STA 790 (Fall 2022) — Bayesian Causal Inference

Chapter 4: Bayesian Approach to Heterogeneous Treatment Effects

Fan Li

Department of Statistical Science
Duke University

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
- 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

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

- 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
- 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 are also of concern

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

- This is exactly 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

- A lot of people in the literature conflate 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
- Some ongoing work in ITE estimation, but we will focus on average (conditional on some feature of the covariate distribution) estimands here

Identifying assumptions

- Estimation of heterogeneous treatment effects 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 an important assumption for heterogeneous treatment effects as well
  - For the same reasons as in ATE estimation

- 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 the subgroup analysis section

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

Basic interaction approaches

- As discussed earlier, under unconfoundedness

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

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

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

Basic interaction approaches

- Many ideas for CATE estimation stem from this approach
- The simplest such way 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

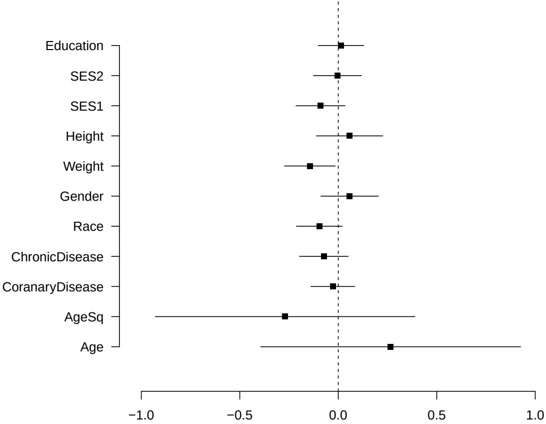
$$E(Y(1) - Y(0)) = \beta_z + \beta_{zx} \mathbb{E}(X)$$

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 $H_0 : \beta_{zx} = 0$

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



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

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

S-Learners

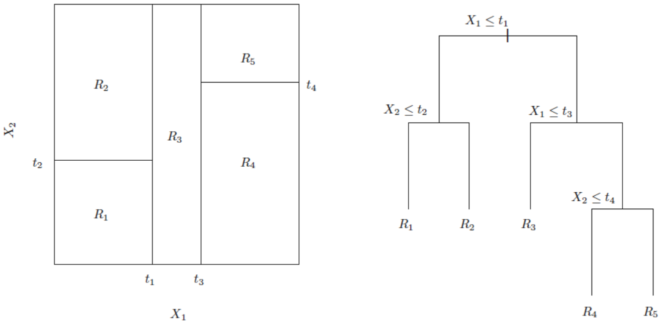
- One class of approach is sometimes referred to as S-learners
- Exploit the fact that

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

- 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

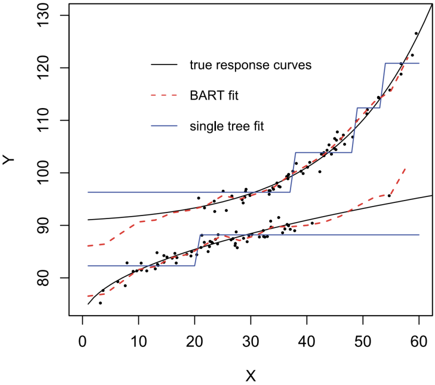
- The prior probability of splitting decreases with tree depth
  - Probability of splitting at node depth k is $\gamma(1+k)^{-\beta}$ with $\gamma, \beta > 0$

- 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)
- 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
- 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

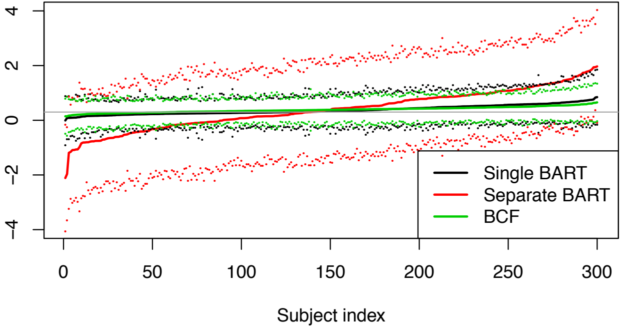
- A couple advantages to this approach
  - Extremely flexible
  - Works well when $f_z(x)$ differs greatly across $z = 0, 1$
- Some drawbacks as well
  - Too flexible! Highly variable
  - Difficult to estimate $f_z(x)$ when treatment group z has few individuals
  - Again no control of $\tau(x)$

