Statistical Science
2019, Vol. 34, No. 1, 82–85
https://doi.org/10.1214/18-STS689
Main article: https://doi.org/10.1214/18-STS667

Comment on "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition"

Susan Gruber and Mark J. van der Laan

Abstract. Dorie and co-authors (DHSSC) are to be congratulated for initiating the ACIC Data Challenge. Their project engaged the community and accelerated research by providing a level playing field for comparing the performance of a priori specified algorithms. DHSSC identified themes concerning characteristics of the DGP, properties of the estimators, and inference. We discuss these themes in the context of targeted learning.

Key words and phrases: Targeted learning, causal inference, TMLE.

1. INTRODUCTION

Dorie and co-authors (DHSSC) are to be congratulated for initiating the ACIC Data Challenge. Their project engaged the community and accelerated research by providing a level playing field for comparing the performance of a priori specified algorithms. The focus of the challenge was on estimation of the statistical parameter, $\psi_0 = E[E(Y \mid Z = 1, X \mid Z = 1) - E(Y \mid Z = 0, X \mid Z = 1)]$, with a causal interpretation as the sample average effect of treatment among the treated (ATT) guaranteed by the organizers. DHSSC designed data generating processes (DGP) that posed a variety of challenges to the different estimators. The DGPs varied according to six main characteristics, or *knobs*. DHSSC identified themes concerning characteristics of the DGP, properties of the estimators, and inference. We discuss these themes in the context of targeted learning.

Targeted learning is concerned with the construction of data adaptive estimators of a parameter of the probability distribution (P_0) of the data, while relying only on realistic statistical assumptions (van der Laan and Rose, 2011). Our entry in the competition, SL+TMLE, used data adaptive super learning (SL) to estimate the response surface and the treatment assignment mechanism, denoted with Q_0 and g_0, respectively. SL is an ensemble machine learning algorithm for prediction (van der Laan, Polley and Hubbard, 2007, Polley and van der Laan, 2010). However, SL's optimality properties are with respect to a global loss function. Targeted minimum loss-based estimation (TMLE) is a double-robust efficient plug-in estimator designed to reduce bias in the SL-based estimate of the parameter of interest, often a much lower dimensional object (one-dimensional in our example) (van der Laan and Rubin, 2006). TMLE uses information in the data with respect to g_0 to update the initial SL-based estimate of Q_0.

2. MODELING THE RESPONSE SURFACE

DHSSC note that non-linearity of the response surface posed significant challenges to many estimators. Like DHSSC, we were not surprised to see that ensemble methods were most successful when faced with this challenge. SL exploits a diverse collection of algorithms that model information in the data in different ways. Of course, when the model space is large an exhaustive search over the solution space is not feasible. Domain expertise can help focus computational resources on likely areas within the model space. For example, insights into relationships among covariates and

Susan Gruber is Principal, Putnam Data Sciences, LLC, 85 Putnam Avenue, Cambridge, Massachusetts 02139, USA (e-mail: sgruber@putnamds.com). Mark J. van der Laan is Professor in Biostatistics & Statistics, University of California at Berkeley, Berkeley Way West, 2121 Berkeley Way, #5302, Berkeley, California 94720-7360, USA (e-mail: laan@berkeley.edu).

data quality can inform procedures for pre-processing covariates (e.g., creating plausible interaction terms) and covariate selection algorithms. Experts can also be asked to specify parametric models for inclusion in the SL library along with machine learning algorithms that more aggressively adapt to the data.

