On two approaches to weighting in causal inference

Commentary on "The right tool for the job: choosing between covariate balancing and generalized boosted model propensity scores," manuscript EDE16-0210R2, Epidemiology

David A. Hirshberg and José R. Zubizarreta

There are two general approaches to weighting in causal inference. One of them centers on modeling the data — on accurately and flexibly modeling the probabilities of receiving treatment. The other focuses on diagnostics — on directly minimizing and/or constraining covariate imbalances and the dispersion of the weights. In this issue of the Journal, Setodji et al.[1] offer an interesting comparison of two weighting methods for estimating average treatment effects with time-invariant binary treatments: the use of generalized boosted models to estimate the probability of receiving treatment,[2] and the use of the covariate balancing propensity score[3] to promote covariate balance between the treatment group and the weighted control group.

Claiming at the outset that neither the covariate balancing propensity score nor generalized boosted models should be preferred in all circumstances, Setodji et al. aim to provide guidance on how to choose between these two methods. In a sense, each stands as a proxy for one of the more general approaches described above: the described motivations for using the covariate balancing propensity score apply just as well to the balancing methods of Hainmueller[4] or Chan et al.,[5] and the motivations for using generalized boosted models apply as well to the use of any flexible nonparametric regression technique to estimate the propensity score.

To get to the heart of the issue, we look at the error in estimating the average treatment effect on the treated, θ, when treatment assignment is unconfounded. This estimand is used frequently in settings with time invariant binary treatments in which we expect treatment to be made available to a new population similar to the one studied. Our remarks largely apply to average treatment effects over other subsets of the population, including the whole population, with slight changes in the details.

Let Y represent the outcome, T the treatment assignment, and X the observed covariates in an iid sample with n_1 treated units and n_0 controls. We write $Y_i = \mu_{T_i}(X_i) + \epsilon_i$ where $\mu_{T_i}(X_i) := \mathbb{E}(Y_i \mid X_i, T_i)$ and $\epsilon_i := Y_i - \mu_{T_i}(X_i)$. Consider the weighting estimator $\hat{\theta} := \frac{1}{n_1}\sum_{i:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} \gamma_i Y_i$ with weights γ. The estimators considered by Setodji et al. are instances of this one using estimated inverse

probability weights. Its error in estimating θ can be decomposed as follows:

$$\hat{\theta} - \theta = \left(\frac{1}{n_1} \sum_{i:T_i=1} \mu_0(X_i) - \frac{1}{n_0} \sum_{i:T_i=0} \gamma_i \mu_0(X_i) \right) + \left(\frac{1}{n_1} \sum_{i:T_i=1} \epsilon_i - \frac{1}{n_0} \sum_{i:T_i=0} \gamma_i \epsilon_i \right).$$

We focus our attention on the first term in parenthesis — the imbalance in the function μ_0, which we'll write $\mathbb{I}_{\mu_0}(\gamma)$. This is what we can control and what we need to control for accurate estimation. As long as the weights are bounded and don't depend on the outcomes $\{Y_i\}$, the second term converges to zero at $\frac{1}{\sqrt{n}}$ rate. Therefore our primary goal, and the motivation for essentially all weighting methods, is to ensure that the imbalance is small without inflating the variance of the second term. See the appendix for details.

If we know the treatment assignment probabilities, inverse probability weighting (IPW) does this fairly well: the imbalance $\mathbb{I}_{\mu_0}(\gamma)$ decays like $\frac{1}{\sqrt{n}}$ under essentially no assumptions on μ_0. This suggests that IPW with estimates of the treatment assignment probability, like the generalized boosted models method discussed by the authors, will perform well. However, there are two major criticisms of this approach: (1) we may not be able to estimate the treatment assignment probability well enough to replicate or improve upon the good behavior of the true inverse probability weights; (2) $\frac{1}{\sqrt{n}}$ or slightly faster decay of the imbalance isn't optimal — it is possible to shrink the imbalance much faster. Methods that choose weights to explicitly minimize or constrain the imbalance, like the covariate balancing propensity score or entropy balancing, aim to address both.

We will focus on the second criticism. The ability to shrink the imbalance much faster than $\frac{1}{\sqrt{n}}$ in some circumstances is a strong reason to prefer these methods. This motivates us to try to detect these favorable circumstances. In order to balance the unknown function μ_0, balancing methods choose weights that minimize or bound the level of imbalance uniformly over a class of functions $\mathcal{F}$, i.e. weights γ that control $\mathbb{I}_{\mathcal{F}}(\gamma) := \max_{f \in \mathcal{F}} \mathbb{I}_f(\gamma)$. These methods are clearly preferable to model-based IPW approaches when we are confident that μ_0 is in or near a set $\mathcal{F}$ for which we can guarantee that $\mathbb{I}_{\mathcal{F}}(\gamma)$ decays fast. However, these classes $\mathcal{F}$ are small and therefore not likely to contain the true μ_0 in practice.

