STA 640 — Causal Inference

Chapter 4. Treatment Effect Heterogeneity (and Machine Learning)

Fan Li

Department of Statistical Science

Duke University

Outline

- Acknowledgement: Large part of this lecture is written by Joey Antonelli of University of Florida
- Outline
  - Motivation and estimands for treatment effect heterogeneity
  - Subgroup analysis
  - Simple, parametric approaches to estimating heterogeneity of the causal effect
  - Flexible, nonparametric approaches
    - Introduction to different algorithms for flexible treatment effect estimation
    - Review of state-of-the-art approaches

Motivation

- There is huge interest in understanding whether a treatment or policy affects certain individuals more than others
  - Referred to as treatment effect heterogeneity or heterogeneous treatment effects
- Personalized medicine is a huge area of interest
  - What treatment should an individual get
  - Physicians are implicitly considering how treatment effects vary when determining what treatment to assign a patient
    - Given their characteristics, treatment history, etc.

Motivation

- There are countless other applications for which heterogeneity of the treatment effect is of scientific interest
- Many cancer treatments only work on a subset of the population
  - Why? What subsets of the population?
- Limited resource settings where not everyone can be assigned treatment
  - Give it to those individuals most likely to benefit
- Helps to transport causal effects from one population to another
  - Two populations might have different characteristics and therefore different ATEs

Motivation

- An additional issue is that sometimes average or marginal treatment effects can mask the effect of a policy
- What if a policy has a positive impact on some individuals and a negative impact on others?
  - ATE will likely be very close to zero
  - Hypothesis tests indicate no treatment effect
  - In truth the treatment is very important
- Looking at heterogeneous treatment effects provides more scientific information than marginal effects alone
  - Immediately recover marginal effects from heterogeneous ones

Motivation

- There are many questions one can answer in a study of heterogenous treatment effects
  - Which covariates modify the treatment effect?
  - Is there any heterogeneity whatsoever?
  - For a given X, what is the expected treatment effect (CATE)
  - For a given individual, what is their treatment effect (ITE)
- Choice of statistical approach will depend on the goals of the study

Estimands of interest

- The most common target estimand is the conditional average treatment effect

$$\text{CATE} = \tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

- Note the ATE is simply the average CATE

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \int_x \tau(x) f_X(x) dx$$

- This shows how the CATE provides additional information over the ATE
  - Once we know the CATE, we immediately know the ATE

Estimands of interest

- Another relevant estimand refers to subgroup analysis
- Assume we have a subset of the covariate space defined by C, e.g. specific age or gender or medical history
- A subgroup specific estimand is given by

$$\mathbb{E}[Y(1) - Y(0)|X \in C]$$

- Commonly we will have non-overlapping regions given by $C_1, \ldots, C_G$, and we estimate

$$\mathbb{E}[Y(1) - Y(0)|X \in C_g] \text{ for } g = 1, \ldots, G$$

- And again we can easily recover the ATE by marginalizing over these

Estimands of interest

- Sometimes the CATE is not of interest, but focus is on a subset of predictors given by $V \subset X$:

$$\mathbb{E}[Y(1) - Y(0)|V = v]$$

- Maybe we simply care whether a particular covariate modifies the treatment effect
- This construction is really useful in high-dimensional settings where X is high-dimensional, but we care more about heterogeneity by certain covariates
  - Still need to account for X when adjusting for confounding, but not when estimating heterogeneous treatment effects

Estimands of interest

- Individual treatment effects (ITE) are also of concern

$$\tau_i = Y_i(1) - Y_i(0)$$

- For example, this is the question that personalized medicine looks to address
  - How will the treatment affect this particular individual
- Generally speaking, these are much harder to estimate
  - More uncertainty
  - Prediction intervals are wider than intervals for a mean
  - Stronger assumptions

Estimands of interest

- The literature often conflates the ITE and the CATE
- Clearly, we have that

$$Y_i(1) - Y_i(0) \neq \mathbb{E}[Y(1) - Y(0)|X = X_i]$$

- Related concepts, and certainly the CATE evaluated at X_i is a good point estimate for the ITE of individual i
- Under outcome modeling approach, all estimands are estimated in the same fashion

Identifying assumptions

- Estimation of heterogeneous treatment effects (HTE) differs from that of marginal treatment effects, but identification is effectively the same
- Easy to see that under SUTVA and unconfoundedness we have

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|Z = 1, X = x] - \mathbb{E}[Y(0)|Z = 0, X = x] \\ &= \mathbb{E}[Y|Z = 1, X = x] - \mathbb{E}[Y|Z = 0, X = x]\end{aligned}$$

- Unconfoundedness allows us to use data with $Z_i = 0$ to estimate $\mathbb{E}[Y(0)|X = x]$ in the whole population
  - Same for $Y(1)$

Identifying assumptions

- Overlap is still a fundamental assumption for heterogeneous treatment effects as well
  - With little overlap, causal inference is problematic conceptually and has large uncertainity operationally
- Suppose we have certain regions of the covariate space that are always treated
- We have to then extrapolate our estimates of $\mathbb{E}[Y|Z = 0, X = x]$ to these individuals with different covariate values
  - Heavily reliant on model specification
  - Difficult to understand the degree of extrapolation
  - Unclear impacts on uncertainty quantification
- We will discuss overlap a bit more in subgroup analysis

Identifying assumptions

- In this section, we will mostly cover estimation issues
  - There are a lot!
- A lot of other issues inherent to a causal analysis apply here as well
  - Considering plausibility of causal assumptions
  - Sensitivity analysis (to be covered in a couple of weeks)
  - Overlap and balance checks
- When these issues differ in ways unique to heterogeneous treatment effect estimation, we will cover them as they come up

Outline the lecture

- Motivation and estimands for treatment effect heterogeneity
- **Subgroup analysis**
- Simple, parametric approaches to estimating heterogeneity of the causal effect
- Flexible, nonparametric approaches (including machine learning)
  - Introduction to different algorithms for flexible treatment effect estimation
  - Review of state-of-the-art approaches

Subgroup analysis

- The simplest form of heterogeneity is subgroup analysis (SGA)
- Again suppose we have non-overlapping subsets of the covariate space given by $C_1, \ldots, C_G$
- Our goal is estimation of

$$\mathbb{E}[Y(1) - Y(0) | X \in C_g] \text{ for } g = 1, \ldots, G$$

- We will see that many of the same estimation strategies we've already learned about can be utilized here analogously

Subgroup analysis

- Important that these groups are chosen beforehand, often in a one-variable-at-a-time fashion
  - Might look old-fashioned, by still informative and widely used in practice, e.g. medical research
- There are data-driven approaches for finding the subsets of the population that benefit from treatment
- Generally speaking using the data to find subgroups complicates analyses
  - Valid inference becomes challenging
  - Post selection inference issues
  - Can use data splitting to alleviate these issues
- We will focus for now on situations where these groups are known beforehand

Subgroup analysis (SGA): weighting

- All balancing weights can be directly applied to SGA. Below we will focus on IPW for simplicity
- Recall the original IPW estimator of the ATE

$$\frac{1}{N}\left\{\sum_{i=1}^{N}\frac{Y_i Z_i}{e(X_i)}-\sum_{i=1}^{N}\frac{Y_i(1-Z_i)}{1-e(X_i)}\right\}$$

  which is used as a sample estimate of

$$\mathbb{E}\left[\frac{ZY}{e(X)}-\frac{(1-Z)Y}{1-e(X)}\right]$$

- We are using the empirical distribution from the sample to approximate this expectation that is with respect to the overall target population

Subgroup analysis: weighting

- Now our target population is the subset of individuals within C_g:

$$\mathbb{E}[Y(1) - Y(0) | X \in C_g] = \mathbb{E}\left[\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)} \middle| X \in C_g \right]$$

so a natural estimator of this is simply

$$\frac{1}{N_g} \left\{ \sum_{i:X_i \in C_g} \frac{Y_i Z_i}{e(X_i)} - \sum_{i:X_i \in C_g} \frac{Y_i (1-Z_i)}{1-e(X_i)} \right\}$$

where $N_g = \sum_{i=1}^{n} \mathbb{I}(X_i \in C_g)$

Subgroup analysis: weighting

- We simply use the IPW estimator but instead average over just the individuals in the desired subgroup
- The same procedure applies to other balancing weights, e.g. overlap weights, ATT weights
- Can apply this procedure separately within each subgroup to estimate subgroup specific effects
- Remember that balancing weights are intended to construct a weighted population for which the covariates are balanced across treatment groups
  - Does that happen here?