Domain expertise was not available in the Data Challenge, however all participants were able to examine sample covariate data. Given the mix of binary, count, and continuous covariates, we strove to create a general SL library that would be flexible enough to successfully model unknown covariate relationships with the binary treatment indicator and a continuous outcome. We defined covariate augmentation and screening procedures that narrowed the search at run-time to different relevant portions of the solution space. For example, instead of restricting lasso to main terms models, we augmented the dataset with squared and dichotomized versions of continuous covariates. In contrast, the multivariate adaptive regression spline algorithm was not supplied with squared terms. We used other combinations of dimension augmentation and reduction routines geared towards each prediction algorithm in the library. The highly adaptive lasso (HAL) algorithm is guaranteed asymptotically efficient (van der Laan, 2017), but was excluded from our competition entry due to its computational burden.

We had not anticipated BART's impressive performance. It's properties are consistent with targeted learning's emphasis on avoiding unwarranted assumptions. The additive sum of trees approach can closely approximate complex functional forms, and the regularizing Bayes prior minimizes overfits for robust finite sample performance. The post-competition results show that adding BART to the SL library improved performance of our SL+TMLE entry. We would not rely solely on BART, however, since BART might not be consistent for some DGPs, or its convergence rate might be slower than that of some other algorithm. In the post-competition evaluation DHSSC used TMLE to target an initial BART estimate. This provided an important layer of insurance grounded by theory, with very little down side: machine learning algorithms themselves are generally not asymptotically linear estimators for target estimands, while targeting them makes them asymptotically linear under the condition that the machine learning algorithm converges fast enough so that a second order remainder becomes negligible. DHSSC demonstrated that even when the initial BART estimate was already close to unbiased, targeting did not increase bias or compromise efficiency.

3. MODELING THE TREATMENT ASSIGNMENT MECHANISM

DHSSC use the term *alignment* to describe the correspondence between the treatment mechanism and the outcome. They point out that bias caused by failing to condition on a covariate present in the dataset depends on the strength of its relationships with treatment and outcome, and the functional forms of those relationships. High alignment indicates a large overlap in the two sets of predictive covariates. In other words, a covariate selection procedure that focuses on the outcome will produce a similar set of covariates as one that focuses on treatment. Estimators that appropriately condition on the selected set will be unbiased. DHSSC's findings highlight the difficulty of analyzing high dimensional datasets that contain unrelated covariates, or instrumental variables (IV) predictive of treatment only. When there is low alignment it can be hard to rule out superfluous covariates. As a result, estimators that unnecessarily condition on IVs when modeling the treatment mechanism will have inflated variance, and potentially inflated bias as well.

On a related note, DHSSC found that flexibly modeling the treatment mechanism did not markedly improve performance. We posit this is because most methods for doing so will not have the correct goal in mind. Our collaborative TMLE (C-TMLE) data-adaptively models g_0 with respect to a loss function for Q_0 (van der Laan and Gruber, 2010, Gruber and van der Laan, 2010). One version of C-TMLE provides a stepwise procedure for building a propensity score (or missingness) model that trades off bias reduction and variance inflation of the target parameter. It is particularly useful when there is sparsity in the data. For example, for the ATT parameter, when there are areas of the covariate distribution among the treated that have low density among the controls. For this competition, we did not use this stepwise version of C-TMLE because it can be very time consuming. However, in a next round, we would plan to incorporate more scalable versions of C-TMLE, such as tuning a lasso-parameter of the treatment mechanism with C-TMLE, or, more generally, using a scalable C-TMLE that first uses the data to provide an ordered sequence of g_0-estimators (Ju et al., 2017). We did, however, automate a covariate pre-screening procedure. In an attempt to exclude IVs from the response surface and treatment models, we examined the p-value on the coefficient in front of the single covariate in a model regressing Y on Z and each covariate in X in turn. Covariates with p-values > 0.5

were dropped from the dataset, with the caveat that a minimum of five covariates having the smallest p-values be retained.

