Capstone: Benchmarking Causal Inference Algorithms

Josh Broomberg

February 23, 2020

Contents

Part I

Introduction

Chapter 1

New Approaches, New Challenges

The last decade has seen enormous progress in the development of advanced algorithms that extract useful information from rapidly expanding stores of data Lecun, Bengio, and Hinton, 2015. These algorithms - referred to as methods of machine learning - broadly operate in the paradigm of predictive inference: implicitly or explicitly learning a joint probability distribution over independent and identically distributed (IID) data and using the implied correlations to make conditional predictions Schölkopf, 2019. As a result of this reliance on statistical correlation to infer the relationship between variables, these methods are useful but inherently limited. As is pointed out in Pearl, 2009, the accuracy of predictive methods fail when, rather than observing and predicting IID from outside of a static data generating process, the target of prediction is an (intended) intervention that changes the data generating mechanism and, as a result, the observed data distribution. Here is a simple example of this failure: In an observed data set, the presence of rain and a wet roof are correlated. Knowing the roof is wet will reduce one's uncertainty about the weather. But, understanding that actively wetting the roof will not result in changes to the weather - despite the high *predictive* power of a wet roof for the presence of rain - requires understanding that rain *causes* the wet roof and not the reverse. This pattern holds in less contrived settings. If a new medicine tends to be prescribed to healthier (less at risk) patients during a trial period, then there will be a strong correlation between the new medicine and quick recovery regardless of the efficacy of the medicine. Given this correlation, predictive methods would (correctly) infer that prescribing the medicine increases the probability of recovery (*within the IID observed data)*. But, upon intervening by giving the medicine to less healthy patients, one would find no positive effect - the predicted recover was the result of the previous treatment policy, not the medicine itself. These help establish an important truth: predicting the result of an intervention that changes some part of an observed system requires information beyond pure statistical correlation. I will show below that correlation does, in fact, play an important role in inferring the outcome of intervention but must be augmented.

The challenge presented by this limitation is that intervention, of one kind or another, lies at the heart of many of the most important questions we would like to answer with growing stores of data. Social, economic and health policy is designed to *change* systems to produce better outcomes for members of society. Companies take actions that *change* market conditions to produce commercially-desirable outcomes. The intention to intervene means that predictive inference, applied to observed examples of different policies/actions, cannot be used to guide choices of which policy/action is best. Rather, what is required is methods of *causal inference.* These methods can be thought of as uncovering correlations at the fundamental, mechanistic level rather than the correlations which appear at the (arbitrary) phenomenological level. These fundamental correlations persist under interventions that change the observed setting and, therefore, allow for the evaluation of the *true,* causal effect of that intervention. Schölkopf, 2019.

Historically, randomized experiments have been the gold standard in uncovering these fundamental, mechanistic correlations Meldrum, 2000. Randomizing the 'intervention' to which some unit is exposed, means that any observed (average) difference in outcome can only and will only reflect the effect of differential exposure to interventions and no other underlying cause. The problem with this approach is that running an experiment is often impossible (for nation-scale policy changes), expensive, slow, and quite possibly unethical (in cases where there is a reasonable expectation the intervention will mitigate suffering but to an unknown extent) Meldrum, 2000. Even if experimentation is viable, the validity of the effect recovered is limited to the sub-population which is targeted for the experiment which, in many cases, will not be representative of the broader population and, further, is only correct on average but not for any single individual Rothwell, 2006. For these reasons, there is great value in methods that are able to infer fundamental, mechanistic effects from non-experimental, non-randomized, *observational* data at both the average and individual level. Methods that are capable of this inference are able to guide policy/action decisions based on the reservoirs of data already available to governments and corporations without the need for experimentation.

The last few decades have seen the rapid development of just such methods of *observational causal inference.* As a general rule, these methods build on existing tools of predictive inference but - through a combination of weak assumptions and modified estimands - recover fundamental mechanistic relations. These methods can be divided into two classes based on the nature of the targeted estimand. First, there is a mature literature on the estimation of *average causal effects.* Methods in this class estimate the average effect of an intervention by undoing the *confounding bias* in (average) group outcomes induced by factors that affect both the outcome of some intervention and the likelihood of receiving it (as in the medical treatment example above). This estimand is useful for government policy evaluation which focuses on the macro-efficacy of some intervention. Common approaches include inverse probability weighting (Horvitz and Thompson, 1952; Hirano, Imbens, and Ridder, 2003) and matching estimators (Rubin, 1974;Rosenbaum and Rubin, 1983). See Imbens and Wooldridge, 2009 and Athey and Imbens, 2017 for extensive reviews of the numerous methods which target average causal effects. The second class of methods is comprised of a newer body of work focused on estimating *individual causal effects* - the effect of an intervention on specific individuals with the possibility of heterogeneous effects across the population of individuals. Estimation of individual effects requires disambiguating the confounding bias, as defined above, which operates at the level of intervention groups, from the heterogeneity of individuals' responses to intervention at the sub-group level. The individual effect estimate is important where there is known heterogeneity in the response to intervention and the impact on individuals is relevant to policy decisions. For example, medical treatment often involves choices between hundreds of available options with highly varied individual responses and, as such, requires an understanding of how each treatment will affect a specific individual (Lu, Sadiq, Feaster, and Ishwaran, 2018). As a result of the inferential challenge implied by both confounding bias and heterogeneous response, methods for individual effect estimation tend to be newer and use modern, flexible semi/nonparametric estimators that require substantial computational resources. These include: regression trees (Su, Edu, Wang, Nickerson, and Li, 2009; Athey and Imbens, 2016), Random Forests (Wager and Athey, 2018; Athey, Tibshirani, and Wager, 2019), Least Absolute Shrinkage and Selection Operator (Lasso) (Qian and Murphy, 2011; Tian, Alizadeh, Gentles, and Tibshirani, 2014; Chen, Tian, Cai, and Yu, 2017), Neural Networks (Fredrik D Johansson, Shalit, and Sontag, 2016; Fredrik D. Johansson, Kallus, Shalit, and Sontag, 2018; Schwab, Linhardt, and Karlen, 2018, Li and Fu, 2017, Künzel et al., 2018) or Bayesian machine learning (Hill, 2011; Taddy, Gardner, Chen, and Draper, 2016).

The diversity in methods, both for average and individual effect estimation, is to be expected. Causal inference fundamentally rests on the same inferential mechanics as predictive inference: accurately estimating distributions, or expectations over those distributions, from observed data. This presence of predictive inference is clear in both of the dominant frameworks for causal inference: the Rubin-Neyman Potential Outcomes Framework Holland, 1986 and Pearl's Structural Causal Modelling (SCM) Pearl, 2009. Causal inference is, however, distinguished from purely predictive inference by the addition of mechanisms and assumptions to counteract the confounding bias induced by the shift in observational setting created by a (potential) intervention. These mechanisms and assumptions are simply combined with the more fundamental inference tasks rather than replacing them. This framing of causal inference methods, which will be formalized in the chapters below, provides an intuitive justification for the existence of many equivalent estimators for the average and individual effect estimands. There are two distinct sources of equivalency evident in the framing above. First, for any given mechanism (and set of assumptions) for reversing observational bias, there tend to be many equivalent predictive inference methods that can then be used to operationalize the estimation of the target effect. Second, for any given data set, there are usually multiple equivalent mechanisms (and requisite assumptions) that can be used to mitigate confounding bias. Much like in pure predictive inference, different, equivalent inference techniques and bias-mitigation mechanisms may exhibit very different performance and properties in finite samples while all converging to the same estimands in the asymptotic limit.

This raises the challenge that is the core focus of this paper. Given numerous equivalent methods for observational causal inference, how does one know which one is best? This question is crucially important - both for researchers who are working to improve these methods and for practitioners who wish to apply the best method available to make important policy decisions in their fields. It is also a question that is hard to answer. Unlike in predictive inference, the ground-truth is not present in the training data. Under the Potential Outcomes framework, this stems from the fact that we never see the same unit under different interventions and thus can never know the true causal effect of the intervention on that unit. Equivalently, in Pearl's SCMs, the observed data is distributed differently to the data under intervention. In either case, the result is the same - the training data does not contain the ground-truth. Further, while some methods may lend themselves to formal, asymptotic convergence proofs, these are not universally available, rely on assumptions which often render them mutually incomparable between methods and, importantly, do not apply to performance in the finite data regime which is of the most relevance for method selection Knaus et al., 2018. This situation is further complicated by a large space of different causal inference challenges corresponding to different distributional settings in the observed data. It will be demonstrated that inference methods are likely to perform quite differently on finite samples corresponding to different locations in the *problem space* of different distributional settings. This implies there may not be one method of causal inference which is universally superior across all finite data samples. So, in short, answering the question of which causal inference algorithm is best is hard because it requires finite sample evaluation, without access to ground-truth, and with the challenge of heterogeneous performance under different distributional settings. This paper tackles this challenge by proposing an evaluation method that allows for finite sample evaluation of arbitrary causal inference methods across an arbitrary selection of distributional settings.