Subgroup analysis: variance-bias tradeoff

- An important question is how the PS is estimated
  - Using the entire sample
  - Using just the individuals in C_g
- Using the full sample aims to ensure balance in the entire target population, not the subgroup specific one
- Nonetheless, if the PS is correctly specified, using the full sample should work well
- Using just the individuals in C_g will improve balance within the subgroup
  - Less efficient. Bias/variance trade-off
- A simple logistic propensity score model with only main effects of all covariates is not usually adequate in SGA

Subgroup analysis: outcome modeling

- Similar issues occur for outcome modeling or doubly robust estimators
- Recall the outcome modeling estimator

$$\frac{1}{N}\left\{\sum_{i=1}^{n}\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right\}$$

- Can similarly replace this with

$$\frac{1}{N_g}\left\{\sum_{i:X_i \in C_g}\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right\}$$

in order to estimate the subgroup effect

Subgroup analysis

- Similar decisions need to be made here
- Fitting the outcome model only on individuals with $X_i \in C_g$ is more flexible, but also less efficient
- One alternative is to fit a model on the full sample, but include interactions between covariates and indicators of subgroup index
  - Similar to fitting separate regression models
  - Can use penalization on interaction terms to shrink/regularize towards the standard model fit on the full data
  - Balance bias and variance concerns
- One solution is the subgroup balancing propensity score (Dong et al. 2020): estimating PS that reaches a comprise between global and subgroup balance
- In general, cumbersome to implement

Post-LASSO Algorithm to Balance Bias-Variance Tradeoff

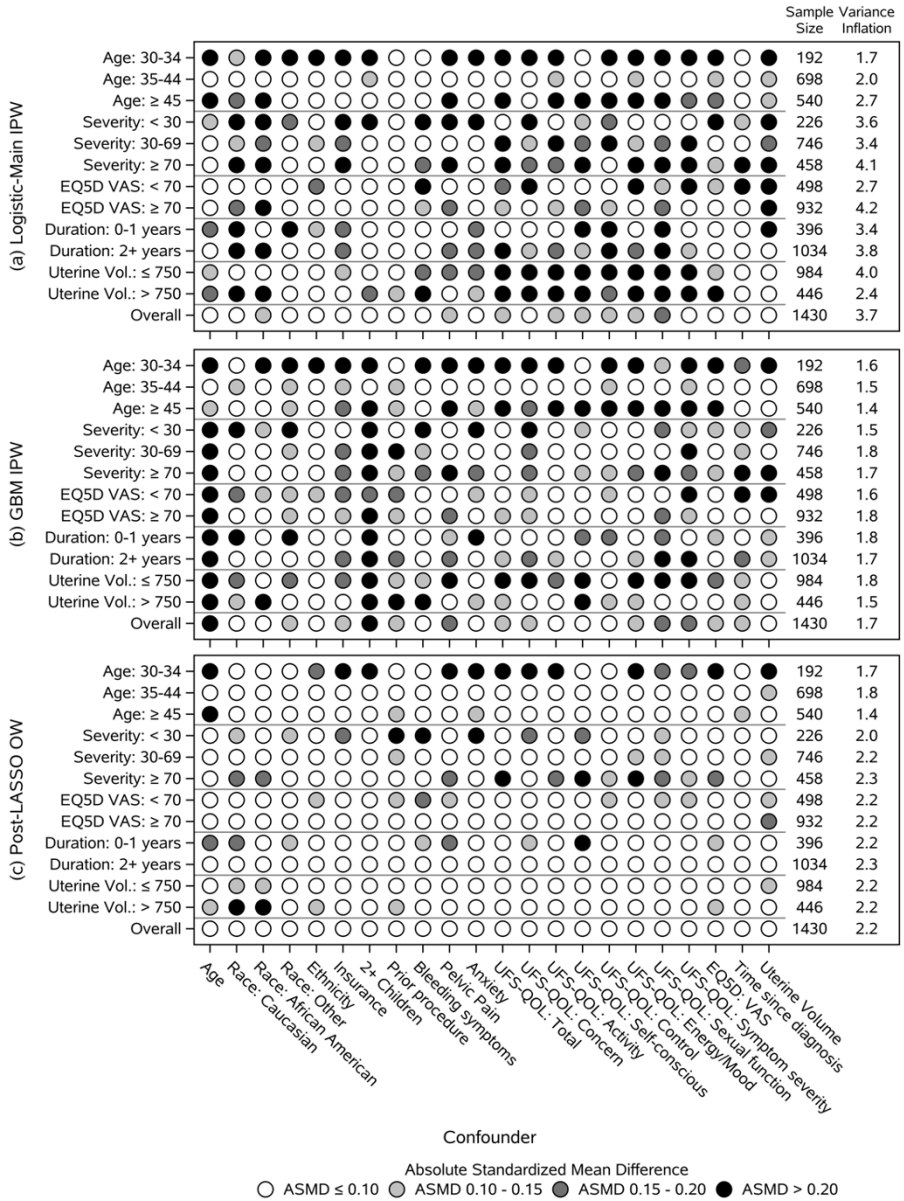
- Yang et al. 2021 (SIM) proposed to use post-LASSO estimate PS. The procedure is:
  - S1. Fit a logistic PS model with all pre-specified covariates and subgroup variables along with pairwise covariate-subgroup interactions, and perform LASSO to select covariate-subgroup interactions (without penalizing the main effects in the model).
  - S2. Estimate PS by refitting the logistic regression with all main effects and selected covariate-subgroup interactions from S1.
  - S3. Calculate a chosen type of weights (e.g. IPW or OW) based on the PS estimated from S2, and check subgroup balance before and after weighting.
  - S4. Estimate the causal effects for all prespecified subgroups using the Hajek estimator within subgroup with the weights from S3.

Visualizing Subgroup Balance: Connect-S plot

- Difficult to visualize subgroup balance. For K subgroups and p covariate, there are Kp standardized differences
- One can draw K love plots, each for p covariates, but still cumbersome
- Connect-S plot (Yang et al. 2021) visualizes Kp balance statistics all at once
  - each row represents a subgroup variable, (e.g. a race group)
  - each column represents a confounder/covariate that we want to balance (e.g. age).
  - Each dot corresponds to a specific subgroup and confounder, and the shade of the dot is coded based on the corresponding balance statistics, with darker color meaning more severe imbalance.

Connect S plot: example of COMPARE UF

Yang et al. 2021



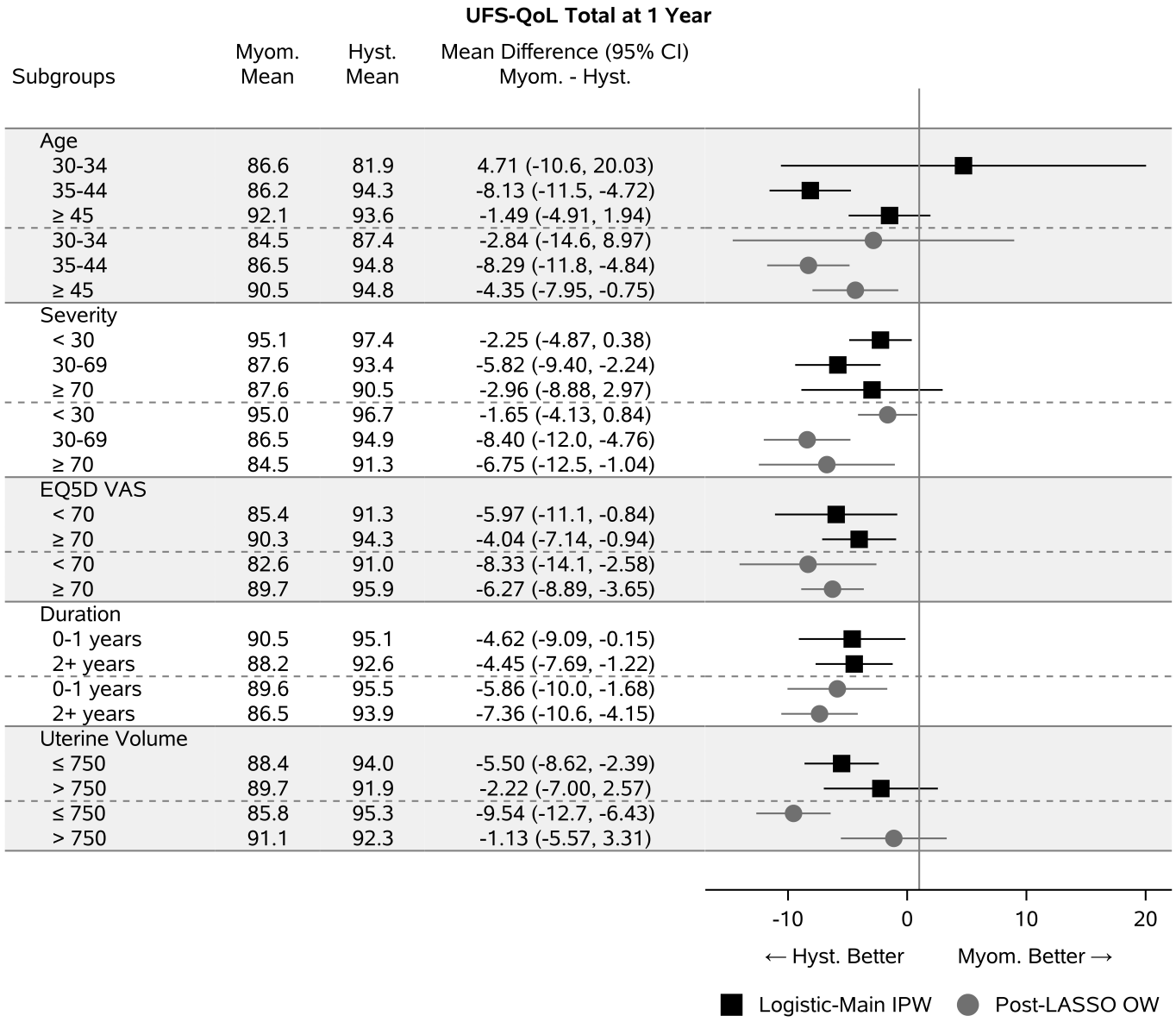
Case study of subgroup analysis - COMPARE UF

