Contributed Discussion

Estevão B. Prado[†,‡,§,*], Eoghan O'Neill[¶], Belinda Hernández[‖],
Andrew C. Parnell[†,‡,§], and Rafael A. Moral[†,‡]

1 Discussion

We congratulate the authors for their stimulating and excellent work on applying Bayesian trees to causal inference modelling. In this discussion, we extend the authors' work by evaluating the models on higher dimensional data sets. We remark in passing that it seems odd that the paper only contains model performance metrics on training data, but to allow for valid comparisons, we follow their approach. In particular we show that, for higher dimensions, some of the existing (non-causal) models have equivalent or superior performance to BCF for the simulations used in the paper.

We carried out a small simulation study to investigate the performance of BCF, ps-BART, Causal RF and other methods. We extend the simulations carried out in the paper by setting $n = 250$, $p = (5, 50, 100, 500)$ and consider the structure for $\tau(\mathrm{x})$, $\mu(\mathrm{x})$ and $\pi(\mathrm{x})$ presented in Section 6.1. By varying the number of covariates, we aim to see how BCF and ps-BART behave as well as motivate other algorithms that are designed to deal with large p. For instance, Hernández et al. (2018) propose a BART-based algorithm suitable for high-dimensional data (in particular when $p > 10,000$), named BART-BMA, that uses Bayesian model averaging and does not utilise an MCMC algorithm. Via simulation studies, they show that BART-BMA outperforms the standard BART when the number of covariates is large. In this context, Linero (2018) introduces Dirichlet BART (DART) that modifies the variable selection in BART by updating the probability of a predictor being selected as a split variable via a Dirichlet distribution. With this change, DART tends to produce more accurate predictions than BART in situations where p is large. We also explored MOTR-BART (Prado et al., 2020), which is an algorithm that generalises BART by generating the predictions based on piece-wise linear functions rather than terminal node constants. Here, we are specially interested in seeing how MOTR-BART (with 10 trees) performs when $\mu(\mathrm{x})$ is linear, as the regularised linear regression with the horseshoe prior presented the best results in Table 2 of Section 6.1.

In Figure 1, we present the RMSE obtained from 50 Monte Carlo simulations for the Conditional Average Treatment Effect (CATE), and measured the methods' performance on the training sets considering an estimate of the propensity score as a covariate for the non-causal algorithms. When $p = 500$, it was not possible to run BCF

[*]Corresponding author: estevao.prado@mu.ie.
[†]Hamilton Institute, Maynooth University, Ireland
[‡]Department of Mathematics and Statistics, Maynooth University, Ireland
[§]Insight: The National Centre for Data Analytics, Maynooth University, Ireland
[¶]Econometric Institute, Erasmus University Rotterdam, Netherlands
[‖]School of Medicine, Trinity College Dublin, Ireland

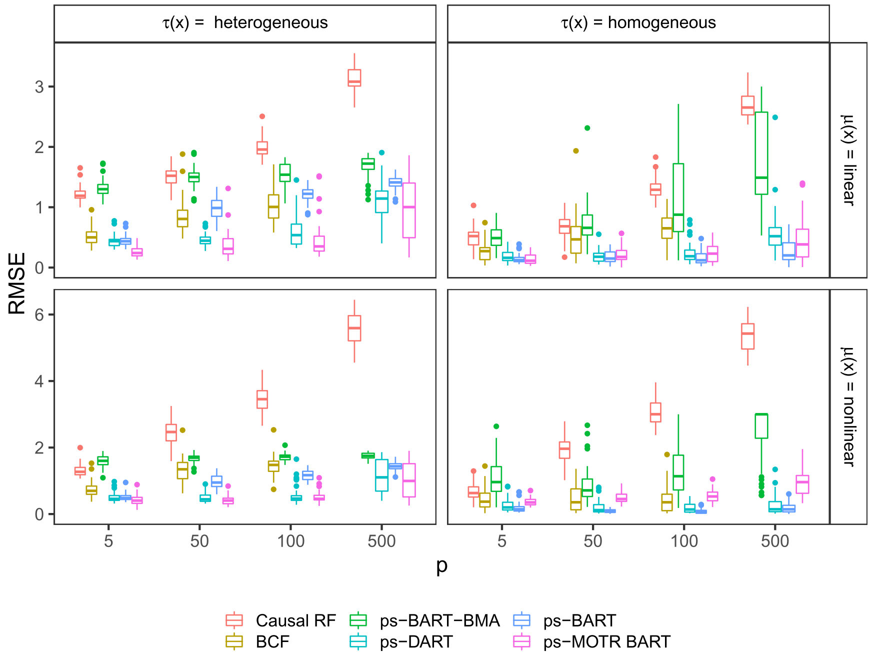

Figure 1: Simulation study results of RMSE for Conditional Average Treatment Effect (CATE).