Researchers have typically addressed the evaluation problem outlined above through two complementary evaluation methods, both of which provide access to a ground-truth causal effect with some trade-off in evaluative efficacy. Empirical evaluation methods use (randomized) experimental datasets and test an estimator's ability to reproduce the experimental result when using a non-randomized control group (to simulate an observational setting). This strategy was pioneered by Lalonde, 1986 with iterative development by Heckman and Todd, 1998, R. H. Dehejia and Wahba, 1999, R. H. Dehejia and Wahba, 2002, R. Dehejia, 2005, and Smith and Todd, 2005 with notable contributions from Hill, Reiter, and Zanutto, 2005 and Shadish, Clark, and Steiner, 2008. The data used in these evaluations is realistic - in so far as it is drawn from data collected as part of formal, real-world studies - but only a small number of such experimental datasets (with appropriate observational controls) are available and these cover a small and poorly defined subset of the problem space[1]. In contrast, the second approach, synthetic evaluation, relies on synthetic data generated by a hand-crafted data generating process (DGP). These DGPs can simulate any location in the distributional problem space but are often unrealistic in terms of the number and type of variables used as well as the functional forms used to simulate the outcome and intervention data. The mechanisms behind these two approaches - and the strengths and weaknesses that these mechanisms imply - are formalized in the chapters below.

The primary contribution of this paper is a hybrid evaluation method designed to overcome the weaknesses of both of the evaluation methods above. The method proposed is based on the *sampling* of synthetic DGPs defined over real, observational data. This hybrid approach combines the diagnostic clarity of synthetic evaluation methods with the realistic distributions of empirical evaluation methods. Appropriate parameterization of the sampling process allows for the generation of DGPs which are in well-defined, specific locations in problem space without the problematic implications of hand-crafted, 'targeted' (but potentially biased) design. Additionally, the sampling approach also allows for the generation of many distinct instances of evaluative datasets in the same problem class, allowing for repeat evaluation of a method and, thus, for convergence to an accurate distribution of performance - revealing average/best/worst case results rather than just single-point samples from an unknown distribution.

The hybrid approach proposed is not entirely novel. A number of other authors have proposed similar methods albeit with small but significant idiosyncrasies. These methods are reviewed in **chapter 4** below. A common shortcoming of all the existing work in this subfield is the lack of tooling to make the proposed evaluation method available to others such that it can be applied to methods not covered in the original papers and used with base data and problem-instance settings relevant to different academic fields of study. Existing code, if available, is tightly coupled to specific empirical source data and inflexibly/confusingly parameterized. With this context in mind, this paper

[1] The true data generating process in these datasets is unknown which means it is unclear what distribution properties the data displays.

makes two contributions. First, a synthesis of the hybrid approaches proposed in the literature with the goal of proposing and justifying a single, strong hybrid evaluation method. And, second, an accompanying tool that makes it easy to apply hybrid evaluation to arbitrary causal inference methods using arbitrary base data and arbitrary problem-instance settings. The goal of this tool is to provide a consistent, universal means by which causal inference researchers - and users - can develop, compare and select methods of causal inference.

The rest of this paper proceeds as follows [2]: Chapter 2 introduces causal inference and formalizes the notation and terminology used throughout this work. Chapter 3 establishes the causal inference problem space - the space of distributional settings that impact the performance of different causal inference estimators. This is the space over which estimators should be benchmarked. Chapter 4 presents a review of the literature on causal estimator benchmarking, comparing asymptotic and Monte Carlo based evaluation methods. Chapter 5 presents the design of a benchmarking method which synthesizes the best of the methods reviewed in Chapter 4 and is capable of evaluating estimators across the problem space from chapter 3. Chapter 6 introduces CausalBench, a tool which implements the design from Chapter 5. Chapter 7 concludes.

[2] This chapter breakdown reflects the chapters in this draft rather than the full break down of the final paper.

Part II

Causal Inference Theory

Chapter 2

A Framework for Causal Inference

This chapter provides a theoretical overview of observational causal inference. This serves as a foundation for the rest of the work, introducing the notation and framing which are used throughout. The framing of causal inference introduced in this chapter is primarily based on Rubin's Potential Outcome Framework as outlined in Holland (1986) with augmentation from Pearl's Structural Causal Models as outlined in Pearl (2009).

Per Holland, observational causal inference applies statistical methods to measure the "effects of causes" in non-experimental settings. IE, methods of observational causal inference seek to estimate the quantitative effect of some *treatment* on *units* with the potential - realized or not - to be exposed to that treatment without the use of randomized exposure[1]. Note that if the unit has no potential to be exposed to the treatment or is always exposed, then understanding the effect of the treatment on that unit is philosophically impossible: the effect of a cause is defined based on the observed outcome relative to some (hypothetical) *counterfactual* in which the cause did not occur. These ideas can be formalized by expressing them through a quantitative, statistical lens as follows.

2.1 Notation and Estimands

Let there be a population of units - U - with individual units from U indexed as u_i . Let $Z\left(u_i\right) = Z_i$ be an indicator variable that tracks whether a unit was exposed to the treatment [2]. Based on the binary nature of the treatment, each unit has two *potential outcomes* - $Y_1\left(u\right)$ and $Y_0\left(u\right)$ - which measure the outcome the unit experiences under treatment or absence of treatment (referred to as the *control* condition, following experimental nomenclature). Further, each unit has a (potentially vector-valued) covariate measurement X_i . The name covariate is used to imply that the variable X may co-vary (or be *correlated*) with the outcomes, Y_1 and Y_0 , and the treatment status Z across the population U . In the real world, a correlation between these variables may

[1] The term treatment is adopted from experimental literature for clarity.

[2] This implies a binary treatment regime in which treatment is either present or absent for each unit but the framework present in the text can be extended to multiple discrete or single continuous treatments.

result from the covariates in X causally affecting the outcome/treatment assignment or the inverse. For simplicity, I assume for the rest of this section that the covariates are measured pre-treatment and, thus, are not causally affected by a unit's treatment assignment or outcome(s) [3]. There are two types of covariates depending on the underlying causal mechanism: *confounders* are causally related to both the treatment and outcome of units in the population while *sources of heterogeneity* are related to either, but not both, the treatment or outcome. The reason for this naming will become clear shortly.

Given the above, the effect of a cause on an individual can be defined as $\tau_i = Y_1\left(u_i\right) - Y_0\left(u_i\right)$. *The Fundamental Problem of Causal Inference,* per Holland, 1986, is that "it is impossible to observe the value of $Y_1\left(u_i\right)$ and $Y_0\left(u_i\right)$ for the same unit and, therefore, it is impossible to observe the effect of t on u" . For any given unit, we only observe a single outcome defined by $Y\left(u\right) = Z_i \times Y_1\left(u\right) + \left(1 - Z_i\right) \times Y_0\left(u\right)$ - the outcome under the treatment which the unit experienced. At this point, it may appear that the idea of causal inference is hopeless. However, introducing statistical tools (more specifically expectations), provides a path forward. The first step is to define average effect estimands in terms of observed and unobserved potential outcomes. These causal inference *estimands* can then be targeted by causal inference *estimators* defined over only observed data. Define the *population average treatment effect* (PATE) as:

$$PATE = E\left[\tau\right] = E\left[Y_1 - Y_0\right] = E\left[Y_1\right] - E\left[Y_0\right]$$

This is the most basic causal estimand. It is possible to construct analogous expressions for samples rather than populations and for the treated/control subgroups of the population/sample. However, for the purposes of this paper, the only other important estimator is the *conditional average treatment effect* (CATE) which refers to the causal effect conditioned on a covariate observation X_i :

$$CATE = E\left[Y_1 - Y_0|X_i\right] = E\left[Y_1|X_i\right] - E\left[Y_0|X_i\right]$$

Note that the PATE can be recovered by averaging the CATE over all units in the population based on the formula below. This allows us to focus our attention on this estimand.

$$PATE = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} E\left[Y_1 - Y_0|X_i\right]$$

Given than two units with an identical observed X_i are effectively indistinguishable in experimental terms, the CATE is often referred to as the *individual average treatment effect* [4].

[3] Resolving the truth of this assumption is crucially important in practice - see Pearl, 2009 for the risks of using covariates which are causally affected by treatment/outcome (even if the measurement is pre-treatment). However, this consideration is not important for establishing the fundamental operation of causal inference.

[4] The word average is used to indicate that units with identical X_i values may, in fact, have different treatment effects in which case the CATE is an average over these effects for units indistinguishable on the observed covariates.

2.2 The Challenge of Accurate Causal Inference

The estimands above imply that if we can accurately approximate the expectation of the two potential outcomes across the population, or a covariate level, then we can approximate the average treatment effects. IE, it is possible to estimate the unobservable counterfactual at the individual level using expectations for the potential outcomes defined over the population of units. However, estimating these expectations is challenging: In any experimental or observational setting - rather than observing $E\left[Y_1\right]$ and $E\left[Y_0\right]$, we observe $E\left[Y_1|Z=1\right]$ and $E\left[Y_0|Z=0\right]$. This expresses the idea that we observe the outcome conditioned on treatment assignment. There is no guarantee that $E\left[Y_1\right]=E\left[Y_0|Z=1\right]$ and, in fact, this is highly unlikely in observation settings. This is where the notion of selection bias become important.