T-learner

- Suppose we estimate $f_z(x)$ separately in each group and we have that
$$\mathrm{Var}(\widehat{f}_1(x)) = v_1, \quad \mathrm{Var}(\widehat{f}_0(x)) = v_0$$
- Due to independence of individuals
$$\mathrm{Var}(\widehat{\tau}(x)) = v_0 + v_1$$
- 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

Choice of priors

- BART is a special case of the Bayesian nonparametric model. There are others, e.g. Gaussian Process (GP), Dirichlet Process mixture (DPM)
- Which one to choose? Depends on the degree of overlap
- A desirable prior should accurately reflect uncertainty for various degree of overlap
- Simulation evidence:
  - In regions with good overlap, all methods perform similarly
  - In regions with poor overlap, choose a robust prior that adaptively captures uncertainty according to different degree of overlap.
  - With poor overlap, BART appears to struggle whereas GP and DPM do better in reflecting uncertainty

Choice of Priors: a simulated example

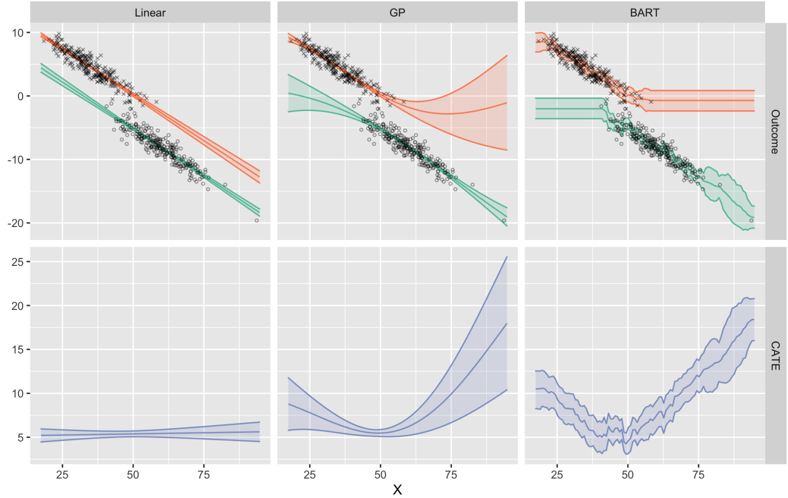
- (An example first due to Surya Tokdar, details in Li et al. (2022) review paper)
- A study with 250 treated and 250 control units
- A single covariate X following Gamma distribution with mean 60 in the control and 35 in the treatment group, and with SD 8 in both groups.
- A true outcome model with constant treatment effects:

$$Y_i(z) = 10 + 5z - 0.3X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

- here the CATE $\tau(x) = 5$ for all x. Covariate overlap is good

  between the groups in the middle of the range of X (around 40 to 50), but deteriorates towards the tails of X.

Choice of Priors: a simulated example



Regularization Induced Confounding

Hahn et al.(2018); Linero (2021)

- Define selection bias: $\Delta(z) = E[Y_i \mid Z_i = z] - E[Y_i(z)]$
- Prior dogmatism (Linero, 2021): With a naive Bayesian shrinkage prior for the outcome and PS model, the prior dist of $\Delta(z)$ concentrates sharply around 0 as p increases

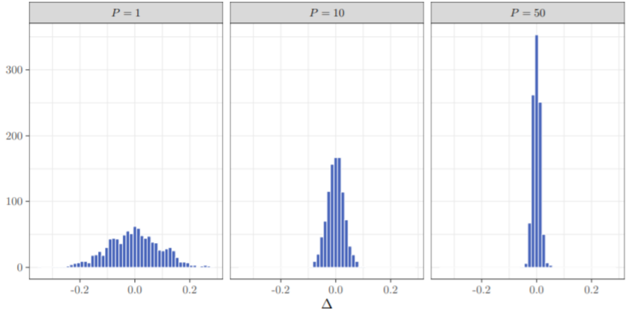


Figure: Figure 3 in Linero and Antonelli (2022)