4. TREATMENT EFFECT HETEROGENEITY

For our SL+TMLE submission, we choose an assumption-free way to account for treatment-covariate interactions by modeling the response surface separately for treated and untreated subjects. The SL+ TMLE JOINT estimation approach saves computation time and space by fitting only one instead of two models of the response surface. However, it places more burden on the data adaptive algorithms in the library to discover and correctly model important interactions. Although we were not involved in specifying the SL+TMLE JOINT estimator, when there are heterogeneous treatment effects its performance would likely improve by explicitly including pre-computed treatment-covariate interaction terms in an augmented dataset passed in to SL, or, by incorporating a machine learning algorithm that targets the conditional treatment effect function $E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X)$.

5. CONFIDENCE INTERVAL COVERAGE

The coverage of the SL+TMLE method typically fell between 0.85 and 0.95. This was superior to most methods, but failed to achieve the nominal rate. This coverage could be further improved by improving variance estimation (influence curve variance based estimators are often anti-conservative in the presence of positivity violations), using a bootstrap method that incorporates second order terms in the expansion of the TMLE to deal with lack of normality, using a more adaptive SL for Q_0 to remove remaining bias, for example, including the HAL algorithm (van der Laan, 2017; Benkeser and van der Laan, 2016), and using an adaptive C-TMLE algorithm. Although these improvements will cause tension with the required computing time in a competition setting, they are certainly worthwhile when analyzing data from an expensive, multiyear study.

6. CONCLUSION

DHSSC report that a method's performance on a given DGP is not easily predicted, even given oracle knowledge of characteristics of the DGP. In practice, that means we cannot know in advance what is the best method to analyze a given dataset. And if we choose only one approach, we cannot comfortably rely on the result. Therefore, one should aim for a method that is asymptotically grounded by theory, and finite sample robust across all allowed data generating distributions in the statistical model. This is precisely why targeted learning rests on a strong mathematical foundation, emphasizes minimal assumptions, and relies on super learning to exploit optimality properties of cross-validation for data adaptively estimating relevant nuisance parameters.

Theory provides an important guide for constructing estimators, but even when estimators possess the same asymptotic properties, finite sample performance can differ. Much of the work on TMLE has been motivated by the need to robustify the estimator in the face of unforeseen challenges in the data. Unexpected results send us back to the theory to understand why, and then incorporate that understanding into the estimation process: examples are the development of C-TMLE, HAL, CV-TMLE (Zheng and van der Laan, 2011), adaptive truncation using C-TMLE (Ju, Schwab and van der Laan, 2017), among others. DHSSC demonstrate that we no longer need to take comfort in the adage that *all models are wrong, but some are useful*. Their findings show that we obtain more reliable answers when we combine machine learning with domain knowledge using semiparametric and nonparametric methods. By thinking ahead and automating the steps in a targeted theoretically grounded data adaptive algorithm, we move statistical estimation further from craft towards science.

REFERENCES

Benkeser, D. and van der Laan, M. (2016). The highly adaptive lasso estimator. *Proc. Int. Conf. Data Sci. Adv. Anal.* **2016** 689–696.

Gruber, S. and van der Laan, M. J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int. J. Biostat.* **6** Art. 18, 31. MR2653847

Ju, C., Schwab, J. and van der Laan, M. J. (2017). On adaptive propensity score truncation in causal inference. Preprint. Available at arXiv:1707.05861.

Ju, C., Gruber, S., Lendle, S. D., Chambaz, A., Franklin, J. M., Wyss, R., Schneeweiss, S. and van der Laan, M. J. (2017). Scalable collaborative targeted learning for high-dimensional data. *Stat. Methods Med. Res.*

Polley, E. C. and van der Laan, M. J. (2010). Super learner in prediction. Technical Report 200, Univ. California Berkeley Division of Biostatistics, Working Paper Series.

van der Laan, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso. *Int. J. Biostat.* **13** 20150097, 35. MR3724476

VAN DER LAAN, M. J. and GRUBER, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *Int. J. Biostat.* **6** Art. 17, 70. MR2653848

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer, New York. MR2867111

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

ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning. Springer Ser. Statist.* 459–474. Springer, New York. MR2867139