Following Holland, let us call the average treatment effect arrived at from the observed outcomes the treatment effect prima facie:

$$T_{pf}=E\left[Y_1|Z=1\right]-E\left[Y_0|Z=0\right]$$

Observe that if the treatment assignment is independent of the potential outcomes - $P\left(Z|Y_1,\ Y_0\right)=P\left(Z\right)$ - then $E\left[Y_1|Z=1\right]$ does indeed equal $E\left[Y_1\right]$ and the true average treatment effect is recovered by the treated/control group estimates. This explains the power of randomized control trials. If treatment/control is assignment randomly then the condition above holds and the group outcome averages allow recovery of the true effect. In the absence of randomization, it is possible for the treat/control group average outcomes to be ‘biased’ by the systematic inclusion of units which have a better/worse potential outcome under treatment. This bias can be neatly formalized. Assume a constant treatment effect T . Then for the treated group we have: $E\left[Y_1|Z=1\right]=E\left[Y_0|Z=1\right]+T$. Substituting this into the formula for T_{pf} , we arrive at:

$$T_{pf}=T+\left(E\left[Y_0|Z=1\right]-E\left[Y_0|Z=0\right]\right)$$

The prima facie effect is the true effect plus the expected difference in the outcome under control for the treated and control groups. In the absence of randomization, systemic differences between the treatment and control group induce bias. The challenge of observation causal inference is how to undo the bias induced by these differences. Before proceeding to strategies for achieving this, it is important to explore the mechanisms which create systemic difference in the first place. To do this, I draw on Structural Causal Models (SCMs) introduced by Pearl and summarized in Pearl, 2009. SCMs are useful because they provide an intuitive way to communicate the causal mechanisms which give rise to the bias discussed above. This understanding, in turn, makes it easier to understand the construction of bias-free causal inference estimators.

Pearl's SCMs relate the variables defined above - outcome, treatment and covariates - through a directed acyclic graph. The nodes in this graph are the variables and edges between nodes indicate causal dependency between the child node variable (the target of the directed edge) and its parent(s) (the origin of the edge(s)). More precisely, the value of each variable node is defined in terms of a function that depends solely on the values of the parents of the node in the graph. This implies that knowing the value of all the parent nodes of a child node fully determines the value of the child node[5]. As

[5] In a complete SCM each node has a parent which presents a source of uncorrelated noise which

a result of this causal dependence, variables will be correlated, to some degree, with their parents in the graph, to the ancestors of those parents, and to any other nodes which share common parents/ancestors. With this definition established, I present the simple SCM in Figure 1 below to explain the origin of selection bias in causal effect estimation.

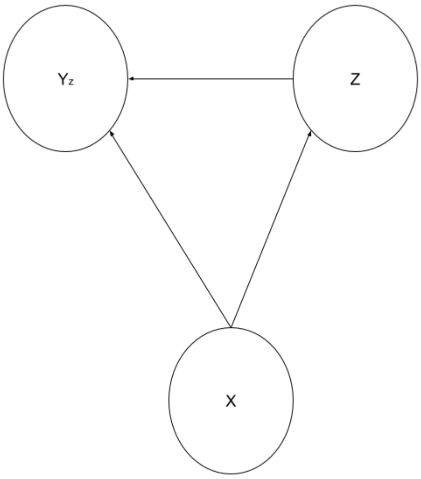


Figure 1: SCM displaying causal relations between the outcome, treatment status and covariates which gives rise to selection bias.

This SCM displays the simplest possible situation in which selection bias is a problem. The edge between Z and Y_z is the primary effect of interest. The value of Y_z will increase by T if $Z = 1$ relative to its value if $Z = 0$. Further, both Y_z and Z are causally related to the covariate X. IE, the treatment status and the outcome are determined by a function which depends only on X and noise. As a result of this causal mechanism, it is easy to see that $P\left(Z|Y_1,\, Y_0\right) \neq P\left(Z\right)$. So, the treatment status is not independent of the outcome as a result of the shared dependence on the value of X and the two groups will have systematically different base outcomes. This simple example provides insight into a more general pattern: where treatment assignment depends on the same covariates as the outcome, there will be a systematic difference in outcome between the groups. We say that the covariates which affect both treatment assignment and outcome confound the estimation of causal effects - hence the name confounders. Beyond clarifying the mechanism behind selection bias, the model in Figure 1 also hints at a viable path for counteracting the bias. If one conditions on X, then for fixed values of X, the only correlation left between Y_z and Z is due to the edge between them. If we treat X_i as an observation of a full set of covariates which full identify a unit, this process is intuitively clear. By conditioning on the covariates, the correlation between Y_z and Z is estimated on subpopulations of identical units. Any change in outcome in these subpopulations must occur as a result of the change in the treatment status. But note that this only holds if we observe all of the confounding covariates related to both the treatment assignment and outcome.

The process outlined here, and the assumptions required to operationalize it on real

explains the absence of perfect predictive power of the observed parents. For simplicity, this noise node is omitted from the models presented in the text. This implies that all the variables are fully determined by their parents.

data, are outlined in the next section.

2.3 Causal Inference as Predictive Inference under Assumptions

The ideas expressed above can be formalized in the Potential Outcome framework. The only nuance is that, under this framework, the outcomes under treatment and control are fixed and therefore there is no directed edge, in the language of SCMs, between treatment assignment and outcome. Despite this small difference, statistical dependency between the potential outcomes and treatment assignment is induced by confounders exactly as above.

Returning to the formalism of the estimands defined above, recall that finding the PATE requires estimating $E\left[Y_1\right]$ and $E\left[Y_0\right]$ but that, in an observational setting, one only has access to $E\left[Y|Z=1\right]$ and $E\left[Y|Z=0\right]$. If one assumes that the covariate vector X_i contains all confounders we have the following:

$$E\left[Y_1\right] = E_x\ \left[E_{Y_1}\left[Y_1|X\right]\right] = E_x\left[E_{Y_1}\left[Y_1|X, Z=1\right]\right] = E_x\left[E_{Y_1}\left[Y|X, Z=1\right]\right]$$
$$E\left[Y_0\right] = E_x\ \left[E_{Y_0}\left[Y_0|X\right]\right] = E_x\left[E_{Y_0}\left[Y_0|X, Z=0\right]\right] = E_x\left[E_{Y_0}\left[Y|X, Z=0\right]\right]$$

This result implies that:

$P\left(Y_1, Y_0|X, Z\right) = P\left(Y_1, Y_0|X\right) =$

This is the statement that the potential outcomes are conditionally independent of the treatment status given the covariates:
$Y_1,\ Y_0\ \perp\!\!\!\perp\ Z|X$
And this is exactly what was demonstrated to be true using the SCM above under the assumption that X does indeed include all confounders. Note that, under this framing, the problem of observational causal inference is reduced to a standard conditional estimation task $E\left[Y_z|X, Z=z\right]$ for the two treatment groups ($z\ \in \{0,1\}$). This is discussed in detail shortly.

Formally, the Rubin Causal Model makes 4 assumptions to recover an unbiased estimate of causal effects:

- **Assumption 1 - Stable Unit Treatment Value Assumption:**

  $Y\left(u\right) = Z_i \times Y_1\left(u\right) + \left(1 - Z_i\right) \times Y_0\left(u\right)$. IE, each unit has a fixed outcome under treatment and control and there is no interaction between the outcomes of different units.

- **Assumption 2 - Exogeneity of Covariates:**

  $X_i\ |Z_i = 1\ \ = X_i|\ Z_i = 0$. IE, covariates are independent of the treatment assignment.

- **Assumption 3 - Conditional Independence:**

$Y_1,\ Y_0 \perp\!\!\!\perp Z|X$. IE, the covariate observation X contains all confounders and there is conditional independence between outcomes and treatment assignment conditional on X .

- **Assumption 4 - Common Support:**

  $0 < P(Z = 1|X = x) < 1$ for all $x \in \textit{support}(X)$. IE, there is some probability that all possible units are exposed to both treatment and control. Philosophically, this is required for the existence of Y_1 and Y_0 because in the absence of the potential for exposure it is unclear what the (impossible) potential outcome would mean. Practically, this ensures that the outcome estimands in each group have common support and can be combined and averaged without inducing bias. This assumption can be relaxed but this must be done with extreme care.

Under these assumptions, it is possible to estimate $E[Y_z|X]$ using $E[Y|X, Z = z]$. So, the estimands defined above become:

$$CATE(X = x) = E[Y|X = x, Z = 1] - E[Y|X = x, Z = 0]$$

$$PATE = \sum_x P(X = x) \times [E[Y|X = x, Z = 1] - E[Y|X = x, Z = 0]]$$

These expressions make it clear that, under certain assumptions, observational causal inference reduces to estimation tasks which are typically *predictive* in that they rely only on standard statistical correlation. In the framing above, the first step of causal inference is to estimate the quantity $E[Y|X, Z]$ - the *response surface* in the words of Hill, 2011 - given the observed data from the two treatment groups. This is a standard predictive estimand. Under causal assumptions, the resultant estimated response surface(s) can then be combined to find average/conditional treatment effects.