Yang et al. 2021

- Goal: determine whether certain patient subgroups should receive myomectomy versus hysterectomy (two treatments)
- Pre-specified 35 subgroups: categories of 16 variables including race, age, and baseline symptom severity
- 20 covariates/confounders: including demographics, disease history, quality of life and symptoms
- Total sample size: 1430, 567 in the myomectomy group and 863 patients in the hysterectomy group
- Outcome: quality of life score after 1 year
- Connect-S plot shows imbalance in many subgroup-confounder combinations

Case study of subgroup analysis - COMPARE UF

Yang et al. 2021



Outline the lecture

- Motivation and estimands for treatment effect heterogeneity
- Subgroup analysis
- **Simple, parametric approaches to estimating heterogeneity of the causal effect**
- Flexible, nonparametric approaches (including machine learning)
  - Introduction to different algorithms for flexible treatment effect estimation
  - Review of state-of-the-art approaches

CATE estimation: basic interaction approaches

- Subgroup analysis: subgroups are pre-specified, static
- The literature has increasingly moved towards identify subgroups with significant effects post-analysis, dynamic
- As discussed earlier, under unconfoundedness

$$\mathbb{E}[Y(1) - Y(0)|X = x] = \mathbb{E}[Y|Z = 1, X = x] - \mathbb{E}[Y|Z = 0, X = x]$$

- This implies we can simply build an outcome model for $f(z, x) = \mathbb{E}[Y|Z = z, X = x]$
- Once we have estimates of this outcome model, we have estimates of the CATE $\widehat{\tau}(x) = \widehat{f}(1, x) - \widehat{f}(0, x)$

CATE estimation: Basic interaction approaches

- In principle, any outcome regression model (e.g. a simple linear regression) can used to calculate CATE
- The simplest approach is with a linear model

$$f(Z, X) = \beta_0 + \beta_x X + \beta_z Z + \beta_{zx} ZX$$

- Related approaches for other models, such as SVMs (Imai and Ratkovic, 2013)
- Easy to see that $\tau(x) = \beta_z + \beta_{zx} X$
- If we center X, then the ATE is simply

$$E(Y(1) - Y(0)) = \beta_z$$

- Otherwise the ATE is given by

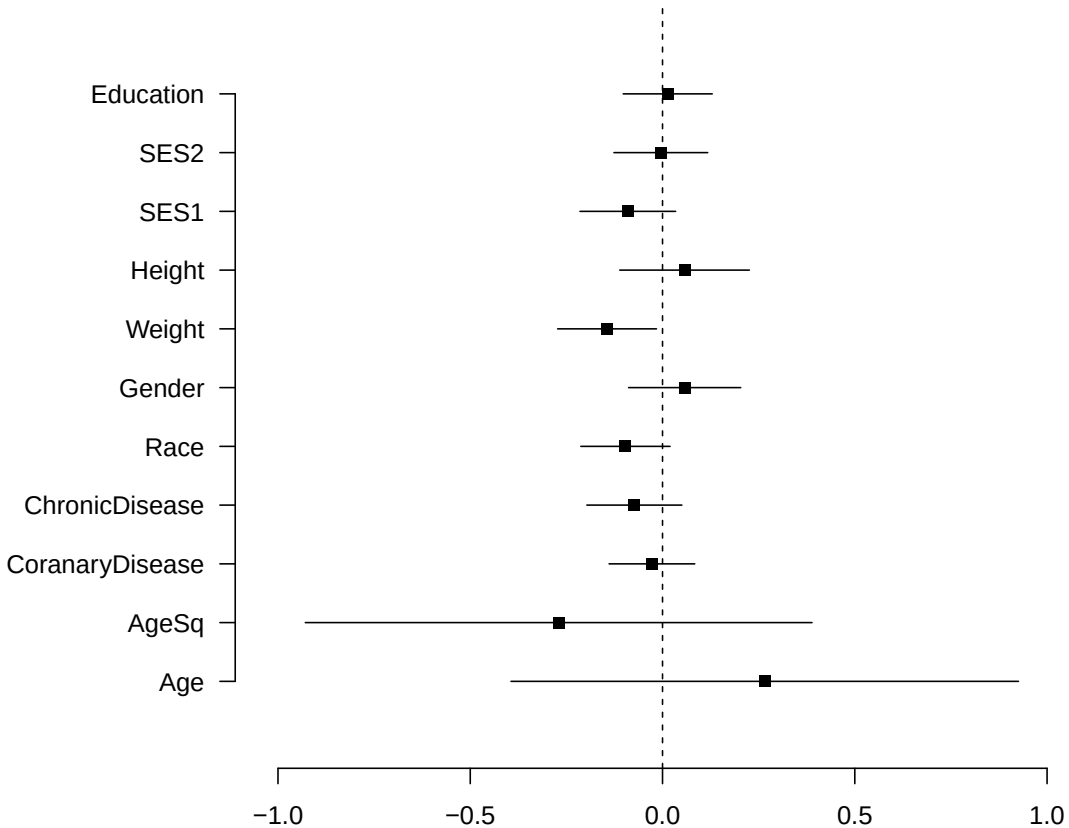
$$E(Y(1) - Y(0)) = \beta_z + \beta_{zx}\, \mathbb{E}(X)$$

CATE estimation: Basic interaction approaches

- Two main reasons why one might like this approach
  - Simple and easy to implement
  - Very interpretable
- A lot of questions are easy to answer in this framework
- Which covariates modify the treatment effect most
  - Examine magnitude of individual β_{zx} values
- Is there any treatment effect heterogeneity?
  - Amounts to testing $H_0 : \beta_{zx} = 0$

CATE estimation: Basic interaction approaches

- Below are estimates of β_{zx} from the NHANES analysis
- Overall ATE is estimated to be -0.08 (-0.19, 0.03)
  - More pronounced, negative effect in individuals with higher weight



CATE estimation: Basic interaction approaches

- A very related approach is to specify separate models in the treated and control groups

$$f(1, X) = \beta_{01} + \beta_{x1} X$$

$$f(0, X) = \beta_{00} + \beta_{x0} X$$

- The CATE is therefore

$$\tau(x) = \beta_{01} - \beta_{00} + (\beta_{x1} - \beta_{x0})x$$

- Treated individuals used to estimate $f(1, X)$ and vice-versa
- In linear models, these two approaches are identical
- Once we jump to nonlinear, flexible approaches these two will behave much differently

Outline the lecture

- Motivation and estimands for treatment effect heterogeneity
- Subgroup analysis
- Simple, parametric approaches to estimating heterogeneity of the causal effect
- **Flexible, nonparametric approaches (including machine learning)**
  - Introduction to different algorithms for flexible treatment effect estimation
  - Review of state-of-the-art approaches

Flexible CATE estimators

- There has been a dramatic increase in semiparametric or nonparametric estimators of the CATE that utilize modern statistical learning tools
  - Bayesian nonparametric approaches
  - Machine learning (Trees, high-dimensional models, etc.)
- Throughout the rest of the lecture, we will review many of these approaches
  - Discuss pros and cons of each
- Some are left out, but this will cover many of the core ideas

