Bayesian Analysis (0000) **00**, Number 0, pp. 1

Discussion of "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects"

Georgia Papadogeorgou[*] , and Fan Li[†]

We congratulate Hahn, Murray and Carvalho (henceforth referred to as HMC) for their important contribution to the growing field of *Bayesian causal inference*. The authors tackle the problem of estimating treatment effect heterogeneity through conditional average treatment effects (CATE), a hard problem that can be framed within the context of high dimensional causal inference and multiple hypothesis testing. HMC adopt a flexible outcome model and they discuss the importance of prior specifications within the Bayesian framework that achieve a two-fold goal: appropriate confounding adjustment for unbiased effect estimation, and sufficient flexibility for estimation of heterogeneity. We regard the separation of these two components within the modeling framework as the most attractive feature of the proposed approach. First, the authors stress the importance of including the estimated propensity scores in the outcome model for more accurate confounding adjustment. Even though the use of propensity scores in Bayesian causal inference is subject to debate (see Section 1), we believe it is an important message to stress the central role of the propensity score in causal inference, irrespective of the mode of inference. Second, we believe that estimation of CATE is rightfully placed within the Bayesian framework in which the model formulation allows for heterogeneity along any covariate while shrinking small signals towards homogeneity. By viewing estimation of CATE within the scope of multiple testing, the Bayesian framework provides an inherent and automatic way for multiplicity control (Scott and Berger, 2010), while acknowledging and documenting uncertainty in estimating heterogeneity along all covariates.

With our discussion, we hope to shed light to the following aspects of this approach: (1) What is the role of the propensity score in Bayesian causal inference? (2) How does shrinkage towards homogeneity drive estimation of CATE? (3) What is the interplay between the choice of non-parametric prior distribution and limited covariate overlap in terms of uncertainty quantification in the estimation of CATE?

1 The role of propensity score in Bayesian causal inference and double-robustness

There has been a long debate of the role of propensity score in Bayesian causal inference (Sims, 2006; McCandless et al., 2009; Robins and Wasserman, 2012; Zigler et al.,

[*] Assistant Professor, Department of Statistics, University of Florida, Gainesville, FL 32611 gpapadogeorgou@ufl.edu url: https://gpapadogeorgou.netlify.app

[†] Associate Professor, Department of Statistical Science, Duke University, Durham, NC 27708 fl35@duke.edu url: http://www2.stat.duke.edu/~fl35

DOI: 0000

2013; Robins et al., 2015). On one hand, under the assumption of ignorable assignment mechanism (and the parameters of the propensity score model and outcome model are *a priori* independent and distinct) (Ding and Li, 2018), the propensity score drops out from the likelihood of the outcomes, and therefore in principle does not matter in Bayesian causal inference. On the other hand, vast empirical evidences (including HMC) suggest that proper inclusion of the propensity score substantially improves Bayesian causal inference. To understand the reason of adding the propensity score as a predictor in the outcome model, it is important to first clarify this seemingly paradoxical phenomenon. First, Robins et al. (2015) pointed out that the propensity score, as a one-dimensional summary of the covariates, is crucial for dimension reduction in causal inference with high dimensional data, irrespective of the mode of inference. A second important insight comes from the frequentist's perspective. Belloni et al. (2014) show that good performance in predicting either the observed outcome or the propensity score alone does not necessarily translate into good performance in estimating the causal effects. Chernozhukov et al. (2018) further pointed out that it is necessary to combine high-dimensional (e.g. machine learning) models for estimating the propensity score and outcome models in order to achieve $\sqrt{N}$ consistency in estimating the average treatment effect. These insights speak to the necessity of combining propensity score and outcome models for estimating causal effects in high-dimensional settings. This is closely related to the class of double-robust (DR) estimators (Scharfstein et al., 1999; Lunceford and Davidian, 2004; Bang and Robins, 2005). An estimator is DR if it is consistent as long as either the propensity score model or the outcome model, but not necessarily both, is correctly specified. Though the concept of DR was originally developed in the form of inverse probability weighting, many different forms have since been proposed, the essence of which is to augment a propensity score estimator (weighting, stratification, or matching) by an outcome model or vice versa.

We view HMC's proposal as a Bayesian analogue of DR. A similar frequentist proposal, due to Rod Little and co-authors, uses the propensity score (in the form of penalized splines) as an additional predictor in the outcome model (e.g. Little and An, 2004; Zhou et al., 2019). In a sense, both HMC and Little's methods can be viewed as a continuous version of the mixed approach of combining propensity score stratification and outcome modeling (Gutman and Rubin, 2013): conducting an outcome regression within the stratum of a specific propensity score range. While the theoretical and empirical advantages of the DR estimators over non augmented-estimators have been well established in the frequentist's paradigm (e.g., Funk et al., 2011), HMC are among the first to do so in the Bayesian paradigm (another example is Antonelli et al. (2018)), which also naturally expands the domain of DR from estimating average treatment effects to CATEs. Moving forward, we believe it would be worthwhile to rigorously define the Bayesian analog of DR (a definition of which is given by Antonelli et al. (2018)) and prove the DR property of the specific prior. This will allow analysts to generalize from the BART prior to other Bayesian nonparametric priors (See Section 3).