While this two step process does clarify the connection between causal and predictive inference, it is a severe oversimplification. Different inference methods can be used to operationalize the 'standard' inference of the response surface - ranging from simple means across covariates blocks to fully-fledged function fitting. Moreover, one can define equally unbiased expressions for the estimands above which depend exclusively on 'standard' inference but have meaningfully different mathematical forms. The different estimators implied by these two complications may be more or less efficient - producing 'better' results on the same dataset - and they may exhibit different performance under different distributional settings. Understanding these differences is crucial - the goal of research into causal inference estimators is to produce methods which are more efficient and work in specific, or across many, distributional settings. The next chapter explores different causal inference estimators and how their performance is affected by (a formalized notion of) the distributional setting of observed data. The formalization of distributional setting is particularly important as this is the constant background against which new estimators must be validated.

Chapter 3

The Causal Inference Problem Space

The previous chapter ended with the claim that many causal estimators can be constructed to target the same causal estimands and that these estimators may display varied performance under different *distributional settings.* The primary goal of this chapter is to formalize the notion of a *distributional setting* and motivate why the *distributional setting*, as defined, affects estimator performance. In so doing, I will introduce the notion of a causal inference problem space which is the space of all performance-relevant distributional settings.

I will proceed in three steps. First, I will provide a brief overview of the landscape of causal inference estimators. While this overview will include instructive examples of different approaches, it is not meant as a complete (or even partial) review of different methods. Rather, the analysis will provide initial insight into the importance of the properties of the underlying data distribution for the performance of different estimators. The discussion of distributional properties will remain informal at this stage. Second, I will formalize the notion of a distributional setting by expressing causal problems in terms of a joint distribution over observed data. In this context, a distributional setting can then be defined as a joint distribution over the observed data with some set of properties that affect estimator performance. Finally, I will establish the axes of an abstract space over distributional settings with each axis representing different possible values of some property of the joint distribution. This space is the Causal Inference Problem Space.

3.1 The Sensitivity of Causal Estimators to Properties of the Observed Data

In the previous chapter, I arrived at an expression for the average causal effect estimands in terms of $E\left[Y|X,Z\right]$ and referenced this as the 'response surface'. I also asserted that it was possible to find expressions for the same estimands using meaningfully different, but purely statistical/predictive, estimators. An intuitive explanation for this statement is that causal processes are made up of different *causal mechanisms* that relate the various variables[1]. And, in order to recover the unbiased effect, it is

[1] Mathematically, these mechanisms take the form of functions which take as inputs some subset of the variables for some unit and output the value of some other variable for that unit. These functions

sufficient to accurately model only one of these mechanisms. Hill (2011) echos this explanation and refers to two mechanisms: the *treatment assignment mechanism* - which relates the covariates X to the treatment status Z - and the *response mechanism* - which relates the covariates X and the treatment status Z to the observed outcome Y . The authors point out that an accurate model of either mechanism is sufficient to recover average treatment effects. Künzel, Sekhon, Bickel, and Yu, 2019 further divide the *response* mechanism into an *outcome mechanism* which relates the covariates X to the outcome without treatment Y_0 and the *treatment effect mechanism* which relates X and Z to τ which, in this framing, is the change in outcome (from Y_0 to Y_1) as a result of treatment [2]. Hypothetically, accurate specification of any of the now three mechanisms is sufficient to recover causal effects. But, given that it is impossible to directly model the treatment effect mechanism without a model of one of the other two mechanisms, it is only useful in so far as it clarifies a potential source of heterogeneity in the combined (outcome and treatment effect) response mechanism as defined by Hill, 2011. With this idea of causal mechanisms in mind, different causal estimators can be understood as arising from the targeting of either, or both, the treatment assignment and response mechanisms. The targeted mechanism also explains the sensitivity of different estimators to different properties of the observed data distribution. With this established, I proceed to the brief overview of different estimators in order to make the points above more tangible.

Looking first at methods which target the treatment assignment mechanism: Matching (Rosenbaum and Rubin, 1983; Abadie and Imbens, 2006) aims to mitigate selection bias by creating treatment and control groups with the same distribution of covariates, reversing the effect of the treatment assignment mechanism. This is done by pairing units with similar covariate values but different treatment exposure to create two groups with similar overall distributions[3] [4] [5]. This process is non-parametric relative to the treatment assignment mechanism because it targets the result of the mechanism without attempting to directly model it[6]. This means it is not sensitive to the exact functional form of the mechanism. However, it is still sensitive to other aspects of the observed data. Broomberg, 2017 describes the potential sensitivity as follows. First, as the number of covariates grows, the probability of finding similar units, for any fixed notion of similarity, shrinks exponentially, thus requiring exponentially larger observed donor groups to achieve balance. This holds even in the absence of strong selection pressure producing different covariate distributions between the groups. Second, if there is strong selection pressure, then it is likely that the distribution of covariates will be meaningfully different between the groups. The larger the imbalance, the less

represent some real generative process which relates the measured quantities in the real world. In order to represent real stochastic processes the functions may themselves be stochastic and take the form of distributions parameterized by the input variables.

[2] The covariates appear in the treatment effect mechanism because the treatment effect may be non-homogeneous/non-constant in which case it will depend on the covariate values of the exposed unit.

[3] The reality of the pairing is slightly more nuanced. The standard case is that observed units from a treated group are matched with control units from a 'donor pool'. This produces an estimate of the average effect for the treated.

[4] Defining a metric in covariate space is a complex task, so there is no single definition of 'similar' units. A naive approach is simply to measure the Euclidian distance between covariate vectors in real-number space. Ultimately, a successful metric can be anything which, when applied to large donor pools of observations, results in matches that produce equal covariate distributions across the groups.

[5] Imai, King, and Stuart, 2008 show that if the distribution in the two groups is indeed the same (balanced), then a simple average over the observed outcome is an unbiased estimator of the average treatment effect.

[6] Matching methods may indeed have parameters which determine their operation but the correctness of these parameters is independent of the functional form of the treatment assignment mechanism.

likely it will be to find equivalent units in the opposite group and the larger the set of observed units that will be required to find matches. So, in short, matching is sensitive to the number of observed units, the number of covariates, and the functional form of the treatment assignment mechanism (in so far as it affects balance).

Propensity score based methods address some of the weaknesses above by explicitly modeling the treatment assignment mechanism and assigning each unit a probability (propensity) of treatment. Austin, 2011 summarizes the various ways in which the propensity score can be used to remove bias. The four classes of methods reviewed by the authors are matching/stratification on the propensity score, inverse probability of treatment weighting, and propensity-based covariate adjustment. These methods share at least two failure modes related to the observed data: one, Hill, 2011 points out that the common methods used for estimation of the treatment mechanism are parametric and therefore sensitive to the functional form of the underlying mechanism and the number of covariates. If this form is misspecified, or there are too many covariates, then the resultant propensity scores will not properly mitigate bias. Two, Knaus et al., 2018 points out that small propensity scores - which inevitably arise when there is an imbalance in covariate distribution across groups - can produce high variance (or even degenerate) estimates in weighting-based methods. So, again, there is sensitivity to the number of covariates, the functional form of the treatment assignment mechanism, and the covariate balance.

Turning to methods which estimate the response surface, a similar pattern of sensitivity is evident. The simplest methods in this class are referred to as conditional mean estimations by Knaus et al., 2018 or as T-learners by Künzel et al., 2019. These methods estimate the conditional mean $E\left[Y|X, Z=z\right]$ separately for both the treatment and control groups (a method which was motivated in the last chapter). Much like with direct modeling of the treatment assignment, this approach is sensitive to specification of the function form used for the estimation. This means there is sensitivity to the underlying functional form (and complexity) of the response mechanism. This sensitivity has inspired a huge number of flexible semi/nonparametric estimation schemes. See, for example, Causal Forests by Athey, Imbens, and Wager, 2018, BART Hill, 2011, and X-learner Künzel et al., 2019. But even these flexible estimation methods are sensitive to aspects of the data distribution. In the case of outcome function fitting, imbalance means that, in both groups, there are certain regions of the covariate support with fewer observations. This makes it hard to accurately fit functions in these regions. This is less problematic in parametric methods with fewer degrees of freedom but seriously affects non-parametric methods which rely on the data for accurate local estimation. Further, response mechanism based approaches only work if they correctly capture the response dependency on all confounders and the treatment assigning. This means that, much like in any function fitting task, there is sensitivity to the explainability of the response (the signal-to-noise ratio) and the number of, and redundancy between, predictive variables (the covariates). Incorrect inference of the response mechanism is possible in cases where there is a low signal-to-noise ratio and/or the presence of many covariates with similar but non-identical correlations with the prediction target

The final class of methods explicitly, or implicitly, combines the modeling of both the treatment and response surface, usually through weighted estimation targets. Examples which include explicit models of the two mechanisms are Rubin and Thomas, 2000 and Robins and Rotnitzky, 1995. Scharfstein, Rotnitzky, and Robins, 1999 show that if semiparametric estimation is used for modeling both the treatment assignment and response mechanisms, then the resultant estimators are valid if either the treatment assignment *or* the response mechanism are correctly modeled, leading to the name "Double Robust" estimators for this class of methods. Note that while this double robustness is appealing, the models of the two mechanisms are still sensitive to the same aspects of the data as outlined above. The combination simply improves the probability of valid results by hedging the accuracy of the one model with the other. More modern methods may implicitly combine models of both mechanisms rather than explicitly combining them. For example, Fredrik D Johansson et al., 2016 jointly optimize a neural network based estimator to produce balanced covariate representations and accurate outcome predictions using this representation. While demonstrating exact sensitivity is hard for complex estimators such as this, it is reasonable to assume that the same sensitivities outlined above will have some impact on accuracy. This is based on the idea that these sensitivities represent fundamental differences in the amount of information available to estimators and the complexity of extracting this information. No estimator, regardless of flexibility or complexity, is immune to the impact of decreasing predictive information.