Covariate adjustment in Randomized Experiments

Bloniarz et al., 2016, PNAS

- Recap: In randomized experiments, covariate adjustment via OLS regression (with treatment, covariates and full set of trt-cov interactions) can reduce variance of ATE (Lin, 2013, AOAS)
- What if the covariates are high dimensional? Dimension reduction is necessary
- Bloniarz et al. (2016) recommend to replace OLS with LASSO:
  - Use post-LASSO (Belloni and Chernozhukov, 2013), i.e. first use LASSO for variable selection, and then perform OLS on the selected variable
  - Post-LASSO is performed separately in control and treatment group
- The above LASSO-adjusted regression estimator is consistent and more efficient than the unadjusted difference-in-means

S-Learners

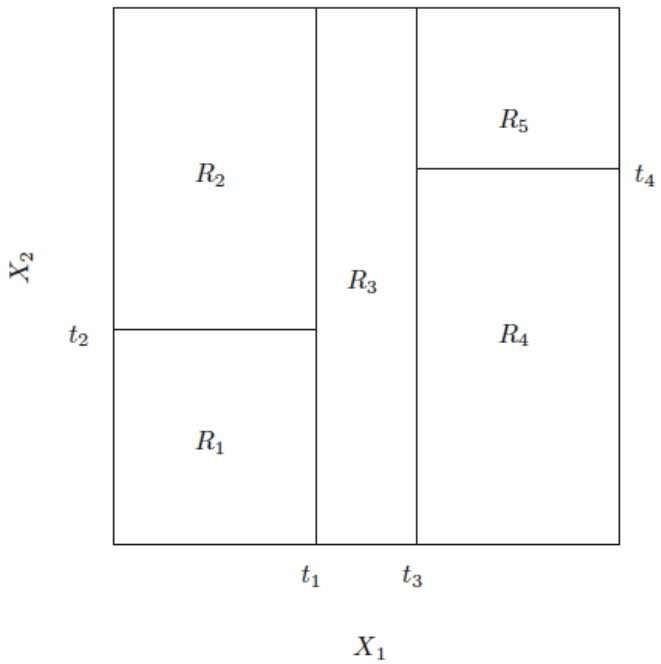
- One class of outcome modeling approach is sometimes referred to as S-learners (S refers to single)
- Exploit the fact that

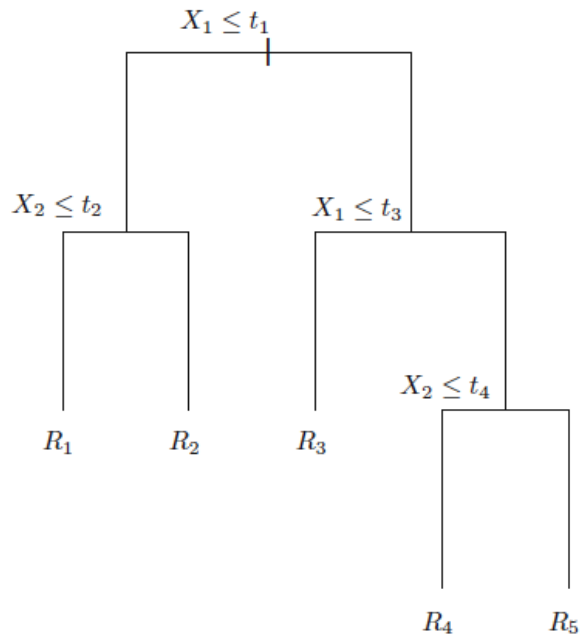
$$\tau(x) = f(1, x) - f(0, x)$$

- Focus solely on flexible estimation of $f(z, x)$
  - CATE estimation is automatic after this
- There are countless machine learning approaches to estimating $f(z, x)$
- One of the seminal papers in this regard is by Jennifer Hill (2011)

Brief review of regression trees

- Regression trees partition the covariate space into non-overlapping regions
- Predictions in each region based solely on data that falls in that region, R_j





James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

BART approaches to CATE estimation

- Main idea in Hill (2011) is to use BART to estimate $f(z, x)$
- BART assumes that

$$f(z, x) = \sum_{t=1}^{T} g(x, z; \mathcal{T}_t, \mathcal{M}_t)$$

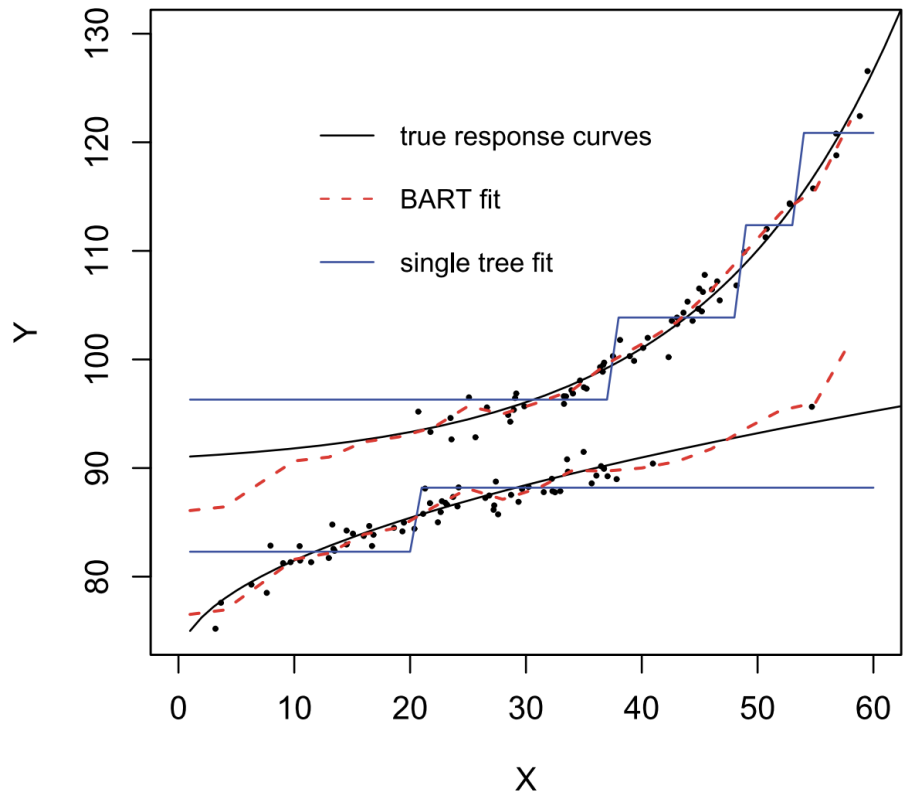
- Here, $g(x, z; \mathcal{T}_t, \mathcal{M}_t)$ is a tree that partitions the space of x and z
  - $\mathcal{T}_t$ represents the tree structure (where splits are)
  - $\mathcal{M}_t$ are parameters for predictions in each terminal node of the tree
- $\mathcal{M}_t = (\mu_{t1}, \ldots \mu_{tL_t})$ where L_t is the number of terminal nodes

BART approaches to CATE estimation

- BART is a Bayesian approach, and certain priors are placed on the parameters of the tree
- The prior probability of splitting decreases with tree depth
  - Probability of splitting at node depth k is $\gamma(1+k)^{-\beta}$ with $\gamma, \beta > 0$
- Shrinkage of mean parameters in each terminal mode are shrunk by a factor of T
  - $\mu_{tl} \sim \mathcal{N}(0, \sigma_u^2/T)$
- My experience is that this greatly outperforms random forests
  - Inference also easy in the Bayesian paradigm
  - Effectively tuning parameter free (defaults work well)
  - For more details, read Chipman et al. (2010)

BART approaches to CATE estimation

- Also much better than using a single regression tree
  - Not surprising given performance of boosting or RFs compared to a single tree



Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." Journal of Computational and Graphical Statistics 20.1 (2011): 217-240.

BART approaches to CATE estimation

- This approach is flexible, automatic, and easy to use
- There are some potential drawbacks
- Putting a BART prior distribution on the response surface $f(z, x)$ has unknown implications for the parameter of interest, $\tau(x)$
- Generally speaking, especially in flexible models, we should be careful about the implications of our prior specification on the parameter of interest
  - Do we expect the CATE to be as complex as $f(z, x)$?

