Potential outcome framework and assumptions

- Observational studies, not experiment
- Fundamental problem of causal inference, we never observe $\tau_i(X_i) = Y_i(1) - Y_i(0)$ directly => introduce identification strategies under certain assumptions
- ATE: is $\beta_0 = E[Y_i(1) - Y_i(0)]$ in $Y_i = \alpha_i + \beta_0 D_i + X_i\delta + \epsilon_i$
- CATE: condition on x's to capture heterogeneity:
  $\tau_0(x) = E[Y_i(1) - Y_i(0)|X_i = x] = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$
- propensity score: probability of being treated

Assumptions

- unconfoundedness => potential outcome is independent of treatment assignment
  - we need this because in observational studies, assignment is not determined ex-ante (in experiment fashion)
- common support / full overlap => propensity score is bounded away from 0 and 1 => guaranteeing existence of hypothetical counterfactual for each treatment assigment

Notation

- $g_0(D_i, X_i) = E[Y_i|D_i = d, X_i = x]$:= expected outcome given treatment & control variables
- $l_0(x) = E[Y_i|X_i = x]$:= conditional mean regression marginalized over treatment?
- propensity score: $m_0(x) = E[D_i|X_i = x] = Pr(D_i = 1|X_i = x)$

Double/Debiased ML

- The Double/Debiased Machine Learning (DML) framework provides valid point estimates and confidence intervals of a low dimensional parameter in the presence of high dimensional nuisance parameters
- Author considers Lasso, Regression Trees, RFs, Boosting & Neural Nets, as well as BART
- Main contribution of DML: accurate point estimates + valid inference (CI + p values)
- Final DML estimates are the median value of the point estimate, median SE, and modified SE incorporating uncertainty of sample splitting
- Knaus et al use DR estimator to derive point estimates for CART up until individual level

Generalized Random Forest

- estimate any quantity of interest by local moment conditions accounting for presence of nuisance parameters + providing valid statistical inference
- Causal Forest is a specific implementation of GRF using tree estimates on random subsets of data
  - each tree identifies Heterogeneous Treatment Effects (HTE)

    - CF is therefore used to estimate CATE
- Asymptotic analysis of GRF:
    - the pseudo-forest estimates are asymptotically normal
    - under weak assumptions, there is a sequence $\sigma_n(x)$ for which $\frac{\hat{\beta}_n - \beta(x)}{\sigma_n(x)} \sim N(0,1)$
- $\sigma_n(x)$ can be used to construct asymptotically valid Gaussian CI for $\beta(x)$
    - but estimator for $\hat{\sigma_n(x)^2}$ is derived via ridiculously hard Delta method (eq 27)
    - An alternative is to use Bootstrap of Little Bags ("blb" in code), which I recognize from the DML causal forest model
    - We should check whether GRF.CausalForest() does this (I suspect it uses bootstrap of little bags, check for model param: inference='blb')

Theory on Causal Forest and HTE estimation in section 3.4.3 / page 18

- We can straight up copy/use this for our paper.

Parameters config:

- 8000 trees and l = 2 (set of half-samples to compute CIs)
- use weighted ATE estimator that incorporates the probability of being treated:
    - $\hat{\beta_{ATE}} = \Sigma_i^n m(X_i)(1 - m(X_i)E[Y(1) - Y(0)|X_i = x]/\Sigma_i^n m(X_i)(1 - m(X_i)$ where $m(x) = Pr(D_i = 1|X_i = x)$
    - this is used in cases of poor overlap (so v skewed treatment distribution)

BART

- Section 3.5
- Prior that you choose is
    - the tree structure
    - values in terminal nodes
    - noise of residuals standard deviation?
- Chipman recommend number of trees to 200

Posterior inference:

- Backfitting generates a sequence of draws of the sum tree functions, which should converge to the true posterior distribution of $f(\cdot)$
- Ensure convergence: 200 burn-ins (?) & 1000 steps per chain
- Given K samples post-burn-in, you estimate $f(x)$ by computing the posterior mean estimate, i.e. the average of the sum-of-trees draws generated by BART: $\frac{1}{K}\Sigma_{k=1}^K f_k^*(x)$

Use BART to estimate treatment effects:

- Given the outcome variable Y_i, treatment variable Z_i and confounding variables X_i, BART can directly estimate treatment effects by fitting the conditional expectation functions

$E[Y_i(1)|X_i = x] = f(1, x)$ and $E[Y_i(0)|X_i = x] = f(0, x)$.

- See equation 33
- Each step of MCMC backfitting generates new draw of $f(\cdot)$
- Retrieve Individual Treatment Effects at the k-th draw by:
  - $\tau_i^k = f^k(1, x_i) - f^k(0, x_i) \quad \forall i = 1, \ldots, N$
- Calculate ATE at k-th draw by taking the average over the corresponding population
- Iterating over all K draws yields the Monte Carlo approximation of the posterior distribution of the parameter of interest =>
- Compute ATE estimate by posterior mean $\beta_{ate} = \frac{1}{K}\Sigma_{k=1}^{K}\beta_{ate}^k$
- posterior intervals via normal approximation by the posterior mean plus/minus the critical value times the posterior standard deviation.

DGP properties

- Each DGP setting represents a combination of 5 criteria:
  1. degree of non-linearity in the response & treatment function
     1. Contain polynomial terms, interactions between covariates, and indicator or step functions
  2. percentage of treated
  3. overlap => full overlap indicates that the propensity score is bounded away from zero and one, so that common support is satisfied
     1. To violate, consider simulation settings with penalized overlap, by forcing observations to have a propensity score of zero.
  4. alignment => the degree to which the response and treatment model share the same confounding effects or contain noise.
     1. Some confounders might only affect the response model and others only the treatment model, or the same confounder might impact both treatment and response in different functional forms.
  5. treatment effect heterogeneity
     1. 'none' denotes that the treatment effect is constant conditional on covariates, 'low' denotes interactions with three covariates and 'high' denotes interactions with six covariates

Evaluation

- RMSE on ATE and coverage rate of interval estimates over R simulations (See equation 35)
- Bias by taking difference between true param & estimator
- Interval length (ub - lb)
- They do GATEs too but that's unimportant for us
- ITEs by using DR MOM, Causal Forest, and BART MChains (!) => we can do the latter two nicely
  - Compute ITE on validation set by holding out 50% of the data at each replication

> - The true ITE are calculated for each replication by $\tau_{r,i}(X_i) = E[Y_{r,i}(1)|X_i] - E[Y_{r,i}(0)|X_i]$
> - The evaluation of ITEs is done by Precision in the Estimated Heterogeneous Effects (PEHE), i.e. the RMSE of the ITE: $(\frac{1}{N}\Sigma_i^N(\tau_r(\hat{X}_i) - \tau_r(X_i))^2)^{0.5}$ where N is number of individuals
> - Then run the precision over R simulations and take the mean

See table 1 page 25 => i think we can either do only ATE or ATE + ITE for Causal Forest + BART MChains