 1985; McCormick, Gortmaker, and Sobol 1990). The study includes observations on 985 infants whose weight at birth was less than 2500 grams, across eight different sites. About one third of the infants were randomly assigned to treatment ($Z_i = 1$), which consisted in routine pediatric follow-up (medical and developmental), in addition to frequent home visits to inform parents about child's progress and communicate instructions about recommended activities for the child. Following Hill (2011), the outcome variable (Y_i) we use is the score in a Stanford Binet IQ test, whose values can range from a minimum of 40 to a maximum of 160, taken at the end of the intervention period (child's age equal 3). The available final sample, obtained after removing 77 observations with missing IQ test score, consists of $N = 908$ data points, while the number of pretreatment covariates amounts to $P = 31$. A full list of the variables included in the analysis, together with a short description, can be found in the supplementary materials.

First, we estimate propensity score using a 1-hidden layer neural network classifier. Then we run Shrinkage BCF with default agnostic prior for 15,000 MCMC iterations in total, but we discard the first 10,000 as burn-in. As output, we obtain the full posterior distribution on CATE estimates and splitting probabilities relative to each covariate. The left-hand pane graph of Figure 3 shows the estimated CATE posterior distribution for the individuals in the sample whose estimated propensity

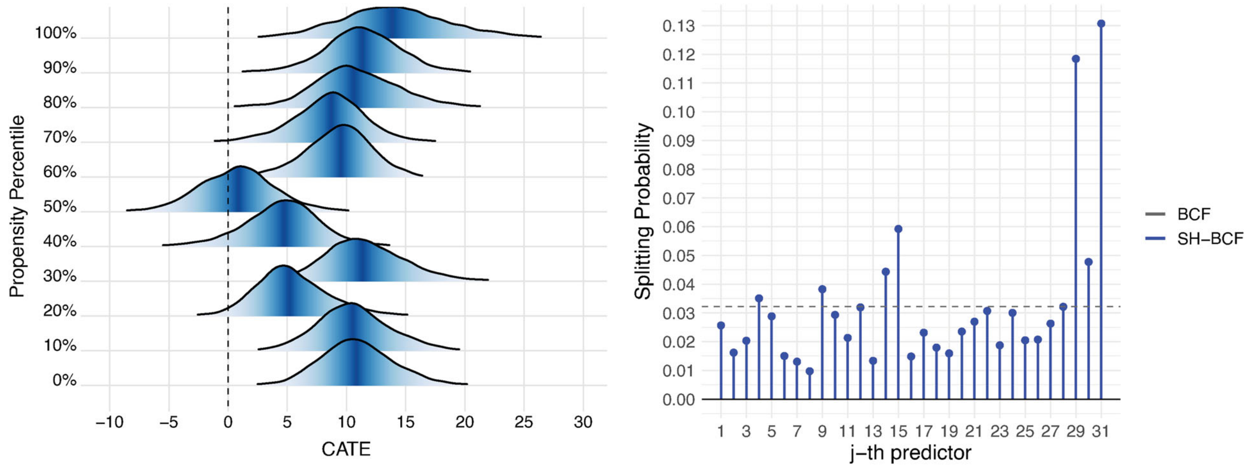

Figure 3. Left panel: Posterior distributions for the CATE estimates, obtained using Shrinkage BCF, corresponding to the approximated propensity percentiles (i.e., for individuals in the sample whose estimated propensity corresponds or is closest to the percentiles). Fill color is darker around the median. Right panel: Shrinkage BCFs posterior splitting probabilities on $\tau(\cdot)$, averaged over the post burn-in MCMC draws.

