Statistical Science
2019, Vol. 34, No. 1, 90–93
https://doi.org/10.1214/18-STS682
Main article: https://doi.org/10.1214/18-STS667

Comment: Contributions of Model Features to BART Causal Inference Performance Using ACIC 2016 Competition Data

Nicole Bohme Carnegie

Abstract. With a thorough exposition of the methods and results of the 2016 Atlantic Causal Inference Competition, Dorie et al. have set a new standard for reproducibility and comparability of evaluations of causal inference methods. In particular, the open-source **R** package `aciccomp2016`, which permits reproduction of all datasets used in the competition, will be an invaluable resource for evaluation of future methodological developments.

Building upon results from Dorie et al., we examine whether a set of potential modifications to Bayesian Additive Regression Trees (BART)—multiple chains in model fitting, using the propensity score as a covariate, targeted maximum likelihood estimation (TMLE), and computing symmetric confidence intervals—have a stronger impact on bias, RMSE, and confidence interval coverage in combination than they do alone. We find that bias in the estimate of SATT is minimal, regardless of the BART formulation. For purposes of CI coverage, however, all proposed modifications are beneficial—alone and in combination—but use of TMLE is least beneficial for coverage and results in considerably wider confidence intervals.

Key words and phrases: Bayesian additive regression trees, TMLE, propensity score.

1. INTRODUCTION

The 2016 Atlantic Causal Inference Conference (ACIC) competition provided a platform for competing causal estimation methods to be compared on a wide range of common datasets. Dorie et al. have written a very detailed exposition of the competition set-up and results [3]. I congratulate the authors on setting a new standard for reproducibility and generalizability when comparing competing causal methods. In addition, the open-access **R** package `aciccomp2016`, which can generate the 7700 datasets used in the 2016 ACIC competition (and new datasets besides) is a tremendous tool for future methodological innovations to be benchmarked against a wide array of other methods, including those from this competition [5, 2].

The results with competition and post-competition methods give a fascinating look at the subtleties of model choice. A key result explored the contributions of different features of estimation methods (e.g., weighting or fitting a propensity score). It is intriguing that there are few relationships between the bias of different methods and these features, once a basic requirement of flexibility in fitting the response surface is satisfied. It is worth noting, however, that both ignorability and overlap assumptions were satisfied in all datasets used for this competition; the role of modeling treatment assignment may well be more important when either or both of these assumptions are not fully met.

The prominence of flexible response surface fitting suggests that further exploration of nonparametric and ensemble methods will be valuable. Dorie

Nicole Bohme Carnegie is Assistant Professor of Statistics, Department of Mathematical Sciences, Montana State University, PO Box 172400, Bozeman, Montana 59717, USA (e-mail: nicole.carnegie@montana.edu).

et al. propose a number of modifications to models with Bayesian Additive Regression Trees (BART, [1]), mostly in parallel. In my evaluation, I explore whether a similar approach could reveal what components of the BART fit are most predictive of changes in bias and confidence interval coverage, and whether combining elements of the modifications (TMLE, multiple chains, cross-validation, including propensity score, and symmetric intervals) would further improve performance.

2. EVALUATING BART MODIFICATIONS

I will refer to the original BART submission—with a single chain, quantile-based intervals, and no other modifications—as "base BART". For purposes of comparison, I included all combinations of the following settings:

- With and without propensity score as covariate (PS) [4]
- With and without TMLE + IPTW (TMLE) [6]
- With 1, 4, or 10 chains in estimation (CH)
- With quantile-based (q) or symmetric (sym) confidence intervals (CIs)

This yielded a total of 12 model types, each with two different CIs. Each type was fit to the 7700 combinations of 77 data generating processes (DGPs) and 100 replications used in the original competition.

For each model fit, I evaluated the bias, root mean squared error (RMSE), confidence interval coverage (nominal 95%), and length of confidence intervals. All measures except CI coverage are in units of the standard deviation of the response, to ensure comparability across DGPs.

Figure 1 gives the results for bias (circles) and RMSE (triangles), ordered by increasing average bias. The bias is generally small, with the inter-quartile range (IQR) of biases across the DGPs and replications far outweighing the differences between methods. The width of the IQR, and hence the RMSE, is larger when TMLE is used. Bias appears to be smaller when the propensity score is included in the covariate set.

When we consider confidence intervals, however, coverage is closest to nominal when including the propensity score as a covariate *and* using TMLE with symmetric intervals, as can be seen in Figure 2. The methods are ordered by increasing coverage (represented with circles), and CI length is represented with triangles. The coverage in this best-case scenario is only marginally greater than using propensity score and symmetric intervals but no TMLE, however, and the intervals are substantially wider.