Often the weights are chosen so that $\mathbb{I}_{\mathcal{F}}(\gamma) = 0$.[4;5] We call these exact balancing methods. In order for exact balancing methods to work at all, it is necessary that $\mathcal{F}$ has dimension no larger than n_0. Perhaps more importantly, unless the dimension of $\mathcal{F}$ is much smaller than n_0, it is likely that we will need very large weights to balance $\mathcal{F}$ uniformly; these methods force us to make a tradeoff between misspecification of $\mathcal{F}$ and large weights that inflate the variance dramatically. In practice, these methods

often suffer from both problems at once: we see both excessive variance due to large weights and bias because balancing $\mathcal{F}$ does little to balance μ_0.

The form of the covariate balancing propensity score used in the article by Setodji et al., which balances linear functions of the covariates, is essentially of this type. By incorporating the score of a parametric model of the probability of treatment into the estimating equations, it gains robustness in the event that the likelihood is not badly misspecified. This property is in fact shared with many exact balancing methods, which fit an implicit treatment probability model.[5;6] In the simulation study of Setodji et al., when both the treatment probability model and the class of functions balanced are misspecified, the covariate balancing propensity score performs substantially worse than the nonparametric treatment probability estimator based on generalized boosted models. On the other hand, the covariate balancing propensity score is shown to have only a slight edge even when the treatment assignment model is correctly specified and the class of functions it balances contains μ_0. We believe this behavior to be characteristic of settings in which the propensity score is fairly easy to estimate. In settings in which it is more difficult, such as high dimensional settings, methods like the covariate balancing propensity score have more to offer as they will at least balance a projection of μ_0. And because in these settings bias tends to dominate, the high variance we expect from exact balancing methods does not have much impact.

Balance checks are a useful tool for assessing our risk of a large imbalance $\mathbb{I}_{\mu_0}(\gamma)$, whether they arise from a misspecified treatment probability model and/or explicitly balanced class $\mathcal{F}$, or from slow convergence of a treatment probability estimate. However, most commonly used balance checks measure the maximal bias $\mathbb{I}_{\mathcal{F}}(\gamma)$ over small classes of functions. Setodji et al. check balance using the two popular diagnostic statistics, the average standardized absolute mean difference and the maximal marginal Kolmogorov-Smirnov statistic, but these statistics do not sharply signal the inadequacy of balancing approaches when they are misspecified. This should be no surprise: the average standardized absolute mean difference is $\mathbb{I}_{\mathcal{F}}(\gamma)$ for a class $\mathcal{F}$ of linear models and the maximal marginal Kolmogorov-Smirnov statistic is $\mathbb{I}_{\mathcal{F}}(\gamma)$ for a class of additive models. Neither usefully characterizes bias when there are interactions between the covariates. We recommend checking balance by evaluating $\mathbb{I}_{\mathcal{F}}(\gamma)$ for much richer function classes $\mathcal{F}$, like Hölder or Sobolev classes of smooth functions.

Setodji et al. propose another indirect balance check for use in deciding between their fast-balancing method (covariate balancing propensity score) and their nonparametric treatment probability modeling approach (generalized boosted models). It checks the adequacy of balancing the class $\mathcal{F}$ of linear functions by using a linear

model to estimate μ_0 and running a goodness-of-fit test. We do not believe this approach to be ideal for several reasons: (1) there are inferential complications because the outcome data is used to choose the method; (2) it may choose a badly misspecified fast-balancing method when the goodness of fit test lacks power; (3) a test with sufficient power will reject the fast-balancing approach in the almost-certain circumstance that $\mathcal{F}$ is misspecified. A more direct way to benefit from the possible simplicity (e.g. linearity) of μ_0 is to choose between different models for the outcome based on how well they predict the outcome, for example using cross-validation. The chosen outcome model can then be incorporated into a 'double-robust' estimator combining weighting with imputation of missing outcomes.[7;8] This approach, advocated by Van der Laan and Rose[9] and Chernozhukov et al.,[10] does not have the inferential complications of approach proposed in Setodji et al..