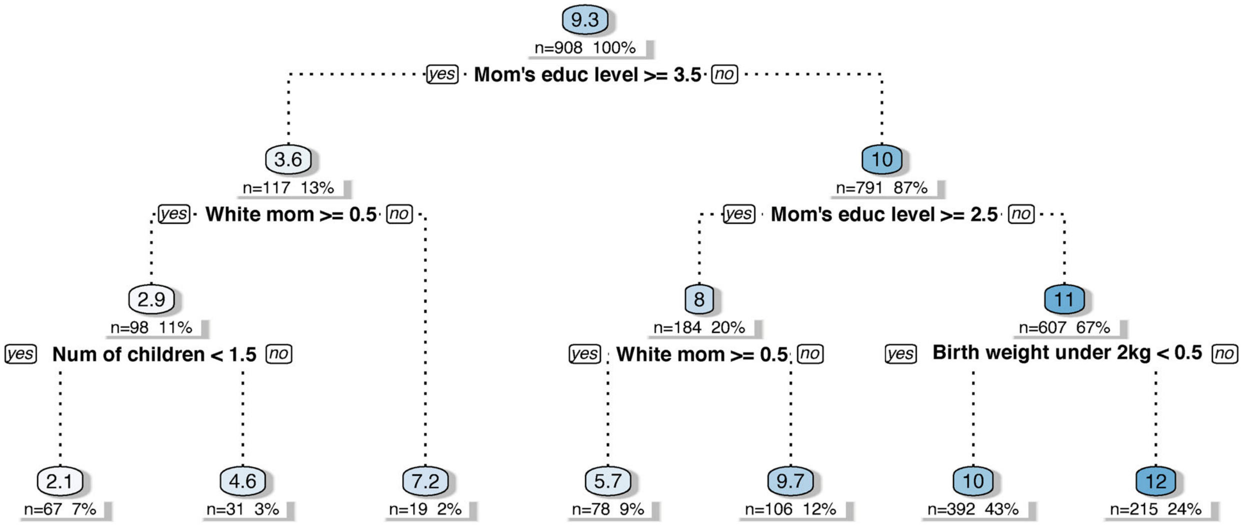

Figure 4. Decision tree identifying the most homogeneous subgroups in terms of treatment response, based on splitting rules involving the available covariates. The nodes report CATE estimates averaged within the corresponding subgroup.

corresponds, or is closest, to the ith percentile of the estimated propensity distribution, where i is 0, 10, 20,…, 100. The represented stratified CATE posterior distribution relative to these propensity values conveys information about the uncertainty around the estimates and depicts an overall positive and rather heterogeneous treatment effects. The estimated average treatment effect is equal to ATE $=$ 9.33 and standard deviation of CATE estimates, averaged over the post burn-in draws, is equal to 3.25, which is another sign of underlying heterogeneity patterns in the treatment response. The analysis would thus, benefit from further investigation about the heterogeneity of treatment effects, with the aim of distinguishing the impact within subgroups of individuals characterized by similar features (i.e., covariates values). Evidence on what the relevant drivers of heterogeneity behind treatment effect are is given by the posterior splitting probabilities on $\tau(\cdot)$ (again averaged over the post burn-in draws), reported in the right-hand pane graph of Figure 3, where few covariates end up being assigned relatively higher weights compared to the others. The two covariates that primarily stand out are the binary indicator on whether the mother's ethnicity is white (29th predictor) and the ordinal variable indicating mother's level of education (31st predictor).

We proceed with a sensitivity analysis of treatment effect subgroups by following the suggestion of Hahn, Murray, and Carvalho (2020); that is, we fit a decision tree partition algorithm using the R package `rpart`, by regressing mean CATE estimates obtained from Shrinkage BCF $\hat{\tau}(\boldsymbol{x}_i)$ (averaged over the MCMC post burn-in draws) on the available covariates $X_i \in \mathcal{X}$. The purpose of this sensitivity analysis exercise is to identify the most homogeneous subgroups, namely the subgroups leading to an optimal partition, in terms of their estimated mean CATE, as a function of the covariates, and to examine how much the emerging partition agrees with the results on posterior splitting probabilities in Figure 3.

Results are depicted in Figure 4 in the form of a decision tree, pruned at four levels. Zero splits trivially return ATE estimate (first node in Figure 4), while shallower nodes show CATE estimates averaged within the subgroup defined by the corresponding split rule. The first split is on the variable "Mother's level of education," specifically on whether the mother has attended college or not. The second level features a split on whether mother's ethnicity is white in one branch, and a split on whether mother has finished high school in the other. These are exactly the same covariates selected by the posterior splitting probabilities. The last set of splits is again on mother's ethnicity, number of children the mother has given birth to and whether child's birth weight is less than 2kg. Within these subgroups, CATE estimates range from a minimum of +2.1 to a maximum of +12.

Both CATE's posterior splitting probabilities as well as subgroup analysis particularly point to covariates related to mother's education and ethnicity, in addition to birth weight (in the subgroup analysis only). Results concerning heterogeneity stemming from mother's ethnicity and child's birth weight are consistent with those in the original (Brooks-Gunn et al. 1992) and follow-up studies Brooks-Gunn et al. (1994) and McCarton et al. (1997), where the treatment effect is found to be lower for white mothers and for children with lower weight. The advantage of carrying out subgroup analysis through models such as Shrinkage BCF lies in the fact that subgroup identification can be done ex-post using CATE estimates, without the need of manually identifying the groups or partitioning the original sample ex-ante.

This illustrative example showed how Shrinkage BCF detects covariates which are responsible for the heterogeneity behind treatment impact in an example of real-world analysis, and how simple a posteriori partitioning of CATE estimates allows the derivation of optimal splitting rules to identify the most homogeneous subgroups in terms of treatment response. The analysis demonstrated that the estimation of individual (or subgroup) effects is a key factor for the correct evaluation and design of treatment administration policies.