The synthesis above is, by no means, a complete picture of all the causal estimators present in the literature. Rather, it serves to make a simple point: All estimators - regardless of the mechanism(s) they model and the parametric/non-parametric inference method used to perform the modeling - are sensitive to aspects of the observed data. From the examples above, it appears that the list below is a minimal subset of the performance-relevant aspects of the observed data:

- The number of observations
- The functional forms of the treatment and response mechanism
- The balance of the covariate distribution across groups
- The number of covariates
- The predictive power of the covariates for the treatment assignment/outcome (signal-to-noise ratio and overlap/correlation between predictors)

The *distributional setting* of a particular causal inference problem collectively refers to the value of all of these performance-relevant aspects of the observed data. The goal of the next section is to formalize the notion of a distribution setting as a set of properties of the joint distribution over the observed data.

3.2 Distributional Settings as Joint Distributions over Observed Data

The section above observed that different estimators target different underlying, component mechanisms and that this targetting produces sensitivity to different aspects of the observed data. This section formalizes this idea by showing that the two component mechanisms outlined above can be thought of as acting together to produce a joint distribution over the observed data, with the properties of this distribution affecting estimator performance. In this framing, a *distributional setting* is a joint distribution over the observed data with specific marginal/collective properties that are relevant to performance. Many joint distributions may have the same property values and therefore represent roughly the same distributional setting.

The first step in establishing this framing is to express the component mechanisms discussed above as probability distributions over the variables involved. This is the paradigm used in Pearl's Structural Causal Models (SCM) Pearl, 2009 which builds on the more general idea of representing generative processes as factorized probability distributions that relate causally-connected variables. Readers unfamiliar with these concepts should see Bishop, 2006 Chapter 8 for an introduction to graphical models and their interpretation as representing causal, generative processes. In this paradigm, each of the mechanisms above is represented as a conditional distribution:

- **The Treatment Assignment Mechanism:** $P(Z|X)$

- **The Response Mechanism:** in a classic SCM, this would be $P(Y|Z,X)$. For consistency with the Potential Outcomes Framework, I further factor $P(Y|Z,X)$ as $P(Y|Z,X) = P(Y_1|\text{Z}\ ,\ \text{Y}_0,\ \tau)\, P(Y_0|Z,X)\, P(\tau|Z,X) = P(Y_1|Y_0,\tau)\, P(Y_0|X)\, P(\tau|X)$. This isolates the two sub-mechanisms identified by Künzel et al., 2019 - the *outcome mechanism* and the *treatment effect* mechanism.

I add to these mechanisms the idea of an *observation mechanism* which produces the sample of units - both treated and untreated - that are available for study. The *observation mechanism* is not directly useful in revealing the causal effects for the units within the sample but it is important for generalizing the results beyond this group. That is, one can only generalize an inferred effect to a wider population if one assumes the observation mechanism is producing representative samples of the wider population. Further, even if it is not used directly in the causal estimators, this mechanism is an important part of the generative process which produces the observed data and can thus affect estimator performance. For simplicity, I will ignore potential conditioning on a wider population and sampling process and simply refer to a generative distribution over observed units X .

- **The Observation Mechanism:** $P(X)$

These three conditional distributions - representing the component causal mechanisms - combine to produce a joint distribution over the observed variables:

$$P(X,Y,Z) = P(Y|X,Z)\, P(Z|X)\, P(X)$$
$$P(X,Y,Z) = P(Y_1|\text{Z}\ ,\ \text{Y}_0,\ \tau)\, P(Y_0|Z,X)\, P(\tau|Z,X)\, P(Z|X)\, P(X) \text{ given that}$$
$$Y = Z \times Y_1 + (1-Z) \times Y_0$$

It is trivially true than any observed dataset can be described by some joint distribution $P(X,Y,Z)$ over the variables X, Y and Z . Above, I have built this joint distribution from the 'bottom up' by defining different component sub-mechanisms of the general causal mechanism, expressing these mechanisms as conditional distributions, and combining them into the joint distribution. An equivalent process for this analysis would be to start from the joint distribution and assert, based on conditional independence implied by some generative model, that the joint distribution can be factored in the way above. The simple but representative generative model in Figure 1 of the previous chapter produces the exact factorization above under the conditional independence relations implied by its structure. The equivalent generation model for the potential outcome based framing is given below in Figure 2.

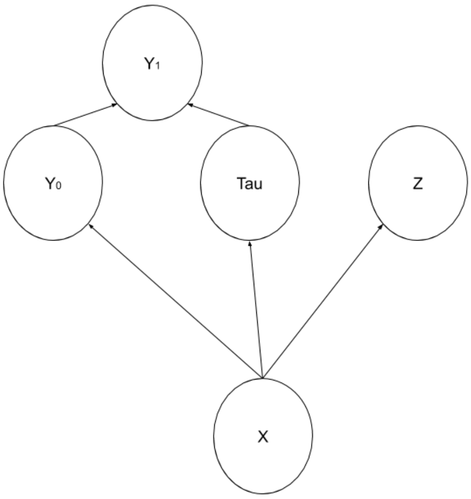


Figure 2: Graphical (Generative) Model for the variables in Rubin's potential outcomes framework.

With this joint distribution and its factorization in hand, the next conceptual step is to connect the aspects of the observed data discussed above to properties, and factored components, of the distribution. This is relatively straight forward. Some of the aspects of the observed data discussed above relate to properties stemming from a single (factorized) component of the joint distribution. For example, the non-linearity of the outcome mechanism dependents on the functional form of $P\left(Y_0|X\right)$. Others stem from the combined effect of multiple components. For example, the covariate balance in each group can be described by $P\left(X|Z\right)$ which by a simple application of Bayes Theorem is given by the formula below. This formula implies that as the treatment assignment mechanism varies away from simple random assignment $P\left(Z|X\right) = 0.5 \forall\, X$, the covariate balance in the groups will diverge (as one expects based on intuition).

$$P\left(X|Z\right) \propto P\left(Z|X\right) P\left(X\right)$$

A complete mapping between the various aspects of the observed data discussed above and properties of the joint distribution is provided later. For now, I simply assert that such a mapping exists based on the motivating examples above. The question then becomes why framing the different aspects of the data discussed above in terms of the joint distribution is useful.

There are three reasons for this. Firstly, it unifies conceptually disparate aspects of the observed data - functional form, covariate balance, predictive power etc - into a single theoretical object. The joint distribution fully determines the aspects of the data which impact performance. Thus, the *distributional setting* can be used to coherently describe a class of joint distributions which reliably result in similar estimator performance characteristics. All joint distributions with similar, quantitatively-defined properties - that is, with the same distributional setting - should exhibit similar performance. Second, this framing provides a clear connection between the process used to generate the observed data, the observed data itself, and the performance of estimators. The introduction briefly mentioned that Monte Carlo methods can sample from flexibly-specified joint distributions. The dependence of estimators on the joint distribution over X, Y , and Z - as established here - is what results in Monte Carlo being a useful evaluation tool. Monte Carlo sampling can be used to generate joint distributions corresponding to different distributional settings in which estimators can be expected to perform quite differently relative to each other and to their performance in other settings. Third, combining these two reasons, a concrete notion of distributional setting allows for the construction of an explorable causal inference problem space. The abstract problem space contains all of the meaningfully different distributional settings (those with different impact on estimator performance) and the use of Monte Carlo generative processes allows for the exploration of this space in order to validate estimators in a (theoretically) exhaustive set of circumstances. Establishing this space is the focus of the next section.

3.3 The Causal Inference Problem Space

The first section in this chapter used examples of estimators to highlight sensitivity to aspects of the observed data. The second section introduced the idea that the performance-relevant aspects of the observed data can be fully described in terms of a *distributional setting* which describes the properties of the joint distribution $P(X, Y, Z)$. This section brings these ideas together by describing the so-called axes which define the space of distributional settings eluded to at the end of the last section. The seven axes introduced below represent the properties of the joint distribution which can vary and, in doing so, produce changes in the performance of estimators when applied to the corresponding data. The resultant problem space is intended to be exhaustive, implying that the axes should span all *distinct problem classes* - classes of joint distributions which yield different estimator performance. The claim that the defined space is exhaustive is hard to prove concretely. Indeed, I expect that iteration on the axes may be required to meet the goal of defining an appropriate basis. However, it is worth noting two points which motivate the choice of axes below:

- The axes were selected in a principled manner by examining the properties of the joint distribution affected by manipulation of the underlying mechanisms described in the last two sections - the treatment assignment mechanism and the outcome mechanism. This is not a guarantee that all relevant distributional properties have been discovered but it does mean those selected are non-arbitrary and arise from varied causal mechanisms.