In keeping with our recommendations for balance checks, we advocate weighting methods that uniformly control the imbalance over large classes of functions $\mathcal{F}$. This requires that we ask for approximate balance instead of exact balance, either by imposing a sufficienly loose constraint on the maximal imbalance $\mathbb{I}_{\mathcal{F}}(\gamma)$ or by minimizing $\mathbb{I}_{\mathcal{F}}(\gamma)^2 + \psi(\gamma)$, where $\psi(\gamma)$ is a proxy for the variance.[11;12;13;14;15] The behavior of these methods is in some ways more similar to methods that weight by the inverse of estimated treatment probabilities than to their exact balancing kin: over large function classes we cannot obtain uniform balance at a fast rate; we instead see decay of the imbalance like or slightly faster than $\frac{1}{\sqrt{n}}$. Furthermore, these weights converge to the true inverse propensity weights under extremely weak assumptions.[13;14]

With an appropriately chosen regularization term $\psi(\gamma)$, the loss $\mathbb{I}_{\mathcal{F}}(\gamma)^2 + \psi(\gamma)$ is a tight bound on the mean squared error for μ_0 known to be in $\mathcal{F}$ when the covariates and treatment assignments are considered fixed. In that event minimizing it gives, in a sense, the best weights for whatever treatment assignment and covariates you observe. However, correct choice of ψ is a very difficult tuning problem; it requires that we know both the scale of μ_0 and the variance $\mathrm{var}(Y_i \mid X_i, T_i)$ for each control observation. To our knowledge, no procedure that estimates these tuning parameters has been proposed.

On the other hand, IPW methods arrive at control of bias and variance less directly, through the asymptotic balancing behavior of inverse probability weighted averages and the asymptotic optimality of the inverse probability weighted noise terms ϵ_i. This indirectness is one reason we believe that approximate balancing methods are a more promising approach. However, it is straightforward to tune an estimator of the treatment probability by cross-validation. While this is useful, it isn't an optimal way of tuning for estimation of treatment effects; the tuning is for accuracy of

the estimated treatment probabilities, not for performance of their inverses as balancing weights. One concrete problem is that these estimated probabilities must be inverted to get the weights, amplifying estimation error badly when the estimated probability is small. The common approach to dealing with this is ad-hoc trimming of the weights; the impact of this choice can have as strong an impact on the behavior of the method as the choice of the method itself. This important practical decision is not studied by Setodji et al.. Hirshberg and Wager[13] show that their approximate balancing weights minimize an unbiased estimate of the mean squared error of the weights as an estimator of the inverse probability weights, plus a regularization term, avoiding this error-inflating inversion step. Despite their apparent disadvantages, well-tuned nonparametric IPW methods are often competitive with approximate balancing methods in settings in which the propensity score is not too hard to estimate.

So far, we've been discussing which weights to use in a weighting-only treatment effect estimator. This, in our opinion, is a second-order consideration. The first choice we need to make is between methods that weight the observed outcomes like those we've been discussing, regression methods that impute the counterfactual outcomes and take an unweighted average over the imputed complete data, and 'double robust' combinations of these two approaches. Double robust approaches for causal inference were proposed by Robins and Rotnitzky,[7] and the previously mentioned approaches of Van der Laan and Rose[9] and Chernozhukov et al.[10] are in this last class. Related approaches that use forms of matching instead of weighting are proposed by Rubin,[16] Rosenbaum,[17] and Abadie and Imbens.[18] We believe that imputation of missing outcomes by regression is an essential tool for treatment effect estimation, and for that reason we highly recommend the use of a double robust approach combining regression with stable weighting. This is corroborated by some of our largest simulation studies: the leading methods in the 2016 Causal Inference Data Analysis Challenge at the Atlantic Causal Inference Conference either combined weighting with outcome regression or used outcome regression alone. Our previous discussion of weighting methods still applies; for the double robust augmented inverse probability weighting estimator, the previous decomposition of the estimation error $\hat{\theta} - \theta$ changes by substitution of the regression error $\hat{\mu}_0 - \mu_0$ for μ_0. In other words, incorporating a regression leaves us needing to balance another unknown function, the regression error; the advantage is that it tends to be a smaller unknown function. In the Figure, we see evidence of this advantage from a small simulation.

We thank Setodji et al. for their article and the editors for the opportunity to put together and share these thoughts. This is an important area of research where much remains to be done, and to which the article of Setodji et al. offers evidence

and insight.