2 Homogeneity-induced bias: The bias-variance trade-off of compromising between separate and simultaneous modeling

BART has a number of advantages, remarkably its robustness to the choice of hyperparameters, making it an attractive tool for causal effect estimation, including in the presence of heterogeneity. So there is substantial merit in understanding *how BART can be used best* for estimating CATEs. In this section, we dive into the implications of imposing a prior distribution that shrinks towards homogeneity of treatment effects.

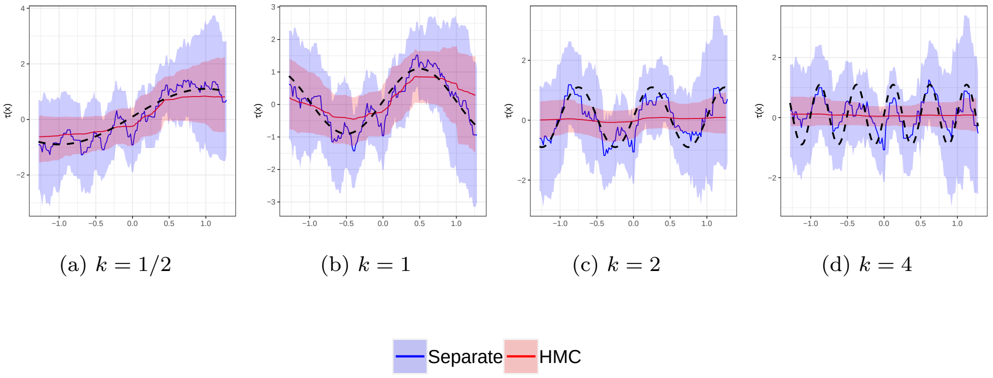
For simplicity, we consider a single covariate X. We focus on the CATE $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ and on three ways to use BART: (Separate) adopt independent BART priors to model Y given X within the treated and control groups, (Simultaneous) specify $E[Y|Z, X]$ using a BART prior, and (HMC) the proposed approach. The separate model provides full flexibility for estimating heterogeneous treatment effects since the mean potential outcome under each treatment arm is modeled separately and without any sharing of information between treated and control groups. In contrast, the simultaneous BART model uses *the same* tree structure for estimating the two potential outcomes, and treatment effect heterogeneity is described via tree branches that split *both* on the treatment *and* the covariate. We view the approach of HMC as a "compromise" between separate and simultaneous modeling. By adopting two BART priors, the authors allow for flexible modeling of $E[Y(0)|X]$ and penalize deviations of $E[Y(1)|X]$ towards homogeneity of treatment effects. In that sense, their model formulation can be thought of as including a tree structure that forces the first split along Z. Therefore, the proposed approach shares information among treated and control groups, but ensures that shrinkage of estimates of $\tau(x)$ are towards homogeneity.

We compare the full flexibility of separate modeling to the flexibility using the approach by HMC in estimating the CATE. We consider a data generation scenario for which the covariate X is generated from $X \sim N(0, 1)$, and treatment and outcome are generated as

$$Z \mid X \sim \text{Bernoulli}(p(X)),\ \text{logit}(p(X)) = 2X$$
$$Y \mid Z, X \sim \text{N}(0.1Z + X^2 + Z \sin(kX\pi), 1).$$

In this case, $\tau(x) = 0.1 + \sin(kx\pi)$. We consider values of $k \in \{1/2, 1, 2, 4\}$ describing increasingly complex CATE. The dashed black lines in Figure 1 show the true values of $\tau(x)$. We estimate $\tau(x)$ based on a sample of size 300, and for values of x for which covariate overlap exists (within the 0.1 and 0.9 quantiles of the standard normal distribution). We use the standard implementations of BART in the `BayesTree` R package for the separate model, and the R package `bcf` by HMC employing the true and *known* propensity score.