BART approaches to CATE estimation

- These issues were addressed in Hahn et al. (2020) - Bayesian Causal Forest (BCF)
- Main idea is to re-parameterize

$$f(z, x) = \mu(x) + \tau(x)z$$

- Nonparametric extension of the basic interaction approaches we saw earlier
- $\mu(x)$ adjusts for confounding by X
- $\tau(x)$ allows for heterogeneity of the treatment effect
- Separate BART prior distributions placed on these two functions
  - Can use simpler trees for $\tau(x)$

BART approaches to CATE estimation

- The authors further advocate for inclusion of the propensity score

$$f(z, x) = \mu(x, \widehat{e}(x)) + \tau(x)z$$

- This improves our ability to adjust for confounding
- Avoids an issue called regularization induced confounding
  - Unintended bias that occurs when we are not careful about how we implement regularization or shrinkage in high-dimensional or nonparametric situations
  - Our model might indirectly shrink degree of confounding bias to zero, which is bad when there is severe confounding

T-learner

- An extension of these ideas that is even more flexible is the T-learner (T refers to “two")
- The previous approach used all of the data to fit one model

$$E(Y \mid Z = z, X = x) = f(z, x)$$

- A T-learner fits separate models to the treated and control groups

$$E(Y \mid Z = 1, X = x) = f_1(x)$$
$$E(Y \mid Z = 0, X = x) = f_0(x)$$

and the CATE is simply

$$\tau(x) = f_1(x) - f_0(x)$$

T-learner

- A couple advantages to this approach
  - Extremely flexible
  - Works well when $f_z(x)$ differs greatly across $z = 0, 1$
- Some drawbacks as well
  - Too flexible! Highly variable
  - Difficult to estimate $f_z(x)$ when treatment group z has few individuals
  - Again no control of $\tau(x)$

T-learner

- Suppose we estimate $f_z(x)$ separately in each group and we have that

$$\text{Var}(\widehat{f_1}(x)) = v_1, \quad \text{Var}(\widehat{f_0}(x)) = v_0$$

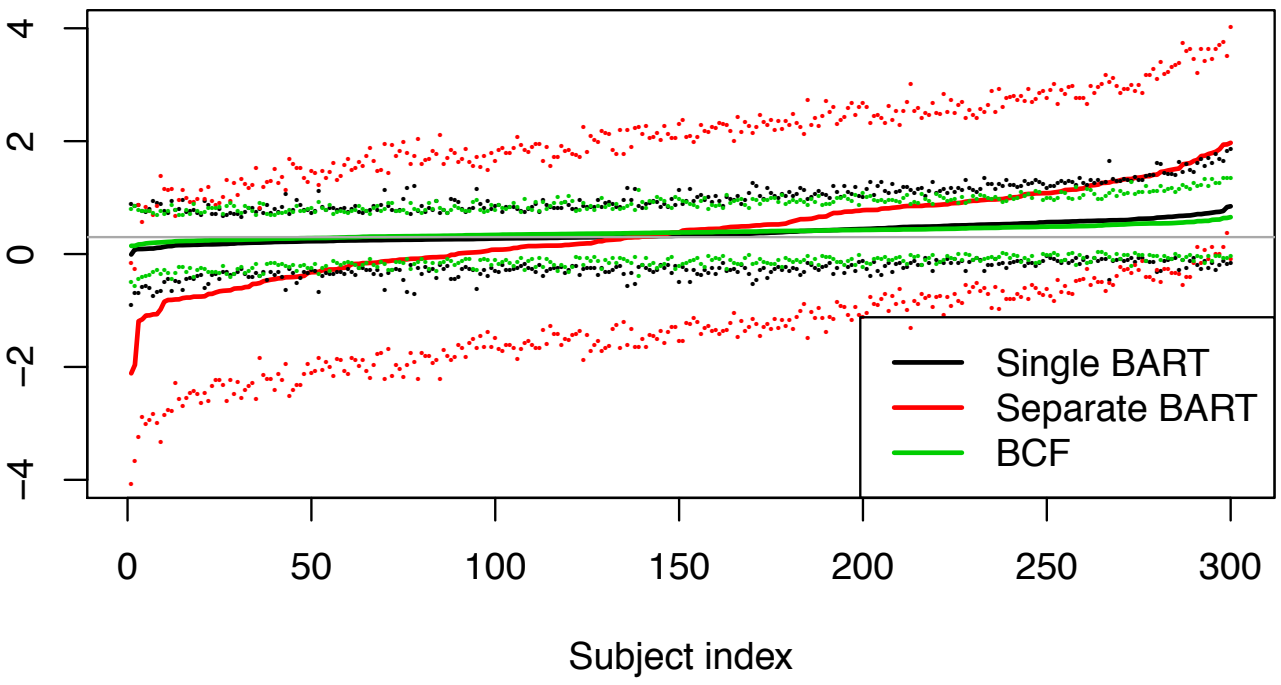
- Due to independence of individuals

$$\text{Var}(\widehat{\tau}(x)) = v_0 + v_1$$

- The variance of the treatment effect is greater than both of the individual functions!
  - Does this coincide with our prior knowledge about the treatment effect function?
  - We generally expect the treatment effect to be as simple, or simpler than $f_z(x)$

T-learner

- Below are estimates and confidence intervals for $\tau(X_i)$ for $i = 1, \ldots n$ in a simulated data set with no heterogeneity
- Separate BART models leads to extremely wide intervals and variable estimates



T-learner

- Now we will discuss a number of ways to address this problem
- One way is to impose some structure on $f_z(x)$
  - Put shrinkage directly on $\tau(x)$ as in Hahn et al. (2020)
  - R-learners, which use a specific loss function and a penalty on $\tau(x)$
  - Multi-task learners put shared structure on $f_1(x)$ and $f_0(x)$
- Another line of approaches constructs pseudo-outcomes and regresses them against X
  - Connections to IPW and DR estimators
- Some approaches directly estimate the CATE
  - Causal forests, related tree-based approaches

Multi-task learning

- Multitask learning views the two potential outcomes as outputs from a function $f : \mathcal{X} \to \mathbb{R}^2$
- The CATE is defined as

$$\widehat{\tau}(x) = \widehat{f}_1(x) - \widehat{f}_0(x) = \widehat{f}^T e, \quad e = [-1, 1]$$

- This looks a lot like the T-learner, which estimated these two functions separately
- Instead, multi-task learning estimates the two of them jointly
  - Borrow information across groups

Multi-task learning with a Gaussian process

- One implementation of this approach uses Gaussian processes to model f (Alaa et al. 2017)
- We assume that the potential outcomes come from

$$Y_i(0) = f_0(X_i) + \epsilon_{i,0}$$
$$Y_i(1) = f_1(X_i) + \epsilon_{i,1}$$

- And we place a Gaussian process prior on f

$$f \sim \mathcal{GP}(0, K)$$

- We won't discuss Gaussian processes in detail, but they are a nonparametric Bayesian formulation to flexibly modeling functions

Multi-task learning with a Gaussian process

- GPs are nice and have been shown to work well in many settings
  - Only assume smoothness of the functions
  - Nearby x values should have similar potential outcomes
- Depending on the choice of kernel function K, this allows the two potential outcome surfaces $f_0(x)$ and $f_1(x)$ to be correlated
- This allows the two surfaces to have different functional forms, but borrows information across both groups and shrinks toward having similar functions

R-learners

- R-learners use a clever parameterization of the problem to directly estimate and regularize the CATE
- As with the Bayesian Causal Forest approach, let

$$Y_i = \mu(X_i) + \tau(X_i)Z_i + \epsilon_i$$

and if we take the conditional expectation of this, we obtain

$$m(X_i) = E(Y_i \mid X_i) = \mu(X_i) + \tau(X_i)e(X_i)$$

R-learners

- As first pointed out in Robinson (1988), these imply that

$$Y_i - m(X_i) = (Z_i - e(X_i))\tau(X_i) + \epsilon_i$$

- Which further implies that

$$\tau(\cdot) = \underset{\tau}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\Big((Y_i - m(X_i)) - (Z_i - e(X_i))\tau(X_i) \Big)^2 \right] \right\}$$

- Nie and Wager (2021) build on these ideas to estimate heterogeneous treatment effects

R-learners

- Their main idea is to estimate the CATE in the following way:

$$\tau(\cdot) = \underset{\tau}{\operatorname{argmin}} \left\{ \widehat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\}$$

where

$$\widehat{L}_n(\tau(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \left(\left(Y_i - \widehat{m}^{-i}(X_i) \right) - \left(Z_i - \widehat{e}^{-i}(X_i) \right) \tau(X_i) \right)^2$$

- $\Lambda_n(\tau(\cdot))$ is a penalty on the complexity of the CATE
  - Many options such as smoothness penalties, lasso, etc.