Insights from Frequentist Literature

- A popular recent trend is to use machine learning (ML) methods for causal inference
- Important insights from related frequentist literature
  - 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
  - (Post-)Double selection of confounders (Belloni et al., 2014, RoES): Specify ML models for both propensity score and outcome models; combine in a doubly-robust estimator
  - Double Machine Learning (Chenozhukov et al. 2018): "Double/di-biased/Neyman orthogonalized" ML and sample splitting;
- Central messages:
  - Causal inference is NOT prediction
  - important to have both PS and outcome models (echo the Bayesian literature)
- An interesting departure from Bayesian paradigm: the role of sample splitting. No need in Bayesian, but crucial in Frequestist

Bayesian Causal Inference: Summary

- "*Any complication that creates problems for one form of inference creates problems for all forms of inference, just in different ways*" – Donald Rubin (2014)
- Bayesian + causal inference: anything special?
  - (Paradoxical) role of propensity score
  - In high-dimensional settings: regularization induced confounding (analogue of the Robins-Ritov problem)
  - Lack of overlap: sensitive to choice of priors and the outcome model
  - Identifibility is no longer all-or-none, a continuum between weak to strong identification

Why (and When) Bayesian?

- Usual arguments: uncertainty quantification, not rely on large sample asymptotics
- Specific to causal inference:
  - Allow easy inference of individual causal effects
  - Combine with decision theory
  - Particularly suitable for complex settings: post-treatment confounding (principal stratification), sequential treatments, spatial and temporal data
  - Advanced Bayesian models and methods bring new tools: Bayesian nonparametrics, Bayesian model selection, Bayesian model averaging

In Rubin's words: On Bayesian causal inference

A Conversation with Donald B. Rubin (p. 448), Stat. Sci. 2014

Fan*: In the 1978 Annals paper (Rubin, 1978a), you gave, for the first time, a rigorous formulation of Bayesian inference for causal effects. But the Bayesian approach to causal inference did not have much following until very recently, and the field of causal inference is still largely frequentist. How do you view the role of Bayesian approach in causal inference?* **Don***: I believe being Bayesian is the right way to approach things, because*

the basic frequentist approach, such as the Fisherian tests and Neyman's unbiased estimates and confidence intervals, usually does not work in complicated problems with many nuisance unknowns. So you have to go Bayesian to create procedures. You can go partially Bayesian using things like posterior predictive checks, where you put down a null that you may discover evidence against, or direct likelihood approaches as in Frumento et al. (2012); if the data are consistent with a null that is interesting, you live with it. But Neyman-style frequentist evaluations of Bayesian procedures are still relevant.

Fan*: But why is the field of causal inference still predominantly frequentist?* **Don***: I think there are several*

reasons. First, there are many Bayesian statisticians who are far more interested in MCMC algebra and algorithms, and do not get into the science. Second, I regard the method of moments (MOM) frequentist approach as pedagogically easier for motivating and revealing sources of information. Take the simple instrumental variable setting with one-sided noncompliance. Here, it is very easy to look at the simple MOM estimate to see where information comes from. With Bayesian methods, the answer is, in some sense, just there in front of you. But when you ask where the information comes from, you have to start with any value, and iterate using conditional expectations, or draws from the current joint distributions. You have to have far more sophisticated mathematical thinking to understand fully Bayesian ideas. There are these problems with missing data (as in my discussion of Efron, 1994) where there are unique, consistent estimates of some parameters using MOM, but for which the joint MLE is on the boundary. So I think it is often easier, pedagogically, to motivate simple estimators and simple procedures, and not try to be efficient when you convey ideas. In causal inference, that corresponds to talking about unbiased or nearly unbiased estimates of causal estimands, as in Rubin (1977). There are other reasons having to do with the current education of most statisticians.

Final Words

- Both design and analysis stage are central to causal inference
- A full Bayesian causal model, by definition, only involves the analysis stage
- Proper Bayesian causal inference must take into account design (i.e. assignment mechanism or propensity score): in either the design or the analysis stage
- The ultimate goal in causal inference is to estimate causal effects; choice of inferential mode is case-dependent
- For causal inference (or anything), being Bayesian should be a tool, not a goal

References

Schnell, Patrick M., et al. "A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects." Biometrics 72.4 (2016): 1026-1036.

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).

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

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.