If we examine the ordering of methods in the plots, it appears that including the propensity score reduces bias and increases CI coverage; using TMLE increases bias, both in terms of average magnitude and variability, but increases CI coverage; and it is less clear what the effect of using multiple chains is. We can examine this further using regression models. First, to evaluate contributions to bias, we fit a linear mixed model with log absolute bias as the response, including random effects for the data generating process, and fixed effects

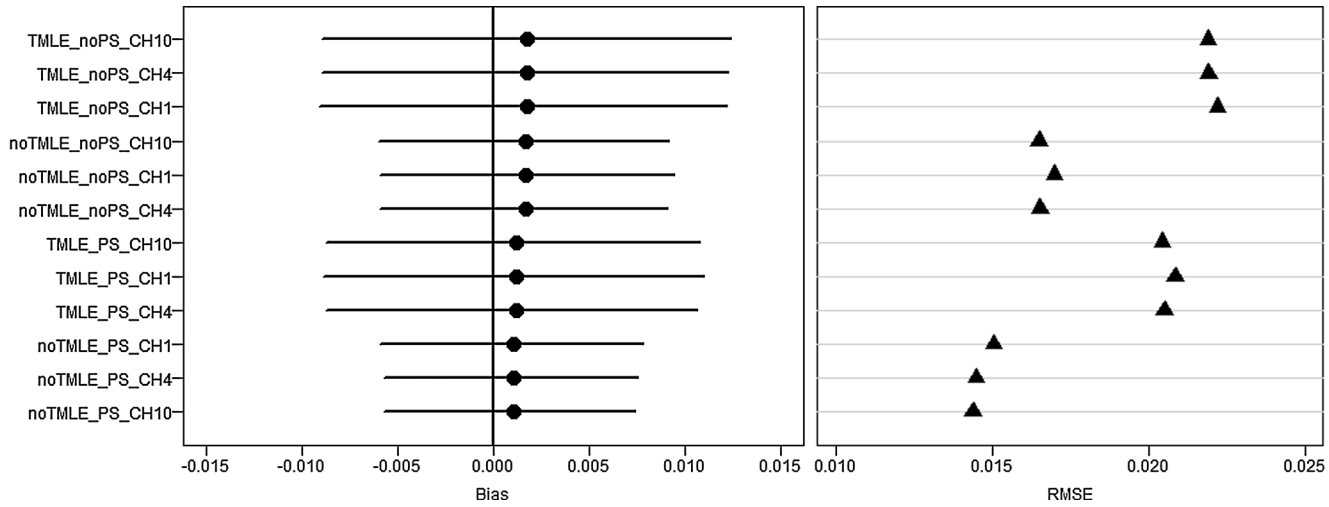

FIG. 1. *Bias (L) and RMSE (R) for all combinations of BART model settings, computed across the combinations of* 77 *data generating processes and* 100 *replications used in the* 2016 *ACIC causal inference challenge. Methods are ordered on the y-axis by increasing average bias.*

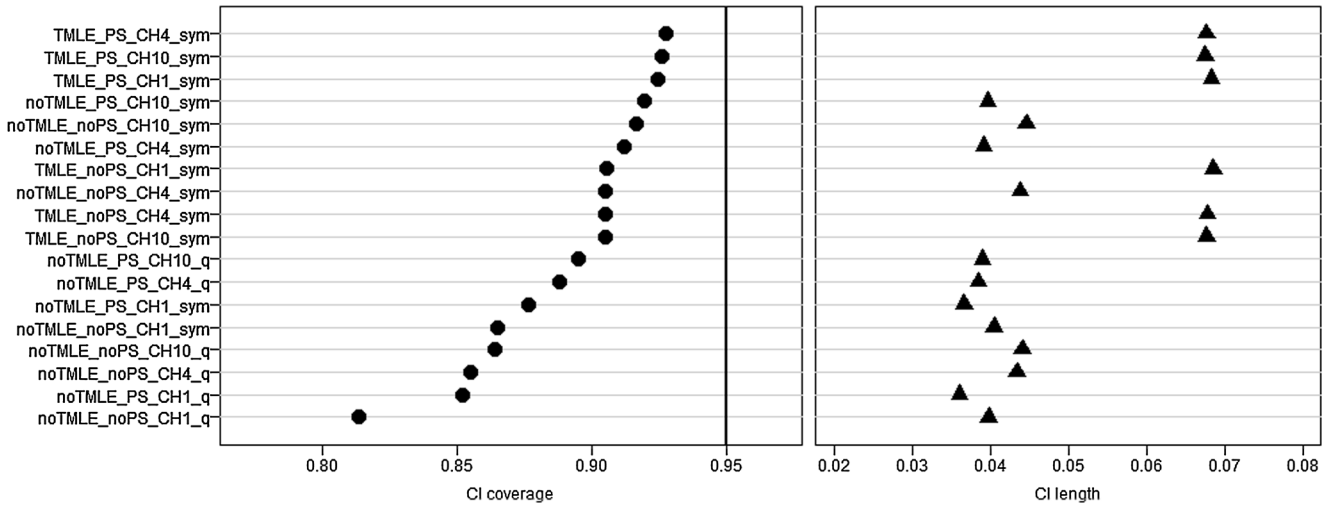

FIG. 2. *Confidence interval coverage rate* (*L*, *nominal* 95%) *and interval length* (*R*) *for all combinations of BART model settings*, *computed across the combinations of* 77 *data generating processes and* 100 *replications used in the* 2016 *ACIC causal inference challenge*. *Methods are ordered on the y-axis by increasing coverage rate*. *A line at the nominal* 95% *coverage rate is included for reference*.

for the number of chains, inclusion of the propensity score, and use of TMLE.

