Comment on "Bayesian Regression Tree Models for Causal Inference" by Hahn, Murray and Carvalho

Stefan Wager
Stanford University

August 11, 2020

The literature on heterogeneous treatment effect estimation has been extremely active over the past few years, and the paper by Hahn, Murray, and Carvalho [2020] is a major addition to it. Hahn et al. [2020] show how to design priors for heterogeneous treatment effects that are robust to what the authors call "regularization-induced confounding" and "targeted selection" and, as such, open the door to considerably more robust and reliable Bayesian inference of treatment heterogeneity under unconfoundedness. The authors convincingly show that their innovations add considerable value over a simpler Bayesian forest approach following, e.g., early work from Hill [2011].

The publication of this paper provides a nice opportunity to reflect on just how fast this area has developed over the past years. When Hahn et al. [2020] released the first draft of their manuscript on arXiv in 2017, there were still major questions about how best to approach the problem of heterogeneous treatment effect estimation as evidenced by, e.g., discussions at that year's Atlantic Causal Inference Conference. By now, in contrast, there appears to be fairly widespread consensus on conceptual ideas that underpin good estimators of treatment heterogeneity. This comment offers one take on 3 ideas which, I believe, have played a major role in pushing the field forward. Of course, these ideas manifest themselves differently depending on methodological context (e.g., frequentist vs. Bayesian methods, trees vs. lasso), but overall they seem to have broad applicability.

Dedicated regularization for treatment effects Following notation from Hahn et al. [2020], we want to estimate the effect of a binary treatment $Z_i \in \{0, 1\}$ on an outcome $Y_i \in \mathbb{R}$ as a function of covariates $X_i \in \mathcal{X}$. Following the Neyman–Rubin causal model [Imbens and Rubin, 2015], we posit potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(Z_i)$, and seek to estimate the conditional average treatment effect function $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$. For purposes of identification, we assume unconfoundedness [Rosenbaum and Rubin, 1983], i.e., that treatment is as good as random conditionally on X_i: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp Z_i \,\middle|\, X_i$.

One can readily check that, under unconfoundedness, we have $\tau(x) = \mu(x; 1) - \mu(x; 0)$ with $\mu(x; z) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x, Z_i = z\right]$. This suggests a simple strategy for non-parametric estimation of the conditional average treatment effect function $\tau(x)$: First learn $\hat{\mu}(x; 0)$ and $\hat{\mu}(x; 1)$ by fitting separate predictive models to the control and treated samples respectively, and then set $\hat{\tau}(x) = \hat{\mu}(x; 1) - \hat{\mu}(x; 0)$. This may, however, lead to problems. In general, modern machine learning methods operate via some type of representation learning; and, if

This work was supported by National Science Foundation grant DMS-1916163.

we use different representations to express $\hat{\mu}(x;\, 0)$ and $\hat{\mu}(x;\, 1)$, then $\hat{\tau}(x)$ may be excessively noisy or biased. As a simple example, consider the case where $\hat{\mu}(x;\, 0)$ and $\hat{\mu}(x;\, 1)$ are both decision trees—but with different splits. In this case, $\hat{\tau}(x) = \hat{\mu}(x;\, 1) - \hat{\mu}(x;\, 0)$ would be quite unstable, and in particular would have a more complicated shape than either $\hat{\mu}(x;\, 0)$ or $\hat{\mu}(x;\, 1)$ on its own. Künzel, Sekhon, Bickel, and Yu [2019] provide further examples of this issue.

A first major advance in the literature on treatment heterogeneity was the realization that, in a high-dimensional or non-parametric setting, it's important to use dedicated regularizers that directly push $\hat{\tau}(x)$ to have a simple form. A simple application of this idea arises in the case of the lasso [Hastie, Tibshirani, and Wainwright, 2015]. Assume that $\mu(x;\, z) = x \cdot \beta_{(z)}$ for some high-dimensional vector z, so that $\tau(x) = x \cdot (\beta_{(1)} - \beta_{(0)})$. A naïve analysis might fit separate lasso regressions on the treated and control units, but such an approach may learn different sparsity patterns for $\beta_{(0)}$ and $\beta_{(1)}$, thus resulting in an unstable $\tau(x)$ estimate. A better approach is to reparametrize $\mu(x;\, z) = x \cdot b + (2z - 1)\, x \cdot \delta$, where $b = (\beta_{(0)} + \beta_{(1)})/2$ and $\delta = \beta_{(1)} - \beta_{(0)}$; then, we can apply separate sparsity penalties on b and δ. This is a simple idea but, by directly pushing the treatment effect parameter δ towards sparsity, it often improves performance considerably.

Hahn et al. [2020] show how to adapt it to Bayesian prior design, while Athey and Imbens [2016] discuss a modification of regression trees that directly target $\tau(x)$. More subtle ideas for algorithmically regularizing the treatment effect function include the X-learner of Künzel, Sekhon, Bickel, and Yu [2019], and refitting predictions from a first step analysis as in the Virtual Twins method of Foster, Taylor, and Ruberg [2011].

The propensity score as a covariate Once we've dealt with egregious instability of $\hat{\tau}(x)$ by using appropriate regularizers, a next concern is whether confounding effects may bleed into treatment effect estimates due to finite sample effects. As is well explained in the section on "targeted selection" in Hahn et al. [2020], this concern arises whenever the propensity score $\pi(x) = \mathbb{P}\left[Z_i = 1 \,\middle|\, X_i = x\right]$ is associated with the baseline effect $\mu(x;\, 0)$. If $\mu(x;\, 0)$ takes on a complicated non-parametric specification that cannot be perfectly captured in finite samples and we use a method that underfits the baseline function such that $\mu(x;\, 0) - \hat{\mu}(x;\, 0)$ is positively (or negatively) correlated with $\pi(x)$, then we may easily have this baseline error push us to over- (or under-) estimate $\tau(x)$.