R-learners

- Note that we used $\widehat{m}^{-i}(X_i)$ and $\widehat{e}^{-i}(X_i)$ in the squared error loss
- These are estimates of the conditional mean outcome regression and propensity score with the i^{th} observation removed
  - Typically done using 5 or 10-fold cross validation, not leave one out
- This approach separated the problem into two separate stages
  - Estimating nuisance functions, $m(\cdot)$ and $e(\cdot)$
  - Estimation of $\tau(\cdot)$ conditional on nuisance function estimates
- Allows for separate penalization in these two steps
  - Allows for the CATE to be much simpler than the outcome regression functions

R-learners

- This approach directly addressed the problems of the T-learner
- They show this approach can be used with many modern machine learning type of estimators for the CATE
  - High-dimensional models
  - Gradient boosting
  - Neural networks
- New research still ongoing in this area

Tree-based approaches

- The last set of approaches we will discuss are tree-based approaches
- The overarching goal of these approaches is to find subsets of the data where the treatment effect varies the most
- No need to specify functional form for $\tau(x)$
  - Assumed constant within areas of covariate space
- Key papers in this area are Athey and Wager (2018), Athey et al. (2019), and Powers et al. (2018)

A quick remark on regression trees

- Before discussing causal trees, we need to discuss one aspect of regression trees
- How do we determine the tree structure?
  - Which covariates to split on?
  - What value of a covariate do we split at?
- In regression trees, we pick splits that reduce the MSE the most among all possible splits
  - Or gini index / classification error for categorical outcomes
- Greedy algorithm that successively creates splits that improve the model the most

A quick remark on regression trees

- Suppose I'm at the top of a tree and haven't split yet
- My current prediction is $\widehat{Y}_i = \overline{Y}$ for all i
- Now we find the values of j and s that minimize

$$\sum_{i:X_i \in R_1(j,s)} (Y_i - \widehat{Y}_{R_1})^2 + \sum_{i:X_i \in R_2(j,s)} (Y_i - \widehat{Y}_{R_2})^2$$

where

$$R_1(j,s) = \{X | X_j < s\} \quad R_2(j,s) = \{X | X_j \geq s\}$$

- And predictions in these regions are just sample averages (within group)

$$\widehat{Y}_{R_1} = \frac{\sum_{i=1}^n 1(X_i \in R_1(j,s)) Y_i}{\sum_{i=1}^n 1(X_i \in R_1(j,s))} \quad \widehat{Y}_{R_2} = \frac{\sum_{i=1}^n 1(X_i \in R_2(j,s)) Y_i}{\sum_{i=1}^n 1(X_i \in R_2(j,s))}$$

Causal forests

- Causal trees are constructed in a similar way
- Key difference is that instead of splitting to reduce MSE the most, we split to maximize heterogeneity of the treatment effect
  - This will lead us to finding areas of the covariate space with different treatment effects
- In regression trees, the estimates in the terminal nodes are sample averages of the outcome
- For causal trees, we estimate the treatment effect within each terminal node separately
- There are multiple causal tree algorithms, but we will mostly focus on the original one from Athey and Wager (2018)

Tune a tree model for causal inference

- In ML models, a crucial step is to use cross-validation to tune hyperparameters: split the data into training (build model) and testing data (check model)
- In prediction problems, the standard performance metric is prediction MSE
- Similarly for a estimator of a causal estimand, say a ITE estimator $\hat{\tau}(x)$, we may use a MSE:

$$L(\hat{\tau}) = E[(Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2].$$

- But wait... this is usually not possible in causal inference problems, because even in the test data we do not know the true causal effect (Rolling and Yang, 2014; Athey and Imbens, 2016)
- So we would need approximations to the truth

Honesty Criterion

Athey and Imbens, 2016, PNAS; Athey Tibshirani and Wager 2018, AOS

- Honesty criterion: a sample can only be used to estimate τ or decide how to build the model (e.g. where to place the splits in trees), but not both.
- Intuition: avoid using data twice
- Implementation: the study sample is divided into three subsamples: two for training (one for building the tree and one for estimating causal effects) and one for testing
- Wager and Athey (2018) devised two tree-based *honest* procedure to estimate ITE: (i) double-sample (outcome) tree, and (ii) propensity tree (discuss later)
- Honesty is important to achieve asymptotic normality and unbiasedness.

Causal forests

- Suppose we have a tree with terminal nodes or leaves given by $L_1(x), \ldots, L_K(X)$
- In leaf k, we can estimate the treatment effect as

$$\frac{1}{|\{i : X_i \in L_k(x), Z_i = 1\}|} \sum_{i:X_i \in L_k(x), Z_i=1} Y_i - \frac{1}{|\{i : X_i \in L_k(x), Z_i = 0\}|} \sum_{i:X_i \in L_k(x), Z_i=0} Y_i$$

- The hope is that within leaf k, individuals have similar covariate values and therefore the treatment is as if randomized
  - And therefore the difference in means estimator is unbiased

Causal forests

- Now suppose that we're considering a split of a parent node into two separate nodes
- The estimated treatment effects in each new node are given by $\widehat{\tau}_l$ and $\widehat{\tau}_r$
- One approach to finding splits is to calculate

$$\frac{|\widehat{\tau}_l - \widehat{\tau}_r|}{\sqrt{\widehat{\text{Var}}(\widehat{\tau}_l) + \widehat{\text{Var}}(\widehat{\tau}_r)}}$$

  and choose the split that maximizes this
- Other approaches explored in Athey and Imbens (2016)

Causal forests

- This will perform well for estimating treatment effects if treatment is unconfounded within leaves
- As suggested in Powers et al. (2018), you can perform additional adjustment
  - Propensity score stratification within leaves
  - Other approaches to confounding adjustment certainly possible
  - Requires larger amount of data within leaves
- Can also incorporate propensity scores into choice of splits
  - Ensures individuals in same leaf have similar PS values
  - No longer maximizes heterogeneity of treatment effect

Causal forests

- Inference in random forests models is typically very hard!
- Athey and Wager (2018) show how inference can be performed for random forests and causal forests
- Sample splitting is used such that
  1. Part of the data is used to find splits, i.e. tree structure
  2. Other part of the data is used to estimate treatment effects within leaves
- They show this leads to asymptotic normality of results with variance estimated by the infinitesimal jacknife (Wager et al. 2014)

Causal forests

- Throughout we've discussed creating splits for a single tree, but generally this is repeated a large number of times and results are averaged over all trees (as in random forests)
- We described the simplest type of causal forest
- Many extensions have been proposed that might perform better empirically
  - See Athey et al. (2019) for some ideas
- See also Powers et al. (2018) for other related algorithms such as boosting and MARS that are based on similar ideas

Causal forests

- The most recent version of these causal forests (that I'm aware of) involves combining causal forests with the R-learner from earlier
- Can create a pseudo-outcome as in the R-learner (also using regression trees to estimate $e(X)$ and $m(X)$)
- As in the R-learner, minimize

$$\tau(\cdot) = \underset{\tau}{\operatorname{argmin}} \left\{ \widehat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\}$$

- And now, the splits of the tree can be chosen to minimize this quantity

Double/De-biased Machine Learning (DML)

- Propensity score plays a central role in causal inference with observational data
- Within the outcome model approach, what's role of PS? Is it ignorable? Obviously not
- Chenozhukov et al. 2018: two main points
  - Point 1: ML methods are remarkably effective in prediction contexts. However, good performance in prediction does not necessarily translate into good performance for estimation or inference about "causal" parameters. In fact, the performance can be poor.
  - Point 2: By doing "double/di-biased/Neyman orthogonalized" ML and sample splitting, one can construct high quality point and interval estimates of "causal" parameters.

