Appendix A: Population Inference

BART for population inference.

This section proposes two approaches to inference for population causal estimands. The first approach stays within the Bayesian framework but uses additional information to modify the actual estimate. The estimation procedure for the second approach is the same as for CATE, but an approximation is developed for the variance of this quantity when viewed as an estimator of the population effect.

While it may be very difficult to obtain a large sample for which the response Y is available, it is often relatively easily to obtain values of the conditioning variables in X for a great many individuals. The first proposal (henceforth BARTpx) is to obtain, or construct, a set $\{\tilde{x}_i\}_1^K$ that represents the population of interest. This set may be much larger than the data used in the BART inference for f. Recall that $c(x, f) \equiv f(1, x) - f(0, x)$ is defined as the treatment effect at $X = x$. Then let $\bar{C}^l = \frac{1}{K} \sum_i^K c(x_i, f^l)$ as in section 3 of the article. The values $\bar{C}^l$ now represent draws from the average population treatment effect, PATE.

There may be far greater interest in how the treatment will impact a different group of individuals (or more generally treatment units) than those in the sample and this approach can directly address such concerns. If predictor information is available for this group, we can use BART to illuminate this question by making predictions across the $\tilde{x}_i$ to form the posterior distribution of the average treatment effect. However, if the set $\{\tilde{x}_i\}_1^K$ is quite different from those in the sample then we are asking the inference for f to extrapolate to unseen regions of the x space. This poses a very difficult problem and confidence in this (or any) procedure is diminished in a way that is difficult to quantify.

The second approach (henceforth BARTsx) relies only on our original sample and estimates the population treatment effect estimands using the same estimates as those described for the conditional average treatment effects, namely $\bar{C} = \frac{1}{L}\sum_{l=1}^{L}\bar{C}^{l}$. What changes is our measure of uncertainty. The proposed estimate combines the posterior variation of the draws f^l with the sample variation of the x_i to get an overall measure of uncertainty relevant to PATE. Let $C_{li} = c(x_i, f^l)$. The C_{li} are conceptualized as draws of $C(X, F)$ where X is drawn randomly from the population of interest and F is drawn independently from the posterior[1]. The obvious estimate of variability is then the usual quantity

$$V_C = \frac{1}{nL}\sum_{l=1}^{L}\sum_{i=1}^{n}(C_{li} - \bar{C})^2$$

where L is the number of kept MCMC draws and $\bar{C}$ is the CATE/PATE estimate, the average of all the C_{li}. This variance estimator represents a sort of "hybrid" between Bayesian ideas and frequentist ideas in that the nature of "random" is different for f^l than it is for the x_i.

Using the standard decomposition $\sum_{i=1}^{n}(C_{li} - \bar{C})^2 = \sum_{i=1}^{n}(C_{li} - \bar{C}^{\,l})^2 + n(\bar{C}^{\,l} - \bar{C})^2$ we have

$$V_C = \frac{1}{nL}\sum_{l=1}^{L}\sum_{i=1}^{n}(C_{li} - \bar{C}^{\,l})^2 + \frac{1}{L}\sum_{l=1}^{L}(\bar{C}^{\,l} - \bar{C})^2.$$

Only the second term is required for inference for CATE. Likewise if using the first approach to population inference then again the second term would fully capture the uncertainty. If the interest is in PATE, however (and we don't have access to information on X for the full population) then the first term captures the additional variation in the treatment effect that arises due to the choice of sample. However the additional assumption that the estimate is at least approximately unbiased for PATE, the square root of V_C can be used as a standard error. Given the strong BART prior, it is unlikely that $\bar{C}$ is completely unbiased for PATE for X drawn from the population and F from the posterior no matter what the true function f may be. For instance if there is an overlap problem this assumption of approximate unbiasedness may be tenuous; however with an overlap problem any

[1] Although the distribution of the f^l draws is likely not independent of the drawn x_i, this dependence may not be strong and we assume it is negligible for our purposes (although the f^l are very dependent on the drawn (x_i, z_i, y_i), the marginal dependence between the drawn f^l and x_i may be less strong).

method is likely to run into trouble[2]. Moreover, the following auxiliary simulations demonstrate that V_C can perform well competitively in practice.

Simulations

This section investigates the properties of the two proposed population estimators with respect to a population estimand, in this case the PATT. The original simulations from Section 4 have been reworked to accommodate this goal. First a large population of observations was generated with a similar joint distribution for the covariates and treatment indicator as in the original IHDP experimental data. To achieve this a general location model is fit to these data (treatment variable and covariates) and this model is used to generate a population of size 60,000 (40,000 treated and 20,000 controls; this mimics the ratio of controls to treated in the original experiment). Just as in the original simulations all children with non-white mothers are removed. Outcome variables for this population were generated using the same specifications for response surfaces A, B, and C (description of simulations for response surface C are included in online appendix B) as in the original simulations, except now for B and C the average treatment effect for the full population of treated (PATT) is coerced to be 4 in each iteration.

For each iteration of the simulations a random sample of 139 treated and 608 controls (the same as in the sample used in our original simulations) was drawn from the full population. Then the two "new" BART methods and three original competitors were fit to the sample data as in the original simulations. The results from these population simulations are displayed in Table A.1. Methods were compared with respect to root mean squared error ("RMSE"), average interval length ("length"), and coverage. The first BART population estimation strategy that uses only the sample data but creates an augmented variance estimator is denoted BARTsx in the table. The second BART population

[2] See related work on BART strategies to identify neighborhoods in the covariate space where there is no overlap (Hill and Su 2010).

Response	Procedures				
A	BARTsx	BARTpx	OLS	Pscore	IPW
RMSE	0.11	0.11	0.11	0.17	0.11
length	0.44	0.43	0.40	0.56	0.39
coverage	95	95	94	91	92
B					
RMSE	0.53	0.25	0.49	0.72	0.55
length	1.75	0.63	1.19	1.75	1.27
coverage	95	88	86	88	84
C					
RMSE	0.35	0.22	0.34	0.46	0.36
length	1.29	0.65	1.00	1.40	1.06
coverage	95	86	87	87	86

Table A.1: Results from population inference simulations.

estimation strategy that makes predictions across the entire population is denoted BARTpx in the table.

As expected, the results for response surface A are almost identical to the corresponding results for the sample because the treatment effect is constant. Response surfaces B and C however exhibit strongly heterogeneous response surfaces. Accordingly, the results for these response surfaces depart noticeably from the results for the original simulations.

For response surface B the root mean squared error is quite similar across BARTsx, linear regression and the weighted estimator. All three of these beat the matching estimator. Perhaps not surprisingly, BARTpx, which makes use of covariate information for the full population, performs markedly better than all of the others with regard to root mean squared error. However, only BARTsx (which has noticeably longer intervals on average compared to linear regression and the weighting estimator) achieves nominal coverage. The difficulty that the other methods are facing is that in order for their associated variance estimators the model specification must be correct, however, this is not that case in these settings.

The results for response surface C largely mimic those for response surface B although the differences in RMSE and interval length across methods are less extreme.

References

Hill, J. L. and Su, Y.-S. (2010), "Assessing common support for causal inference in high-dimensional covariate space," Tech. rep., New York University.