- The axes selected by this principled process end up spanning the evaluation settings present in the benchmarking literature that is reviewed in the next chapter. This means that, at a minimum, these axes span the space of settings present in a wide sample of the relevant literature even if the resultant (sub)space isn't exhaustive of the full problem space.

Finally, it is worth noting that, while I do not present them here, each axis has at least one associated metric which assigns it a specific value for any given joint distribution. If the axes do span the causal inference problem space, then combinations of these values should be able to represent every distinct class of causal inference problem.

Without further ado, I present the proposed axes of the causal inference problem space. For each axis, I explain which distributional component(s) of the overall causal mechanism affect its value. Following each definition (or group of related definitions), I provide simple toy datasets and estimators to demonstrate the impact of the axis on estimator performance. The toy datasets are based on a single covariate X and show the simplest possible version of all mechanisms except those in focus.

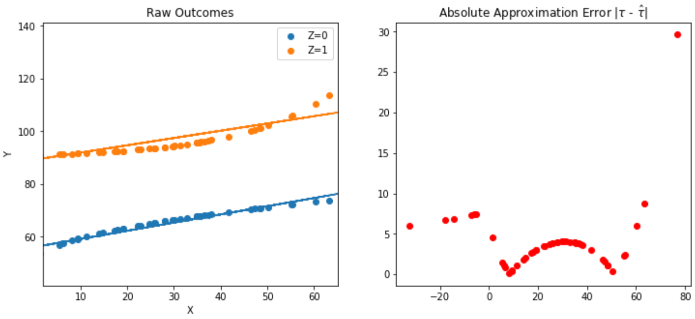
1) **The Nonlinearity/Complexity of the Outcome Mechanism** which defines the distribution over the value of Y_0 (the outcome without treatment) based on a unit's covariates. This corresponds to the functional form of the distribution $P(Y_0|X)$.

2) **The Nonlinearity/Complexity of the Treatment Effect Mechanism** which defines the distribution over the treatment effect for each unit. This corresponds to the form of the distribution $P(\tau|X)$ and specifically the dependence on the covariance X .

Toy Example

The example below demonstrates the impact of a nonlinear response mechanism when using parametric (potentially misspecified) conditional mean regression. This example is inspired by a similar plot found in Hill, 2011. The outcome and treatment effect mechanism are non-linear and this produces an estimation error when using a misspecified linear model.

$$X \sim N(30, 25)$$
$$Y_0 \ \sim 50 + 3 \times \sqrt{X}$$
$$\tau \ \sim 90 + e^{0.05X} - Y_0 \Rightarrow Y_1 = 90 + e^{0.05X}$$
$$P(Z|X) = 0.5$$

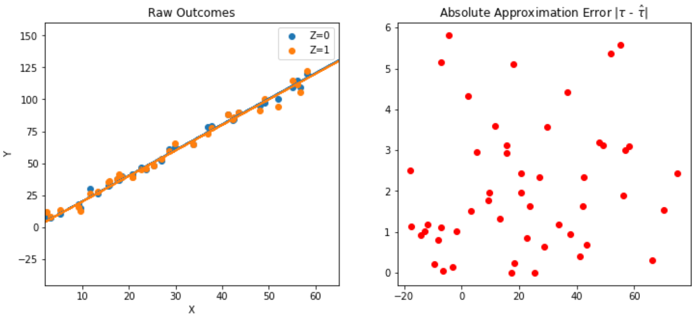


3) **The Magnitude of the Treatment Effect:** the magnitude of treatment effect, measured relative to noise in the outcome mechanism (as defined above). This corresponds to the magnitude of the values produced by $P(\tau|X)$ and the magnitude of the noise in $P(Y_0|X)$ - IE, the changes in Y_0 not correlated with changes in X.

Toy Example

The example below demonstrates the impact of a treatment effect which is small relative to the noise in the response mechanisms. The base outcome is linear and the treatment effect is constant. This is a classically easy problem but, in this case, the noise makes correctly-specified parametric inference highly inaccurate.

$$X \sim N(30, 25)$$
$$Y_0 \ \sim N(2 \times X, 3)$$
$$\tau \ \sim \ 0.5$$
$$Y_1 = N(Y_0 \ + \tau, 3)$$
$$P(Z|X) = 0.5$$



4) **The Nonlinearity/Complexity of the Treatment Assignment Mechanism** which defines the distribution over the treatment assignment value based on a unit's covariates. This corresponds to the form of $P(Z|X)$, the treatment assignment mechanism per the definitions above.

Toy Example

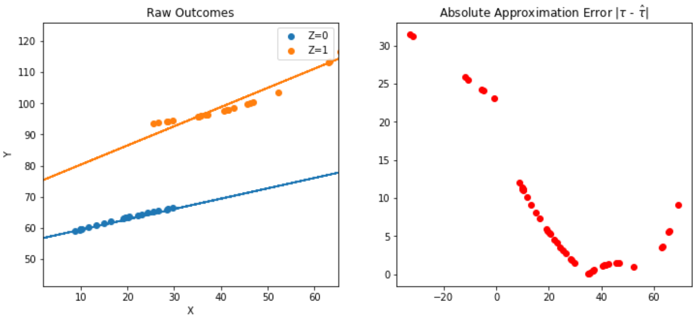
NOTE: I haven't yet constructed this example. As it doesn't conform to the standard template of the other examples.

5) **The Covariate Balance:** The distributional (dis)similarity of the covariates between the groups. As above, this is determined by $P(Z|X)$ and $P(X)$.

Toy Example

The example below demonstrates the effect of covariate imbalance induced through strong selection pressure. The example is the same as the one shown for the non-linear response axes but covariate imbalance is induced between groups using the treatment mechanism below. This result is that the conditional regressions extrapolate beyond the bounds of the outcome information, producing large errors.

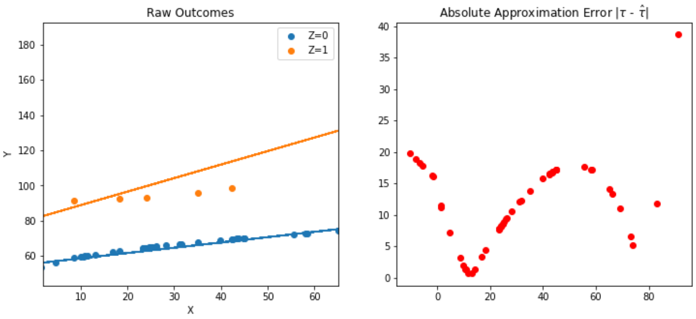
$$P\left(Z|X\right) = Bernoulli\left(p = min\left(1, X/30\right)\right)$$



6) **The Percent of Units Treated:** the (dis)similarity of the number of units in each group. Again, primarily determined by $P\left(Z|X\right)$ and $P\left(X\right)$.

<u>Toy Example</u>

The example below demonstrates the effect of a small percent of units treated. The result is similar to the case of imbalance but holds even in the presence of asymptotic balance. IE, in this case $P\left(Z|X\right) = P\left(Z\right) = 0.1$. With enough samples, there would be no need for extrapolation but in the finite sample limit, there is an accuracy issue.



7) **The Degree of Alignment Between the Treatment Assignment and Response Mechanism[7]:** the number of confounding covariates relative to the number of non-confounding, but still, predictive covariates.

[7] The term alignment is taken from Kern, Stuart, Hill, and Green, 2016.

Toy Example

NOTE: I haven't yet constructed this example. As it doesn't conform to the standard template of the other examples.

Beyond these seven primary axes, there are three further properties of observed data which are relevant to performance but are not strictly properties of the joint distribution. These correspond to the two assumptions - ignorability and overlap - and the number of observed samples. I include these as part of the definition of the distributional setting despite the slight definitional complexity this brings about.

3.4 Conclusion

This chapter laid the theoretical groundwork necessary for the robust evaluation of causal inference estimators. Different estimators target different components of the causal mechanism and this produces sensitivity to different properties of the joint distribution over the observed data. The Causal Inference Problem Space is a space defined by ten axes along which the properties of the joint distribution vary. Understanding this space is important because different estimators are likely to perform differently on distributions which lie in different parts of it. With this foundation established, the next chapter goes on to review the methods used to evaluate causal inference methods in the literature.

NOTE: I am not that pleased with the transition from this chapter to the next. Will smooth it out in the next draft.