due to numerical errors. Firstly, we see that the Causal RF algorithm is highly sensitive to the number of covariates. For all combinations of $\tau(\mathrm{x})$, $\mu(\mathrm{x})$ and estimands, the RMSE values for Causal RF tend to increase as p gets bigger. When $\tau(\mathrm{x})$ is heterogeneous, we see that ps-MOTR-BART is competitive even when the structure of $\mu(\mathrm{x})$ is nonlinear. In addition, ps-DART presents lower RMSE values than ps-BART, which might suggest that ps-DART could be an alternative in situations where p is large (Santos and Lopes, 2018). On the right-hand side of Figure 1, the results of RMSE for CATE are shown when $\tau(\mathrm{x})$ is homogeneous, which we believe is unrealistic in practice. Here, we see that for $p < 500$ and linear $\mu(\mathrm{x})$ that ps-BART, ps-DART and ps-MOTR-BART present excellent levels of accuracy. For instance, when $p = (5, 50 \text{ and } 100)$, they produce similar results with the three generating more accurate estimates than BCF. With $\tau(\mathrm{x})$ homogeneous and a nonlinear structure for $\mu(\mathrm{x})$, however, we note that MOTR-BART does not produce as accurate estimates as ps-BART and ps-DART.

In Figure 2, we see that the results for the Average Treatment Effect (ATE) are similar. That is, ps-BART and ps-DART perform well across all simulations and ps-MOTR-BART performs particularly well in data with heterogeneous effects and linear $\mu(\mathrm{x})$. Also, BCF does not give better results than ps-BART or ps-DART in any setting, although it outperforms Causal RF and ps-BART-BMA.

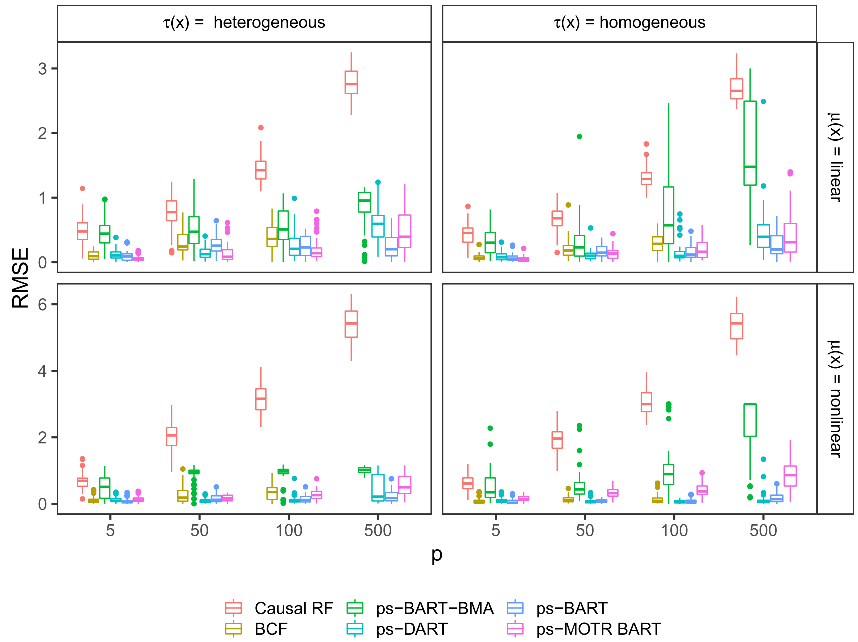

Figure 2: Simulation study results of RMSE for Average Treatment Effect (ATE).

Although not shown, we also explored the methods' performance on test data. Perhaps due to the non-stochastic nature of the simulation equations, we did not observe large differences between the results in training versus test performance. The results presented here can be reproduced by using the R scripts available at https://github.com/ebprado/BCF-discussion-paper.

References

Hernández, B., Raftery, A. E., Pennington, S. R., and Parnell, A. C. (2018). "Bayesian additive regression trees using Bayesian model averaging." *Statistics and Computing*, 28(4): 869–890. MR3766048. doi: https://doi.org/10.1007/s11222-017-9767-1. 1029

Linero, A. R. (2018). "Bayesian regression trees for high-dimensional prediction and variable selection." *Journal of the American Statistical Association*, 113(522): 626–636. MR3832214. doi: https://doi.org/10.1080/01621459.2016.1264957. 1029

Prado, E. B., Moral, R. A., and Parnell, A. C. (2020). "Bayesian additive regression trees with model trees." *arXiv preprint arXiv:2006.07493*. MR3858513. doi: https://doi.org/10.4310/SII.2018.v11.n4.a1. 1029

Santos, P. H. F. d. and Lopes, H. F. (2018). "Tree-based bayesian treatment effect analysis." *arXiv preprint arXiv:1808.09507*. 1030