7. Conclusions

In this work, we introduced a sparsity-inducing version of the popular nonparametric regression model Bayesian Causal Forest, recently developed by Hahn, Murray, and Carvalho (2020), in the context of heterogeneous treatment effects estimation. The new version proposed, Shrinkage Bayesian Causal Forest, is based on the contributions by Linero (2018) and Linero and Yang (2018), and differs in the two additional priors that modify the way the model selects the covariates to split on. Shrinkage BCF allows targeted feature shrinkage on the prognostic score and CATE surfaces, and in addition returns posterior splitting probabilities, an interpretable measure of feature importance. In Section 5, we demonstrated its performance on simulated exercises that mimic confounded observational studies where only some covariates are relevant, while the rest of them constitutes nuisance predictors that can cause bias if included in a fully-saturated outcome model. Shrinkage BCF demonstrates competitive performance and scalability compared to the original version of BCF and to other state-of-the-art methods for CATE estimation, that tend to deteriorate with an increasing number of covariates. We also showed that it effectively tackles strong confounding from targeted selection, a property inherited from the BCF parameterization, and illustrated its use on a real-world study.

In the simulated studies of Sections 4 and 5, in addition to those in the supplementary materials, we have investigated Shrinkage BCF's performance on different sparse DGPs. When we consider nonsparse DGPs instead, with few but all relevant covariates, default BCF might be the preferable option, even though Shrinkage BCF would not incur in much higher error.

Besides the implementation of feature shrinkage per se, the additional advantage of Shrinkage BCF specification is that the pair of Dirichlet priors placed on the splitting probabilities can be tailored to incorporate subject-matter knowledge about the importance and impact of the covariates, separately for prognostic score and CATE. Embedding of prior information in a Bayesian fashion represents a way of avoiding a completely agnostic model, that nonetheless benefits from the excellent predictive properties of a nonparametric regression algorithm such as BART. Hence, the informative version of Shrinkage BCF can be useful in applied studies with limited sample size, where a priori knowledge is possessed and can be efficiently incorporated without losing the benefits of using a powerful nonlinear model.

Finally, as highlighted in different parts of the manuscript, we stress how the main advantage of DART (and consequently Shrinkage BCF) over BART (and BCF) is very much computational and improves performance in sparse DGPs settings exclusively. The MCMC convergence in DART is faster, as the model is urged to split more and more eagerly along the most predictive features. However, DART and Shrinkage BCF do not perform variable selection explicitly. An interesting future direction would be to augment DART priors to include predictor-specific inclusion parameters, to completely select out irrelevant predictors.

Supplementary Materials

Supplementary materials include a pdf file describing additional simulated experiments, and code and data to replicate all the simulated and real-world examples.

Funding

This work was supported by a British Heart Foundation-Turing Cardiovascular Data Science Award (BCDSA/100003).

References

Alaa, A., and van der Schaar, M. (2018), "Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design," in *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80), pp. 129–138. [2,8]

Alaa, A. M., and van der Schaar, M. (2017), "Bayesian Inference of Individualized Treatment Effects Using Multi-Task Gaussian Processes," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3427–3435. [2,8]

Athey, S., and Imbens, G. (2016), "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360. [1]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1,4]

Brooks-Gunn, J., Liaw, F., and Klebanov, P. K. (1992), "Effects of Early Intervention on Cognitive Function of Low Birth Weight Preterm Infants," *The Journal of Pediatrics*, 120, 350–359. [10,12]

Brooks-Gunn, J., McCarton, C., Casey, P., McCormick, M., Bauer, C., Bernbaum, J., Tyson, J., Swanson, M., Bennett, F., and Scott, D. (1994), "Early Intervention in Low-Birth-Weight Premature Infants. Results through Age 5 Years from the Infant Health and Development Program," *JAMA*, 272, 1257–1262. [12]

Caron, A., Manolopoulou, I., and Baio, G. (2020), "Estimating Individual Treatment Effects Using Non-parametric Regression Models: A Review," arXiv:2009.06472. [1,3,8]

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935–948. [1,3,5]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [1,3,4,5]

Dawid, A. P. (2000), "Causal Inference Without Counterfactuals," *Journal of the American Statistical Association*, 95, 407–424. [1,2]

——— (2015), "Statistical Causality from a Decision-Theoretic Perspective," *Annual Review of Statistics and Its Application*, 2, 273–303. [2]

Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. (2020), "Estimation of Conditional Average Treatment Effects with High-Dimensional Data," *Journal of Business & Economic Statistics*, 40, 313–327. [2]

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011), "Subgroup Identification from Randomized Clinical Trial Data," *Statistics in Medicine*, 30, 2867–2880. [1]

Green, D. P., and Kern, H. L. (2012), "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees," *Public Opinion Quarterly*, 76, 491–511. [8]

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018), "Regularization and Confounding in Linear Regression for Treatment Effect Estimation," *Bayesian Analysis*, 13, 163–182. [3]