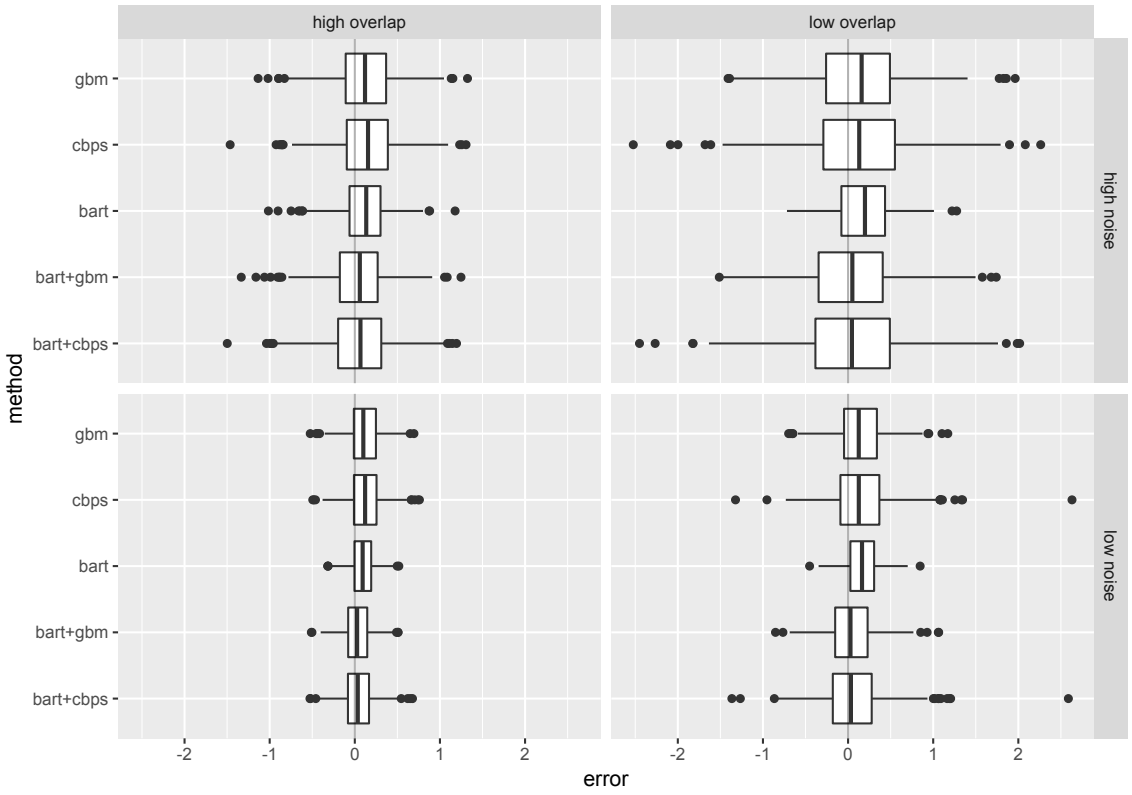

Figure 1: A comparison of the error distributions using weighting with generalized boosted models and the covariate balancing propensity score, regression imputation with Bayesian Adaptive Regression Trees (BART), and their combination as an AIPW estimator. Following convention, the split boxes indicate the second and third quartiles of the data, with the whiskers extending away from the median a distance of 1.58 times the interquartile range and samples beyond the whiskers plotted individually.

In our simulated data, we used a 15-dimensional vector of covariates X distributed as in Setodji et al. and like Setodji et al. used the constant treatment effect -0.4. We departed from Setodji et al. in our choice of a less linear form for μ_0 and $\operatorname{logit} p$. In each simulated dataset, μ_0 was a random convex quadratic function of the form $f(x) = x^T Q x + b^T x + c$ operating on the first $k = 10$ covariates, where $Q = A^T A / \left(\sqrt{k}\lambda_{\max}(A^T A)\right)$ for a $k \times k$ matrix A with iid standard normal entries and b is uniformly distributed on the k-dimensional euclidean unit sphere. $\operatorname{logit} p$ was an independent random function of the same form acting on the first $k = 7$ components of X, with the constant c chosen to vary the amount of overlap of the treatment and control groups. The use of functions of the first 10 and 7 covariates for μ_0 and $\operatorname{logit} p$ respectively were consistent with the simulation scenarios of Setodji et al., with the last 5 covariates left as 'distractors.'

References

[1] C. M. Setodji, D. F. McCaffrey, L. F. Burgette, D. Almirall, and B. A. Griffin. The right tool for the job: choosing between covariate balancing and generalized boosted model propensity scores. *Epidemiology*, page to appear, 2017.

[2] D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[3] K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[4] J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[5] K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society*, 78(3):673–700, 2016.

[6] Q. Zhao and D. Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

[7] J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[8] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[9] M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[10] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

[11] S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

[12] N. Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

[13] D. A. Hirshberg and S. Wager. Balancing out the unobservable: Weighting for uniform balance to attenuate regression errors in estimation of average treatment effects. *Working Paper*, 2017.

[14] Y. Wang and J. R. Zubizarreta. Approximate balancing weights: Characterizations from a shrinkage estimation perspective. *arXiv preprint arXiv:1705.00998*, 2017.

[15] J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

[16] D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.

[17] Paul R Rosenbaum et al. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

[18] Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.