Figure 1 shows estimates of $\tau(x)$ and pointwise 95% credible intervals for the separate and HMC methods. When $\tau(x)$ varies smoothly as a function of x ($k \in \{1/2, 1\}$), the model of HMC is more efficient for estimating the CATE than the separate model. However, when $\tau(x)$ varies quickly with x ($k \in \{2, 4\}$), HMC's approach mistakenly



(a) $k = 1/2$ (b) $k = 1$ (c) $k = 2$ (d) $k = 4$

Figure 1: Estimates of heterogeneous treatment effects $\tau(x)$ and pointwise 95% credible intervals for the separate model (blue) and the model of HMC (red) based on a sample of size 300 and increasing complex heterogeneity structure. The dashed black lines show the true value of $\tau(x)$.

shrinks effect estimates towards homogeneity. These results indicate that, compared to the separate model, the approach by HMC has lower variance, but has the potential to lead to *homogeneity-induced bias*. Even though this bias is expected to be eliminated as sample size increases, this illustration indicates that the homogeneity-inducing BART prior might lead to large biases in the presence of a highly-complex heterogeneity structure.

In most applications, highly complex heterogeneity across a *single* covariate like the one in Figure 1 for $k \in \{2, 4\}$ is unlikely to exist. However, these scenarios are representative of situations with *many* covariates for which heterogeneity along each covariate separately might be small, but the true heterogeneity structure includes interactions among multiple covariates. In these situations, the homogeneity-inducing prior is expected to fail to identify the true complexity of HTE. In that sense, if the “subgroup” for which treatment is most effective is defined in terms of multiple subject characteristics, we suspect that the homogeneity-inducing prior would lead to estimates of $\tau(x)$ that fail to identify it. This bias-variance trade-off between CATE complexity and efficient estimation will be exacerbated by having a limited number observations in an area or the covariate space, and perhaps limited overlap.

3 Choice of Bayesian nonparametric prior distribution

BART is one example of the general class of Bayesian nonparametric priors. A natural question would be “what about other priors?” In the context of CATE estimation, a particular relevant issue is (covariate) overlap and uncertainty quantification. Specifically, HMC demonstrated that the advantages in terms of point estimates of their BART prior over several other methods vary according to the degree of confounding, which is commonly referred to as overlap in the standard terminology of causal inference. Heuristically speaking, overlap means the similarity of the covariate distributions between the treatment and control groups. Overlap is a key concept that separates causal inference from traditional two-sample inference. In the region with good overlap (or low confounding in HMC's terminology), different specification of the outcome model usually leads to similar causal estimates. In contrast, in the region with poor overlap, the uncertainty of causal estimation is much higher and the results are sensitive to the specification of the outcome model. Thus, besides point estimation, an important criterion in choosing the outcome model for CATE estimation is proper uncertainty quantification.

Given its underlying tree structure, intuitively BART may not have the flexibility to capture the additional uncertainty in regions of poor overlap, whereas some other "smoother" Bayesian nonparametric models such as the Gaussian Process may fare better. We demonstrate this through a toy example thanks to Surya Tokdar (who grants us the permission). Consider a scenario where a single covariate 'age' influences both treatment assignment and a continuous outcome; younger people are more likely to receive the treatment and higher outcome scores. Specifically, we generate a sample with 200 treated units and 300 control units. The only covariate (X_i) follows a Gamma distribution with mean 60 and 35 in the control and treatment group, respectively, and with standard deviation 8 in both groups. The true outcome model is as follows:

$$Y_i = 90 \cdot \mathbf{1}\{Z_i = 1\} + 82 \cdot \mathbf{1}\{1 - Z_i\} - 0.2X_i + \epsilon_i,\ \epsilon_i \sim N(0, 1).$$

As shown in the data cloud by group in Figure 2, the above data generating process creates a case where the degree of overlap varies significantly across the range of the single covariate: there is good overlap around age of 40, but lack of overlap increases as we move to the two tails of the range of age.

To estimate the CATE, we fit the following outcome model separately in the treatment and control groups: $Y_i = f(X_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We choose three prior specifications for $f(x)$: (1) a BART prior similarly to Hill (2011) and HMC but without the propensity score; (2) a linear model with Gaussian prior: $f(x) \sim N(\beta x, \delta^2)$; (3) a GP prior (Rasmussen, 2003) with the covariance function specified using a Gaussian kernel with signal-to-noise ratio parameter ρ and inverse-bandwidth parameter λ: $(f(x_1), f(x_2), \ldots, f(x_n))^T \sim N(0, \Sigma)$ where $\Sigma_{ij} = \sigma^2 \rho^2 \exp\{-\lambda^2 \|x_i - x_j\|^2\}$). For each unit, we predict the missing counterfactual outcome by plugging the covariate into the fitted model of the opposite group, and then estimate the individual treatment effect by the difference between observed outcome and predicted counterfactual outcome. Figure 2 shows the predicted counterfactuals and the CATE, as well as the associated uncertainty band as a function of age. The true effects curve is deliberately omitted to focus on issue of uncertainty quantification. In the region of good overlap, all three models lead to similar point and interval estimates of CATE. However, marked difference emerges in the region of poor overlap. Here the linear model appears overconfident in predicting counterfactuals and thus estimating the CATE. GP trades potential bias with increased uncertainty bands as overlap decreases and produces a more adaptive uncertainty quantification. BART produces shorter error bars than GP (but wider than LM), but width