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," *Bayesian Analysis*, 15, 965–1056. [1,2,3,8,10,11,12]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, New York: Springer. [1]

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. [3]

Herren, A., and Hahn, P. R. (2020), "Semi-Supervised Learning and the Question of True Versus Estimated Propensity Scores," arXiv:2009.06183. [3]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [1,6,8,10]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29. [3]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press. [1,2]

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017), "Better Together? Statistical Learning in Models Made of Modules," arXiv:1708.08719. [3]

Johansson, F. D., Shalit, U., and Sontag, D. (2016), "Learning Representations for Counterfactual Inference," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (Vo. 48), pp. 3020–3029. [2]

Künzel, S., Sekhon, J., Bickel, P., and Yu, B. (2017), "Meta-Learners for Estimating Heterogeneous Treatment Effects Using Machine Learning," *Proceedings of the National Academy of Sciences*, 116, 4156–4165. [2,8]

Linero, A. R. (2018), "Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection," *Journal of the American Statistical Association*, 113, 626–636. [2,4,5,12]

Linero, A. R., and Yang, Y. (2018), "Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity," *Journal of the Royal Statistical Society*, Series B, 80, 1087–1110. [2,5,12]

Lu, M., Sadiq, S., Feaster, D. J., and Ishwaran, H. (2018), "Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods," *Journal of Computational and Graphical Statistics*, 27, 209–219. [1]

McCarton, C., Brooks-Gunn, J., Wallace, I., Bauer, C., Bennett, F., Bernbaum, J., Broyles, R., Casey, P., McCormick, M., Scott, D., Tyson, J., Tonascia, J., and Meinert, C. (1997), "Results at Age 8 Years of Early Intervention for Low-Birth-Weight Premature Infants. The Infant Health and Development Program," *JAMA*, 277, 126–132. [12]

McCormick, M. (1985), "The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity," *The New England Journal of Medicine*, 312, 82–90. [10]

McCormick, M. C., Gortmaker, S. L., and Sobol, A. M. (1990), "Very Low Birth Weight Children: Behavior Problems and School Difficulty in a National Sample," *The Journal of Pediatrics*, 117, 687–693. [10]

Nie, X., and Wager, S. (2020), "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," *Biometrika*, 108, 299–319. [2,8]

Pearl, J. (2009a), *Causality: Models, Reasoning and Inference* (2nd ed.), New York: Cambridge University Press. [1,2]

——— (2009b), "Remarks on the Method of Propensity Score," *Statistics in Medicine*, 28, 1415–1416. [2]

——— (2018), "Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining—WSDM '18*. [2]

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018), "Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions," *Statistics in Medicine*, 37, 1767–1787. [2]

Robinson, P. M. (1988), "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. [2,3]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [3]

Ročková, V., and Saha, E. (2019), "On Theory for BART," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (Vol. e 89), pp. 2839–2848. [3]

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58. [2]

Shalit, U., Johansson, F. D., and Sontag, D. (2017), "Estimating Individual Treatment Effect: Generalization Bounds and Algorithms," in *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70), pp. 3076–3085. [2]

Sivaganesan, S., Müller, P., and Huang, B. (2017), "Subgroup Finding via Bayesian Additive Regression Trees," *Statistics in Medicine*, 36, 2391–2403. [8]

Starling, J., Murray, J., Lohr, P., Aiken, A., Carvalho, C., and Scott, J. (2019), "Targeted Smooth Bayesian Causal Forests: An Analysis of Heterogeneous Treatment Effects for Simultaneous Versus Interval Medical Abortion Regimens Over Gestation," *The Annals of Applied Statistics*, 15, 1194–1219. [2]

Traskin, M., and Small, D. S. (2011), "Defining the Study Population for an Observational Study to Ensure Sufficient Overlap: A Tree Approach," *Statistics in Biosciences*, 3, 94–118. [1]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [1,4,8]

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018), "Representation Learning for Treatment Effect Estimation from Observational Data," in *Advances in Neural Information Processing Systems*, 31, 2633–2643. [2]

Zhao, Q., Small, D. S., and Ertefaie, A. (2018), "Selective Inference for Effect Modification via the Lasso," arXiv:1705.08020. [2]

Zigler, C., Watts, K., Yeh, R., Wang, Y., Coull, B., and Dominici, F. (2013), "Model Feedback in Bayesian Propensity Score Estimation," *Biometrics*, 69, 263–273. [3]

Zigler, C. M., and Dominici, F. (2014), "Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects," *Journal of the American Statistical Association*, 109, 95–107. [3]

Zimmert, M., and Lechner, M. (2019), "Nonparametric Estimation of Causal Heterogeneity Under High-Dimensional Confounding," arXiv:1908.08779. [2]