Recursive Partitioning for Heterogeneous Causal Effects

Susan Athey & Guido Imbens

Supporting Information - Appendix

This Appendix describes some additional details of the simulation study, and also presents additional simulation results in Table S1.

The code for our simulations was written as an R software package that is in preparation for public release. It is based on the 'rpart' R package. This package is available at https://cran.r-project.org/web/packages/rpart/index.html. For TOT, we directly use rpart applied to the transformed outcome Y_i^*, and we use 10-fold cross-validation for pruning the tree. For each of our other estimators, we modified several components of the package. In the remainder of this Appendix, discussions of modifications apply to F, CT, and TS estimators. For these estimators, we create new versions of the "anova" functions, functions that in the standard rpart package are used for calculating the total goodness of fit for a node of the tree, evaluating the quality of alternative splits, and estimating the goodness of fit for pruned trees using cross-validation samples. We maintain the overall structure of the rpart package. The rpart package has an important tuning parameter, which is the minimum number of observations per leaf, denoted n_m. We modify the rpart routine to insist on at least n_m treated *and* n_m control observations per leaf, to ensure that we can calculate a treatment effect within each leaf. In the simulations reported in Table 1 of the paper and in Table S1 of this Appendix, we use $n_m = 25$ for all models except TOT, while for the TOT model the minimum leaf size is 50 (without restrictions on treated and control observations).

We make one additional modification to the way the standard rpart splitting function works. We restrict the set of potential split points considered, and further, in the splitting process we rescale the covariate values within each leaf and each treatment group in order to ensure that when moving from one potential split point to the next, we move the same number of treatment and control observations from the right leaf to the left leaf. We begin by describing the motivation for these modifications, and then we give details.

The rpart algorithm considers every value of $X_{i,k}$ in $\mathcal{S}^{\text{tr}}$ as a possible split point for covariate X_k. An obvious disadvantage of this approach is that computation time can grow prohibitively large in models with many observations and covariates. But there are some more subtle disadvantages as well. The first is that there will naturally be sampling variation in estimates of $\hat{\tau}$ as we vary the possible split points. A problem akin to a multiple hypothesis testing problem arises: since we are looking for the maximum value of an estimated criterion across a large number of possible split points, as the number of split points tested grows, it becomes more and more likely that one of the splits for a given covariate appears to improve the fit criterion even if the true value of the criterion would indicate that it is better not to split. One way to mitigate both the computation

time problem and the multiple-testing problem is to consider only a limited number of split points.

A third problem is specific to considering treatment effect heterogeneity. To see the problem, suppose that a covariate strongly affects the mean of outcomes, but not treatment effects. Within a leaf, some observations are treated and some are control. If we consider every level of the covariate in the leaf as a possible split point, then shifting from one split point to the next shifts a single observation from the right leaf to the left leaf. This observation is in the treatment or the control group, but not both; suppose it is in the treatment group. If the covariate has a strong effect on the level of outcomes, the observation that is shifted will be likely have an outcome more extreme than average. It will change the sample average of the treatment group, but not the control group, leading to a large change in the estimated treatment effect difference. We expect the estimated difference in treatment effects across the left and right leaves to fluctuate greatly with the split point in this scenario. This variability around the true mean difference in treatment effects occurs more often when covariates affect mean outcomes, and thus it can lead the estimators to split too much on such covariates, and also to find spurious opportunities to split.

To address this problem, we propose the following modifications to the splitting rule. We include a parameter b, the target number of observations per "bucket." For each leaf, before testing possible splits for a particular covariate, we order the observations by the covariate value in the treatment and control group separately. Within each group, we place the observations into buckets with b observations per bucket. If this results in less than n_m buckets, then we use fewer observations per bucket (to attain n_m buckets). We number the buckets, and considering splitting by bucket number rather than the raw values of the covariates. This guarantees that when we shift from one split point to the next, we add both treatment and control observations, leading to a smoother estimate of the goodness of fit function as a function of the split point. After the best bucket number to split on is selected, we translate that into a split point by averaging the maximum covariate value in the corresponding treatment and control buckets. In the simulations presented in this paper, we do not constrain the maximum number of buckets, and we let $b = 4$. We found that this discretization approach improved goodness of fit on average for the simulations we considered, although it can in principle make things worse.

Another tuning parameter for standard CART as well as the methods proposed here is the number of cross-validation samples. A common convention is to use 10 samples. We deviate from that convention and use 5 cross-validation samples. The reason is that our methods require various quantities to be estimated within leaves. Given a minimum leaf restriction of 25 treated and control units, if we take a cross-validation sample of one-tenth of the original training sample, we might end up with no treated or no control observations in a leaf in a cross-validation sample. In addition, it may be difficult to estimate a sample variance within a leaf. Rather than require larger leaf sizes,

we simply use fewer cross-validation samples.

In the simulations reported in Table 1 of ths paper, we used the infeasible MSE_τ to evalute alternative estimators. In practice, we must estimate the infeasible criterion. In the paper, we propose estimators that rely on the tree structure of our estimator, but we may also wish to compare our performance to estimators that don't rely on partitions. One alternative is the MSE^{TOT} criterion. Given an estimator $\hat{\tau}_i$, it is equal to

$$MSE^{\text{TOT}} = \frac{1}{N^{\text{te}}} \sum_{i=1}^{N^{\text{te}}} (Y_i^* - \hat{\tau}_i)^2.$$

Because $\mathbb{E}[Y_i^*|X_i] = \tau(X_i)$, this is an unbiased (but noisy) estimator for MSE_τ. In Table S1, we present rankings of estimators using this criterion. We see that it ranks estimators in the same way as MSE_τ, but the percentage differences between estimators are smaller than with the infeasible criterion.

Table S1 also includes the full set of estimates for the infeasible criterion MSE_τ, to illustrate how sample size and honest versus adaptive estimation affects the criterion.

Table S1: Infeasible and Feasible MSE Estimates for Simulation Study

Design	1		2		3	
$N^{\text{tr}} = N^{\text{est}}$	500	1000	500	1000	500	1000
Estimator	MSE_τ^{TOT} Divided by MSE_τ^{TOT} for CT-H*					
TOT-H	1.009	1.009	1.006	1.004	1.012	1.004
F-H	1.014	1.005	1.067	1.086	1.073	1.117
TS-H	0.9996	0.9997	1.013	1.008	1.026	1.036
	Infeasible MSE_τ					
TOT-A	0.134	0.104	1.452	1.021	3.701	2.518
F-A	0.152	0.076	2.644	2.588	5.140	5.038
TS-A	0.092	0.066	1.817	1.264	4.357	3.552
CT-A	0.090	0.068	1.518	1.082	3.844	2.610
TOT-H	0.134	0.104	1.452	1.021	3.701	2.518
F-H	0.155	0.077	2.645	2.587	5.141	5.038
TS-H	0.084	0.052	1.578	1.094	4.034	3.233
CT-H	0.086	0.054	1.334	0.955	3.423	2.416

*$\text{MSE}_\tau^{\text{TOT}}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator}}(\mathcal{S}^{\text{tr}}))/\text{MSE}_\tau^{\text{TOT}}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{CT-H}}(\mathcal{S}^{\text{tr}}))$