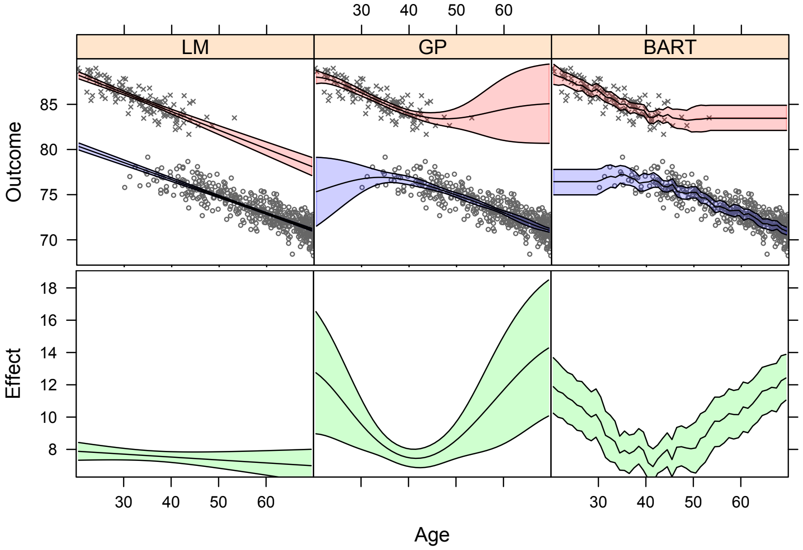


Figure 2: Estimates of missing counterfactuals (upper panel) and CATE (lower panel) and corresponding uncertainty band as a function of the single covariate 'Age' by three different models: linear model (LM); Gaussian Process (GP); BART, in the example of *poor overlap (or high confounding)* in Section 3. ×: treated units; ∘: control units.

of the uncertainty band remains similar regardless of the degree of overlap, which is clearly over confident in the region of poor overlap. This pattern is not surprising given how GP and BART are constructed, on which we do not elaborate here.

Of course, it is prudent not to overly generalize the message from a single example. Nonetheless, we feel that deeper investigation is warranted on how BART-based and other Bayesian nonparametric priors perform in terms of uncertainty quantification of CATE under different degree of overlap, which is a central problem in causal inference with high dimensional data and high dimensional estimation targets.

References

Antonelli, J., Papadogeorgou, G., and Dominici, F. (2018). "Causal Inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties." *arXiv preprint arXiv:1805.04899*. 2

Bang, H. and Robins, J. (2005). "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics*, 61: 962–972. 2

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies*, 81(2): 608–650. 2

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21(1): C1–C68. 2

Ding, P. and Li, F. (2018). "Causal inference: A missing data perspective." *Statistical Science*, 33(2): 214–237. 2

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). "Doubly robust estimation of causal effects." *American journal of epidemiology*, 173(7): 761–767. 2

Gutman, R. and Rubin, D. B. (2013). "Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes." *Statistics in Medicine*, 32(11): 1795–1814. 2

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. 5

Little, R. and An, H. (2004). "Robust likelihood-based analysis of multivariate data with missing values." *Statistica Sinica*, 949–968. 2

Lunceford, J. and Davidian, M. (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." *Statistics in Medicine*, 23: 2937–2960. 2

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). "Bayesian propensity score analysis for observational data." *Statistics in medicine*, 28(1): 94–112. 1

Rasmussen, C. E. (2003). "Gaussian processes in machine learning." In *Summer School on Machine Learning*, 63–71. Springer. 5

Robins, J. and Wasserman, L. (2012). "Robins and Wasserman Respond to a Nobel Prize Winner."
URL `https://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/` 1

Robins, J. M., Hernán, M. A., and Wasserman, L. (2015). "On Bayesian estimation of marginal structural models." *Biometrics*, 71(2): 296. 2

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). "Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion)." *Journal of the American Statistical Association*, 94: 1096–1146. 2

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 2587–2619. 1

Sims, C. (2006). "On an example of Larry Wasserman." *online manuscript, available from http://sims. princeton. edu/yftp/WassermanExmpl/WassermanComment. pdf*, 2(10). 1

Zhou, T., Elliott, M. R., and Little, R. J. (2019). "Penalized Spline of Propensity

Methods for Treatment Comparison." *Journal of the American Statistical Association*, 114(525): 1–19. 2

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. 1