Part III

Benchmarking Causal Inference Methods

Chapter 4

Literature Review

This review explores and synthesizes existing approaches for the evaluation of causal inference estimators. I aim to achieve three primary goals. First, to justify the importance of finite sample evaluation based on Monte Carlo methods by exploring the landscape of alternatives. Second, to establish principles for the validity of finite sample evaluation. Finally, third, to analyze the validity of existing approaches in terms of these principles in order to motivate a better approach. This literature review is comprised of four sections that work toward these goals. First, I briefly establish definitions for statistical inference and estimators which perform this inference. Second, I explore the landscape of different approaches to evaluating the performance of estimators - comparing theoretical, asymptotic evaluation and Monte Carlo-based finite-sample evaluation. Third, I establish the general principles for the validity of finite sample evaluation. Finally, I apply these principles to analyze existing approaches to evaluating causal inference estimators. The first three sections present this author's framing of the literature with light reference to the underlying works. The final section is a close review of the existing methods of Monte Carlo based evaluation.

4.1 Statistical Inference and Estimators

All statistical inference can be characterized in the following simple terms, due to Pearl, 2009: There is some true data generating process (DGP) characterized by a set of (random) variables, functions (possibly probability distributions) that relate these variables, and parameters that determine the exact form of the functions. The DGP gives rise to a joint distribution over the variables which make up the DGP. Any sample of observed data can then be thought of as a sample from this joint distribution. The goal of inference is to estimate the functional relationships and parameters of the original data-generating process using sample data from the joint probability distribution.

As summarized in Calder, 1953, estimators operationalize inference by providing a mathematical mapping between a sample of data from the true joint distribution and an estimate of a target parameter - an *estimand* - that is (or is concretely related to) a parameter from the true DGP. Given that a sample is inherently stochastic, estimators give rise to a *sampling distribution* over the estimand rather than a single value. The quality of an estimator is defined by some quality metric that depends on this sampling distribution. Concretely defining the abstract notion of the 'quality' of an estimator is a non-trivial task, discussed below.

4.2 Evaluating the Performance of Statistical Estimators

Ultimately, the goal of evaluation is to determine the quality of an estimator. There is no single measure for quality. The metric used is sensitive to the eventual application of the estimator and the real-world, normative value associated with various kinds of error. Bishop, 2006 Chapter 1 outlines most of the commonly encountered quality metrics and in which circumstances they tend to be most useful. The common thread which unifies these metrics is that they depend on the estimator's sampling distribution over the estimand. All metrics are, at some level of abstraction, calculated in reference to this sampling distribution Calder, 1953. So, the estimator itself, the sample size available, and the underlying joint data distribution affect the value quality metrics by affecting the sampling distribution.

Given the central role of the sampling distribution, one can think about evaluation approaches in terms of how they derive the sampling distribution for a given estimator. In this framing, there are two approaches that I will refer to as *theoretical* and *experimental*. Theoretical methods derive an analytical expression for the sampling distribution based on the estimator, sample size, and underlying data distribution. Experimental approaches *estimate* the sampling distribution by repeatedly applying the estimator to samples from an underlying data distribution with a known value for the estimand. Repeat sampling results in convergence to the expected value of any metric defined over the sampling distribution Paxton, Curran, Bollen, Kirby, and Chen, 2001. This provides *experimental* insight into the performance of the estimator in the sense that data is collected to estimate the quality of the metric under researcher-controlled settings of the sample size and underlying distribution. In general, the process of using samples from a distribution to estimate the value of parameters over that distribution is referred to as Monte Carlo estimation Hastings, 1970. This is why sampling-based evaluation is referred to as Monte Carlo Evaluation in the literature.

These two approaches - theoretical and experimental - are not mutually exclusive and, for the evaluation of an arbitrary estimator, can both provide useful information. However, honing in on our specific context, Paxton et al., 2001 provide a compelling analysis of the specific usefulness of Monte Carlo evaluation in evaluating econometric estimators. They raise three concerns with the usefulness of theoretical evaluation which are outlined below along with corresponding points made by authors who reach the same conclusions with regard to evaluating causal inference estimators specifically.

1) In many cases, the theoretical sampling distribution for an estimator is not known. As pointed out in Knaus et al., 2018, more complex estimators tend to make for much more challenging theoretical analysis. This fact, combined with the trend toward semi/nonparametric causal estimators outlined in the introduction, means that theoretical sampling distributions are missing for many causal inference estimators which have proven promising in practice.

2) Theoretical evaluation tends to rely heavily on simplifying assumptions - the most common of which is the assumption of the normality of the underlying data distribution. These assumptions limit the applicability of the theoretical sampling distribution: it is unlikely that real-world data distributions meet these assumptions and thus the theoretical metrics may not hold on real data. Even if

the assumptions behind the theoretical analysis of different methods are available, Knaus et al., 2018 point out that the assumptions behind the analysis for different estimators may not overlap such that comparing theoretically derived metric values is impossible.

3) Theoretical analysis tends to hold in the asymptotic limit of large sample sizes. The implications of theoretical results for finite samples is not necessarily well defined. This is echoed in Huber, Lechner, and Wunsch, 2013.

Taken together, these three weaknesses mean that Monte Carlo analysis - which can be performed using arbitrary sample sizes and on any underlying data distribution - plays a crucial role in the evaluation of causal inference estimators. The rest of this review focuses exclusively on these methods of evaluation, exploring the principles which define a good Monte Carlo evaluation and how existing variants measure up against these principles.

4.3 Monte Carlo Evaluation of Causal Inference Estimators

4.3.1 General Principles

This section proposes a synthesized set of principles for what makes a *good* Monte Carlo-based evaluation method. Much like is the case for evaluating individual estimator quality, there is no universal metric for the quality of an evaluation method. Two thematic axes of quality are common in the literature. I refer to these axes as the *axis of specific validity* and the *axis of general validity.* I use the word axis to represent the non-discrete nature of validity. We will see that there appears to be a trade-off between specific and general validity and that different methodological choices will have different expected levels of validity on these axes. The levels of validity are hard to quantify but this should not be confused with the absence of a continuum.

4.3.1.1 The Axis of Specific Validity

Specific validity refers to whether a Monte Carlo evaluation provides a valid measure of an estimator's performance in some specific distributional setting. A distributional setting is defined by a given DGP (joint distribution of the data) and the sample size drawn from this DGP. On the surface, internal validity appears to hold (trivially) for any Monte Carlo evaluation. There is some joint distribution from which a fixed size sample is drawn, the estimator is applied to the sample and the estimand value is collected. If this is repeated enough times, the distribution over the estimand for the distributional setting can be approximated with arbitrary precision. The problem is revealed by the nuance that researchers do not target arbitrary DGPs. The power of Monte Carlo - as established above - is that it does not rely on simple joint DGPs/joint distributions and, thus, can be used to validate *realistic* DGPs (DGPs which could be found in the real world). This means that Monte Carlo evaluations are (and should be) designed with DGPs which aim to be representative of the real world. The quality with which the Monte Carlo DGP represents the *specific* real-world setting targetted is the source of *specific validity.* It is possible for a researcher to target a realistic setting but fail to use a DGP which is representative and, therefore, end up producing a result that is not valid in the *specific* setting. This is one of a set of related failure modes for specific validity which will be explored below. The others include biased specification and the creation of DGPs which do not ensure convergence to the true sampling distribution.

Concern for this validity is present in various papers that evaluate causal inference estimators using methods derived from Monte Carlo evaluation. Dorie, Hill, Shalit, Scott, and Cervone, 2019 refer to the importance of the "calibration of the DGP to the real world" . They observe that if the types of covariates, their marginal and joint distribution, and the functional forms which relate them to the treatment and outcome variables are too simple, then the DGP will not be representative of the real world. Paxton et al., 2001 refer to the need for example-data-driven DGP design in order to avoid the creation of DGPs purpose-built to confirm the efficacy of an estimator despite having "little resemblance to [data and] models encountered in practice" . The common concern here is that a positive evaluation result from an unrealistic DGP, or a DGP which poorly represents some real DGP, will not correspond to evidence that the estimator works in a specific, real-world setting. This is where the idea of *specific validity* originates.

4.3.1.2 The Axis of General Validity

General validity refers to whether a Monte Carlo evaluation provides insight into whether an estimator generalizes well across different distributional settings. Dorie et al., 2019 observe that "while it is natural to make sure that a method works in the specific scenarios for which it is designed, this doesn't necessarily help a general researcher understand how it might perform more broadly" . The implicit premise in this observation is that it is uncommon for practitioners to know the DGP which underlies a given set of observational data. This means it is important to know how a given estimator will perform across a range of relevant distributional settings. The assumption being that an estimator that performs well (or better than other estimators) across a range of relevant distributional settings is the one which is most likely to be best when the true setting is unknown. While superior performance is not guaranteed - unless the estimator is tested against an exhaustive set of distributional settings - it does appear that performance on a range of settings is a sensible and useful heuristic to determine general robustness. This is where the idea of *general validity* originates.

4.3.1.3 The Trade-off between Specific and General Validity and the Optimal Design Curve

Both Paxton et al., 2001 and Wendling et al., 2018 point out that "there is a trade-off between realism and control in any Monte Carlo design" Paxton et al., 2001. This implies a trade-off in specific and general validity. General validity requires evaluation of a representative set of well-defined distributional settings but the control required to specify the DGPs that comprize this well-defined set tends to result in less realistic, and therefore less *specifically valid*, evaluations. The mechanistic reasons for this are explored in the next section.