(Post-)Double selection of confounders

- A popular recent trend is to use machine learning (ML) methods in double-robust (DR) estimators to estimate ATE
- Main idea: specify ML models for both propensity score and outcome models (Farrell, 2015)
- With high-dimensional confounders, Belloni et al. (2014, RoES) proposes a double-selection procedure
  - Select confounders/covariates for the propensity score model and for the outcome model, e.g., by LASSO
  - Use least square estimation of the outcome with treatment indicator plus the union of selected confounders
- "Double-selection" gives $\sqrt{N}$ consistency of ATE, which "single-selection" cannot reach
- Chernozhukov et al. (AER, 2015) extended this procedure to settings with high-dimensional instrumental variables

Double/De-biased Machine Learning (DML)

- Machine learning (ML) estimators depend on regularization, and usually have slower rate of convergence, typically $n^{-\psi_m}$ with $\psi_m < 1/2$ (regularization bias)
- Naively apply ML estimators on IPW or OR estimators leads to higher-order bias that do not vanish asymptotically at $\sqrt{n}$ rate
- Consider directly applying ML to estimate each component of a double (e.g. DR) estimator (DML)
  - intuitively, the regularization bias term is proportional to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{N} \{\hat{e}(X_i) - e_{\text{true}(X_i)}\}\{\hat{m}_1(X_i) - m_{1,\text{true}}(X_i)\}$$

  - vanishes to zero at a (product) faster rate $n^{-(\psi_m+\psi_e)}$, where $\psi_m + \psi_e \geq 1/2$

Double/De-biased Machine Learning (DML)

- But one more issue: we have trained the ML estimator and predicted the missing potential outcomes using the same data
  - bias due to over-fitting
- Sample splitting and cross-fitting to control for such bias
  - Split data into main sample and auxiliary sample
  - Use the main sample to train and auxiliary sample to estimate
  - Swap the role of main sample and auxiliary sample and re-estimate
  - Average across splits
- Can extend to K-fold splitting – analogy with cross-validation
- Sample splitting also leads to valid variance and CI estimators

Bias due to Over-fitting

Chernozhukov et al. 2018

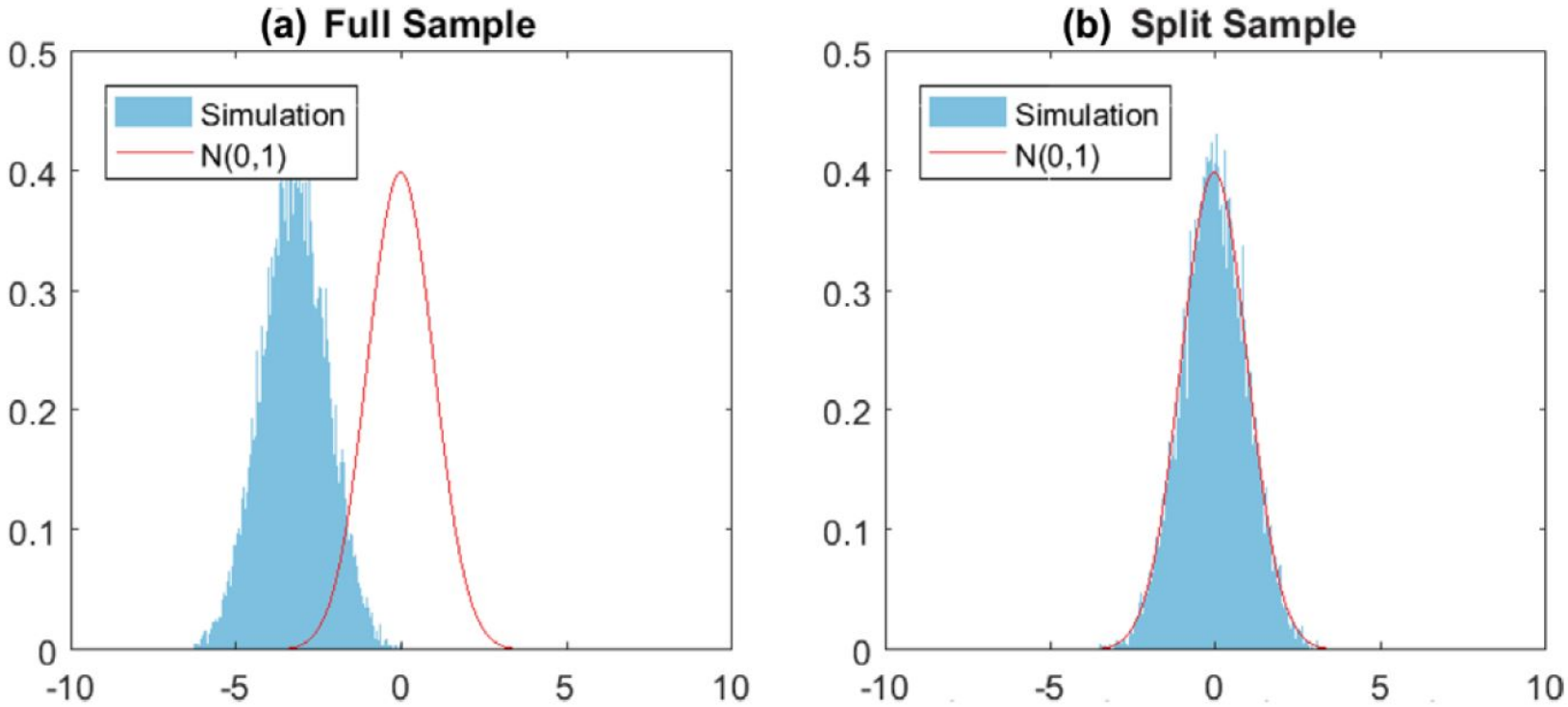


Figure 2. Comparison of full-sample and cross-fitting procedures. [Colour figure can be viewed at wileyonlinelibrary.com]

Operationalize DML for Estimating ATE

1. *split sample*: take a K-fold random partition $\{I_k\}_{k=1}^{K}$ of observation indices $\{1 \ldots, N\}$ such that the size of each fold $\text{card}(I_k) = N/K$. Also define for each k, $I_k^c = \{1, \ldots, N\} \backslash I_k$
2. *estimate PS and OR in training sample:* for $k = 1$, construct ML estimators (random forest or other "best performing ML" tools) for PS and OR on training data I_k^c
3. *estimate ATE in prediction sample:* Solve the doubly robust estimating equations for τ (which satisfy Neyman orthogonality condition) using $I_k \Leftrightarrow$ plug in the estimated PS and OR in the DR estimator, and obtain $\hat{\tau}^k$ using prediction sample I_k
4. *aggregate over K folds:* Repeat 2-3 for $k = 2, \ldots, K$ so that DML uses the full data

Example of DML: Partial Linear Model

Chernozhukov et al. 2018

- Partial linear model: $Y = Z\theta + g(X) + U$, where Z is treatment, and X is a high-dimensional vector of confounders, $\mathbb{E}(U|X, Z) = 0$, θ is the causal estimand.
- X are confounders in the sense: $Z = c + m_0(X) + V$ with $\mathbb{E}(V|X) = 0$
- DML procedure
  - Predict Y and Z using X by $\widehat{\mathbb{E}(Y|X)}$ and $\widehat{\mathbb{E}(Z|X)}$, respectively, obtained using (your favorite) ML model such as random forest
  - Get residuals: $\hat{U} = Y - \widehat{\mathbb{E}(Y|X)}$, $\hat{V} = Z - \widehat{\mathbb{E}(Z|X)}$
  - Regress $\hat{U}$ on $\hat{V}$ to get the estimate of $\hat{\theta}$
- $\hat{\theta}$ is consistent and asymptotically normal.
- Frisch-Waugh-Lovell (1930s) style

DR with Machine Learning: TMLE

- Targeted Maximum Likelihood Estimation or Targeted Minimum Loss Estimation (TMLE) (van der Laan and Rubin, 2006, and series of following work)

  1. Obtain a preliminary estimate of $\{\hat{m}_z^{(0)}(X)\}$ of the outcome $E\{Y(z)|X\}$ based on a ML algorithm (e.g. an ensemble learner), and fit a parametric (or ML) PS model to estimate PS $\hat{e}(X)$
  2. Fit a canonical generalized linear model for $E\{Y(z)|X\}$, with link function $h(\cdot)$, offset term $h\{\hat{m}^{(0)}(X)\}$, and the single covariate – IP weights: $Z_i/\hat{e}(X)$

- TMLE uses (inverse of) PS as the additional covariate: recall discussion earlier on regression with the clever covariate

DR with Machine Learning: TMLE

- The logistic model in step (3) is called a fluctuation working model
- Without the fluctuation model, the algorithm is simply an OR estimator based on $N^{-1} \sum_{i=1}^{N} \hat{m}^{(0)}(X_i)$
- TMLE uses (inverse of) PS to fluctuate the initial regression
  - can show that the score of the stabilized fluctuation model at zero fluctuation ($\hat{\epsilon}_n = 0$) spans the doubly robust estimating function (recall discussion earlier on regression with the clever covariate)
- This is a fully iterated DR estimator

Representation Learning

Johansson, Shalit, Sontag, 2016, 2017 ICML

- With high-dimensional covariates, directly building outcome models $\mu_0(x)$ and $\mu_1(x)$ might be hard
- Representation learning: instead find a low-dimenisional transformation/representation $\Phi(X_i)$, based on which build the outcome model: $\mu_0(\Phi(X_i)), \mu_1(\Phi(X_i))$
- Johansson et al. (2016, 2017): combine predictive power and covariate balance to learn the representations
- Build modeling as a penalized optimization problem, minimizing a loss function consisting of two components:
  - prediction loss on observed data
  - the **distance** between the representations in two groups $p(\Phi(X)|Z_i = 1)$ and $p(\Phi(X)|Z_i = 0)$