The effects of the various modifications are strongly statistically significant (Table 1), as would be expected for even weak associations in a dataset with 92,400 observations. The effects are small, however: the average absolute bias for base BART is only 0.006 SD units after accounting for variation due to setting, with an estimated 2.3% decrease using 4 chains, 2.7% decrease using 10 chains, 10.5% decrease when including the propensity score, and 42.5% increase when using TMLE. Thus, if our primary consideration is bias in effect estimation, it makes little difference which model type we choose.

This is supported by evaluation of the components of variability in log absolute bias using a model with random effects for method, DGP, and their interaction. The results suggest that the vast majority of variation across observations is random error at the replication level: of a total variance of 1.51, we attribute only 0.038 (2.5%) to method, 0.080 (5.3%) to setting, and essentially zero to their interaction. This agrees with our observations from Figure 1; there was very little variation in the average bias across methods, and a great deal more within.

Where method choice appears to make a substantive difference is in uncertainty quantification and CI coverage (Table 2). For examining confidence interval coverage, I used a generalized linear mixed model (GLMM)—specifically, logistic regression with random effects for DGP. For base BART, the model estimates an odds of coverage of 5.0 (corresponding to 83.4% coverage), after accounting for variability due to DGP. All of the modifications considered improve those odds by 22 to 45 percent, on average. The largest effect is from using symmetric intervals, followed by adding chains to estimation, using propensity score,

TABLE 1
Contribution of method elements to log absolute bias. *Results from a linear mixed-effects model with random effects for data generating process*

	Estimate	Std. Error	t-statistic	p-value
Intercept	−5.04	0.033	−150.7	<0.0001
CH4	−0.024	0.009	−2.49	0.013
CH10	−0.027	0.009	−2.89	0.004
PS	−0.111	0.008	−14.3	<0.0001
TMLE	0.354	0.008	45.7	<0.0001

TABLE 2
Contribution of method elements to probability of confidence interval coverage. *Results from a generalized linear mixed model with random effects for data generating process*

	Estimate	Std. Error	t value	p-value
Intercept	2.04	0.06	26.4	<0.0001
Sym	0.37	0.02	17.9	<0.0001
CH4	0.27	0.02	12.6	<0.0001
CH10	0.33	0.02	15.5	<0.0001
PS	0.22	0.02	12.6	<0.0001
TMLE	0.20	0.02	8.82	<0.0001

and, finally, inclusion of TMLE. The best combination is using 4 chains, TMLE, including propensity score as a covariate, and computing symmetric CIs: 92.7% coverage, mean length 0.068 SD units. The best combination without TMLE is 10 chains, including propensity score, and symmetric intervals: 91.9% coverage, mean length 0.040 SD units. Thus, when optimizing CI coverage, it is most important to use symmetric intervals (45% increase in odds of coverage), and using multiple chains in estimation. Both including the propensity score and using TMLE also appear to be advantageous, but the improvements in coverage with TMLE come at the expense of substantially wider intervals.

3. DISCUSSION

In the end, differences in bias between methods were miniscule, both relative to the variability in the response and relative to the IQR of bias magnitudes across DGPs and replications. Not using TMLE resulted in considerably narrower IQRs, and slightly smaller mean bias, but any of the methods considered are likely to do well with respect to bias, assuming ignorability and overlap conditions are met.

When we consider CI coverage and length, all combinations have better coverage than base BART (no propensity score in the covariate set, no TMLE, single chain, quantile-based intervals). The use of symmetric intervals appears to be key in improving CI coverage; it is not perfectly clear why this should be. Using TMLE improves coverage, but also results in considerably wider CIs. We can get comparable coverage with 40% narrower intervals when using multiple chains, including the propensity score as a covariate, and using symmetric intervals. Even in the best-case scenario, however, we do not quite achieve nominal coverage.

As new approaches are developed in the future to deal with coverage issues, the availability of the competition data for evaluation will be invaluable. The breadth of DGPs used make comparisons across methods more equitable; a small simulation study can easily (and unconsciously) be tailored to situations in which a proposed method performs well. In addition, the availability of published results for a wide selection of methods will permit broad comparisons with minimal computational time.

REFERENCES

[1] CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

[2] DORIE, V. (2017). aciccomp2016: Atlantic Causal Inference Conference Competition 2016 Simulation. R package version 0.1-0.

[3] DORIE, V., HILL, J., SHALIT, U., SCOTT, M. and CERVONE, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.* **34** 43–68.

[4] HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at arXiv:1706.09523 [stat.ME].

[5] R CORE TEAM (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[6] VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500