Considerations of this type have attracted considerable attention in the causal inference community for several decades [Robins and Ritov, 1997], and play a key role in discussions of how best to do variable selection when estimating global causal parameters [e.g., Belloni, Chernozhukov, and Hansen, 2014]. The general message is that, in order to be robust to associations between $\pi(x)$ and $\mu(x;\, 0)$, one needs to fit the propensity score $\hat{\pi}(x)$ and adjust for it in the final modeling step. Hahn et al. [2020] propose the simple idea of using a propensity estimate as a feature when estimating the baseline effect, i.e., they fit the baseline as $\hat{\mu}(x,\, \hat{\pi}(x);\, 0)$, and find this to work well empirically.

Treatment-focused loss functions A final idea that unlocks a general suite of tools for heterogeneous treatment effect estimation is the use of loss functions that directly isolate the treatment effect function $\tau(x)$. The idea of using target-specific loss functions has a long history in causal inference, going back at least to van der Laan and Dudoit [2003]. In the context heterogeneous treatment effect estimation, a simple instance of this idea is the "transformed outcome" method, which starts from the following observation. Under unconfoundedness, we can check that $\mathbb{E}\left[\Delta_i \,\middle|\, X_i = x\right] = \tau(x)$, where $\Delta_i = Z_i Y_i / \pi(X_i) -$

$(1 - Z_i)Y_i/(1 - \pi(X_i))$. Thus, in a randomized trial where propensity scores $\pi(x)$ are known a-priori, we can estimate $\tau(x)$ by first forming the modified outcomes Δ_i and then running a non-parametric regression of Δ_i on X_i [Tian, Alizadeh, Gentles, and Tibshirani, 2014]. Baseline effects $\mu(x; 0)$ never even appear in this specification, which largely obviates concerns related to regularizing or under-fitting of this term.

One limitation of the transformed outcome method is that it is not robust to errors in estimating the propensity score when $\pi(x)$ is not known a-priori, and several "robust" loss functions for treatment effect estimation that remedy this issue have recently been proposed. The R-learner [Nie and Wager, 2020] builds on the partially linear model estimator of Robinson [1988] to develop a loss function that is first-order robust to errors in estimating $\pi(x)$ and baseline effects; see also Athey, Tibshirani, and Wager [2019] and Zhao, Small, and Ertefaie [2017] for variants of this idea applied to random forests and the lasso specifically. Meanwhile, the DR-learner [Fan, Hsu, Lieli, and Zhang, 2019, Kennedy, 2020, Zimmert and Lechner, 2019] estimates $\tau(x)$ by regressing the augmented inverse-propensity weighted scores of Robins, Rotnitzky, and Zhao [1994] against X_i. Overall, the promise of such robust loss functions is that they enable accurate estimation of $\tau(x)$ even when when $\pi(x)$ and $\mu(x; 0)$ may be difficult to estimate, thus generalizing well known results on semiparametric inference for "global" targets like the average treatment effect; see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins [2018], and references therein.

Closing thoughts Recent advances in methods for treatment heterogeneity have enabled a large toolkit of practical methods available to applied researchers. As a community, we now appear to be at a point where we can quickly remix these ideas to develop new methods for treatment heterogeneity that can address different application-specific challenges, and have a formal understanding of how dedicated methods are able to accurately target $\tau(x)$. Several open questions remain, however. In particular, while reading the paper of Hahn, Murray, and Carvalho [2020], I was left wondering about the following two:

- When using treatment-focused loss functions, it's possible to show that (under appropriate conditions), the accuracy with which we can estimate $\tau(x)$ is insensitive to our rate of convergence on the nuisance components $\pi(x)$ and $\mu(x; z)$, even when $\hat{\pi}(x)$ and $\hat{\mu}(x; z)$ may converge an order of magnitude slower than $\hat{\tau}(x)$ [Kennedy, 2020, Nie and Wager, 2020]. Do analogous results hold for the approach of Hahn et al. [2020], where $\hat{\pi}(x)$ is used as a covariate when fitting $\hat{\mu}(\cdot)$? In the context of estimating an average treatment effect, Hirano, Imbens, and Ridder [2003] showed that using an estimated propensity score for weighting could sometimes be enough to achieve efficiency. Does any intuition of this type carry over to the problem heterogeneous treatment effect estimation?

- In Hahn et al. [2020], the propensity score is only used as a covariate to make us robust to potential confounding effects. In some applications, however, there may also be interest in the propensity score as an effect modifier. For example, when studying the returns to college education, Brand and Xie [2010] argue that students who are least likely to attend college a-priori may be the ones who benefit the most from attendance; and, in applications like these, it may be of interest to allow $\tau(x)$ explicitly depend on $\pi(x)$. Of particular interest here would be to understand how $\tau(x)$ varies with the true propensity score $\pi(x)$, and not just the estimate $\hat{\pi}(x)$.

Finally, I want to thank Hahn, Murray, and Carvalho [2020] for preparing this very nice paper, and look forward to seeing how the community builds on their results in the future.

References

Susan Athey and Guido W Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

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

Jennie E Brand and Yu Xie. Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review*, 75 (2):273–302, 2010.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68, 2018.

Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *arXiv preprint arXiv:1908.02399*, 2019.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, forthcoming, 2020.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press, 2015.

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

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, forthcoming, 2020.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Mark J van der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, UC Berkeley Division of Biostatistics, Berkeley CA, 2003.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

Michael Zimmert and Michael Lechner. Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*, 2019.