Representation Learning

Johansson, Shalit, Sontag, 2016, 2017 ICML

- Find the outcome model μ_0, μ_1 and the representations Φ via minimizing the loss function

$$\arg\min_{\mu,\Phi} \left\{ \sum_{i=1} L(\mu_{Z_i}(\Phi(\mathbf{X}_i)), Y_i) + \kappa \cdot \mathcal{D}(\{\Phi(X_i)\}_{Z_i=0}, \{\Phi(X_i)\}_{Z_i=1}) \right\},$$

  where L is a loss function, $\mathcal{D}(\cdot,\cdot)$ is a distance metric, and κ is a hyperparameter controlling the importance of distance.
- L: usually squared prediction error $L = \sum_i (Y_i - \mu_{t=T_i}(\Phi(X_i)))^2$
- Above the first term measures the predictive power the representation Φ, the second term measures the distance between the representation distribution in treated and control groups
- Adding the distance penalty: (i) incorporate balance/overlap; (ii) theoretical guarantee to bound the loss of ITE predictions

Representation Learning

Johansson, Shalit, Sontag, 2016, 2017 ICML

- Choice of the distance metric is crucial: Wasserstein distance, Maximum Mean Discrepancy
- Use the entropy from entropy balancing as distance gives double-robustness representations (Zeng et al. 2020)

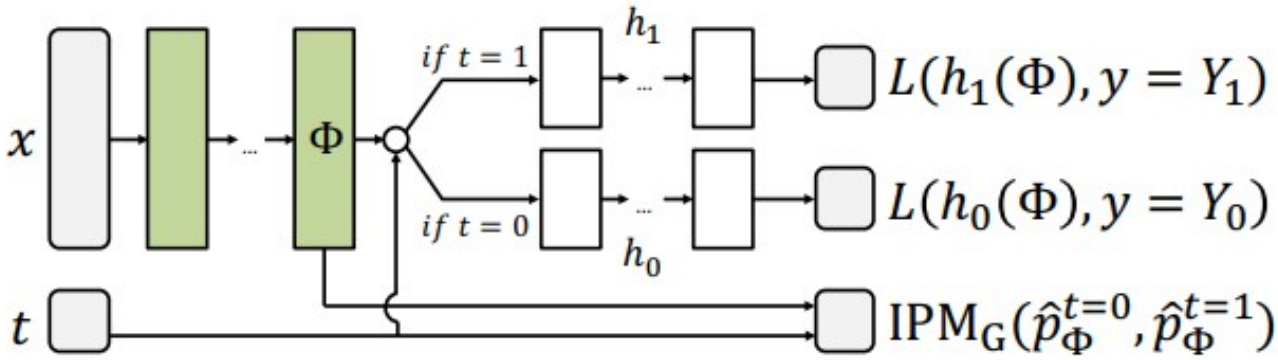


Figure: Neural network architecture for ITE estimation.L is the loss function for factual/observed data, IPM_G (integral probability metric)is the distance metric. $h_0(\Phi(\cdot)), h_1(\Phi(\cdot))$ (e.g. h is equivalent to μ) are the outcome models

Representation Learning

Johanson, Shalit, Sontag, 2016, 2017 ICML

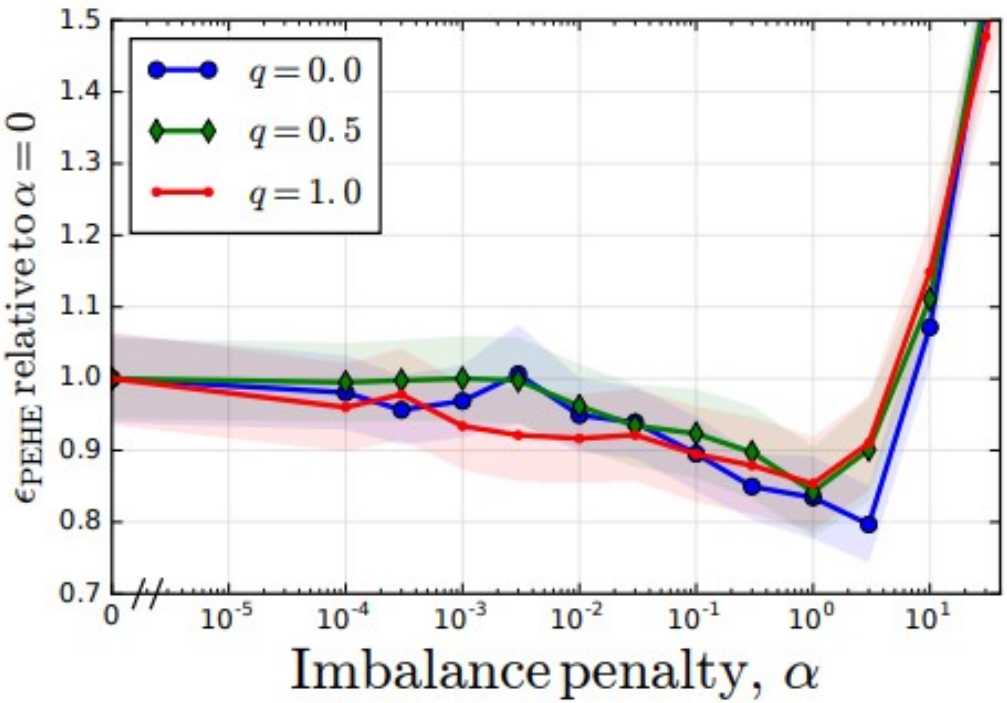


Figure: Change of ITE prediction error versus the importance of distance metrics (penalty on imbalance). A proper choice of hyperparameter is important.

Machine Learning and Causal Inference: Key Insights

- Machine learning greatly expands the toolbox for outcome modeling
- But machine learning does not magically solve the fundamental problem of causal inference
- The key issues in causal inference — overlap, balance, unconfoundedness — remain the same and requires more care
- To adapt machine learning methods to causal inference, one has to adapt to those key issues.
- Key insights:
  - Sample splitting: for building model and for estimating effects
  - Double learning: combine both PS model and outcome model for causal inference with high-dimensional data
  - Flexible (outcome) modeling
- Fast evolving field, we will go through a few popular methods

References

- Belloni, A., Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. Bernoulli, 19(2), 521-547.
- Belloni, A., Chernozhukov, V., Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608-650.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. Econometrica, 85(1), 233-298.
- Bloniarz, A., Liu, H., Zhang, C. H., Sekhon, J. S., Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. Proceedings of the National Academy of Sciences, 113(27), 7383-7390.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5), 261-65.
- Rolling, C. A., Yang, Y. (2014). Model selection for estimating treatment effects. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 749-769.
- Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189(1), 1-23.

References

Imai, Kosuke, and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." The Annals of Applied Statistics 7.1 (2013): 443-470.

Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." Journal of Computational and Graphical Statistics 20.1 (2011): 217-240.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics 4.1 (2010): 266-298.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." Bayesian Analysis 15.3 (2020): 965-1056.

Alaa, Ahmed M., and Mihaela van der Schaar. "Bayesian inference of individualized treatment effects using multi-task gaussian processes." arXiv preprint arXiv:1704.02801 (2017).

Nie, Xinkun, and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects." Biometrika 108.2 (2021): 299-319.

References

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

Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 (2018): 1228-1242.

Athey, Susan, Julie Tibshirani, and Stefan Wager. "Generalized random forests." The Annals of Statistics 47.2 (2019): 1148-1178.

Caron, Alberto, Gianluca Baio, and Ioanna Manolopoulou. "Estimating individual treatment effects using non-parametric regression models: a review." arXiv preprint arXiv:2009.06472 (2020).

Johansson, F., Shalit, U., Sontag, D. (2016). Learning representations for counterfactual inference. In International conference on machine learning (pp. 3020-3029).

Shalit, U., Johansson, F. D., Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning (pp. 3076-3085).

References

van der Laan, M. J., Rubin, D. (2006). Targeted maximum likelihood learning. The international journal of biostatistics, 2(1).

Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016): 7353-7360.

Wager, Stefan, Trevor Hastie, and Bradley Efron. "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife." The Journal of Machine Learning Research 15.1 (2014): 1625-1651.

Powers, Scott, et al. "Some methods for heterogeneous treatment effect estimation in high dimensions." Statistics in medicine 37.11 (2018): 1767-1787.

Dong J, Zhang J, Zeng S, and Li F. (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research*. 29(3) 659â676.

Yang S, Lorenzi E, Papadogeorgou G, Wojdyla D, Li F, Thomas LE. (2021). Propensity score weighting for causal subgroup analysis. *Statistics in Medicine*. 40:4294-4309