For now, I observe that there is no single answer to the right balance between general and specific validity. Researchers working on causal Inference methodology may care more about specific validity in one distributional setting in order to demonstrate that a new method achieves its design purpose. Researchers working on applied problems in social and political science may care more about general validity to ensure robust performance in unknown settings.

One could imagine a set of designs representing optimal trade-offs where, for any given design, the external validity is maximized relative to a minimum constraint placed on internal validity. This would trace out an abstract 'optimal frontier' curve in the design space - as in Figure 3A below. Any design along this curve could not improve its validity on either axis without sacrificing validity on the other. Designs on this curve would all be equally *good* in the sense that they may be useful in different circumstances. Unfortunately, as will be seen in the next section, the majority of the actual designs do not lie on this optimal curve and, with specific design improvements, one/both the specific and general validity could be improved without decreasing the validity of the other. The location of such a design - and a path for unambiguous improvement - is demonstrated in Figure 3B. The rest of this review is focused on locating existing approaches to Monte Carlo evaluation in this design space and exploring how they could be moved closer to the optimal frontier.

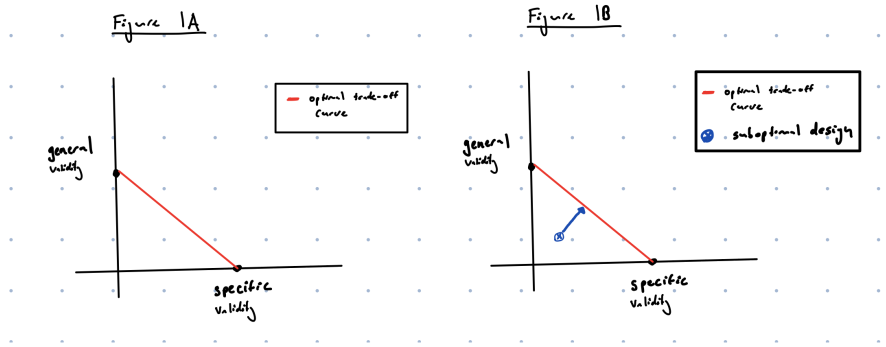


Figure 3: Left panel: A point which is not on the optimal trade off curve. Right panel: A point on the optimal trade off curve.

4.3.2 (Sub)Optimal Approaches

At this stage, it's clear that Monte Carlo based evaluation is a particularly useful method of assessing the performance of causal inference estimators. In the real world, practitioners deal with finite data samples from unknown but complex DGPs. Monte Carlo is capable of evaluating estimators in these settings while theoretical analysis is hard or impossible. It is also clear that Monte Carlo evaluation can vary in validity along two axes - specific and general validity - and that different designs may make (sub)optimal trade-offs between validity on these axes. The analysis which established these two points applied to Monte Carlo evaluation in an abstract sense. This section reviews concrete implementations of Monte Carlo evaluation of causal inference estimators from the literature, analyzing them in light of the axes of validity introduced above. This move toward analyzing concrete implementations brings with it vocabulary which is specific to causal inference. I refer the reader to the theory chapter above for clarification of any unfamiliar terms.

The analysis in this section divides Monte Carlo evaluation methods into two categories based on the underlying design of their DGP: *pure* and *hybrid*. *Pure* methods are further subdivided into purely *synthetic* and *empirical* DGP designs. *Hybrid* designs are hybrids because they combine the two pure designs strategies in a single DGP. This is a 'novel' ontology introduced to summarize the existing variants of Monte Carlo. All of the designs share the same fundamental structure - in the sense that the same variables must be sampled so that a causal inference estimator can be evaluated - but exhibit different validity characteristics when judged through the lens of specific and general validity. I proceed by, first, establishing a generic DGP - an abstract representation of the DGP used in all Monte Carlo evaluation methods. I then express the various designs found in the literature in terms of different strategies for the concretizations of this generic DGP, making comparative analysis straightforward. Finally, I analyze the implication of different concretization strategies in terms of the axes of validity from above.

4.3.2.1 A Generic DGP for the Monte Carlo Evaluation of Causal Inference Estimators

This sections builds towards a formal specification scheme for the DGPs which underpin Monte Carlo evaluation of causal inference estimators. The scheme is defined in terms of a generic DGP, the functional components of which can then be defined/instantiated to describe a specific, concrete DGP design.

First, Table 1 below recapitulates the variables over which a causal inference problem-instance is defined and provides the associated dimensionality and definition. Full explanations for these variables are presented in Chapter 2 above.

$X \in R^{(n,d)}$	X is the covariate matrix that contains measurements of d covariates for a population of n individuals
$T \in \{0,1\}^n$	T is the treatment status of the n individuals for a binary treatment
$Y_0 \in R^n$	Y_0 is the outcome under control (absence of treatment) referred to as the untreated outcome
$Y_1 \in R^n$	Y_1 is the outcome under treatment referred to as the treated outcome
$TE \in R^n$	TE is the individual treatment effect
$Y \in R^n$	Y is the observed outcome

Table 1: Variables which define a causal inference problem instance

The variables in Table 1 are related through four generic functions given in Table 2 [1]. A concrete Monte Carlo evaluation design can, thus, be fully specified by the concrete instantiation of these four functions.

[1] Note that the functions specified here may be stochastic functions, functions which do not necessarily produce the same output for the same input on every evaluation.

$\rho : () \rightarrow R^{(n,d)}$	The **covariate population sampler** which represents the joint distribution over the d covariates and produces samples of the covariate matrix X. Strictly speaking, X contains sampled covariate observations for n individuals such that a 'sampled' X is constructed by repeatedly sampling the underlying joint distribution over the covariates
$\Omega : \ R^{(n,d)} \ \rightarrow R^{n}$	The **treatment assignment sampler** which assigns a treatment status T to the individuals in X
$\Phi : \ R^{(n,d)} \ \rightarrow R^{n}$	The **untreated outcome sampler** which assigns an outcome under control (Y_0) to the individuals in X
$\tau : \ R^{(n,d)} \ \rightarrow R^{n}$	The **treatment effect sampler** which assigns a treatment effect (τ) to the individuals in X

Table 2: Functions which generate and relate the variables in a causal inference problem instance

A complete (but generic) DGP is then captured by the procedure presented in Table 3. This procedure relates the variables defined in Table 1 using the functions defined in Table 2.

Generic DGP Procedure
$X \leftarrow \rho\,()$
$T \leftarrow \Omega\,(X)$
$Y_0 \leftarrow \Phi\,(X)$
$TE \leftarrow \tau\,(X)$
$Y_1 \ \leftarrow Y_0 + TE$
$Y = T \times Y_1 + (1 - T) \times Y_0$

Table 3: The procedure for a Monte Carlo evaluation DGP defined in terms of the variables from Table 1 and functions from Table 2

Note three facts about this generic DGP:

- The designer has a maximum of four degrees of freedom in specifying the DGP. These degrees of freedom correspond to selecting the four functions ρ, Ω, Φ, τ .

- This generic DGP provides all of the information required to evaluate a causal inference estimator. The estimator receives X, T, and Y and produces an estimate of either $E\,[TE|X]$ - the conditional treatment effect - or $E\,[TE]$, the average treatment effect. The ground-truth (sampled) values of TE are then be used to evaluate the estimates.

- This DGP is constructed to ensure ignorability. Following Dorie et al (2019): Assume Φ, Ω and τ are probability distributions, then we have the following factorization as a result of the DGP form above (and specifically as a result of the absence of T as an argument to Φ, τ): $p\,(Y_0,\ Y_1, T|X) = p\,(Y_0,\ Y_1|X) \times$

$p\,(T|X)$ which implies $p\,(Y_0,\ Y_1|T, X) = p\,(Y_0,\ Y_1|\ X)$ or $Y_0,\ Y_1 \perp\!\!\!\perp T|X$. Non-ignorability can be induced by dropping some of the d covariates in X but, by default, it is guaranteed.

Taken together, these three facts mean that specifying the four functions is sufficient to generate a causal inference problem-instance. In fact, the minimal level of constraints on the four degrees of freedom means that this generic DGP can be used to specify a concrete DGP design corresponding to any joint distribution which factors in the way outlined above.

Having established this generic DGP, I turn to the idea of concretization strategies. Different strategies for specifying the four base functions - thus concretizing the generic DGP - give rise to the different design categories which were outlined above. There are three mutually-exclusive and exhaustive strategies for concretely specifying a function:

1) A function can be manually specified by constructing a mathematical (distribution) function from elementary mathematical building blocks and specifying all the parameters which appear in its definition. This produces functions with known form and parameter values. I will label functions specified in this way as $f_{synthetic}$.

2) A function can be estimated from empirical data by specifying a fixed functional form and inferring the unknown parameter values. This produces functions with a known form but unknown ground-truth parameters. Functions produced by this strategy will be labeled $f_{estimated}$.

3) A function can be 'specified' to exactly match empirical data. In this case, the mathematical form of the function and any associated parameter values are unknown. This strategy produces functions labeled $f_{empirical}$.

The latter two strategies necessitate access to empirical data from some target DGP. Variables that are sourced empirically for this purpose will be labeled as $A_{empirical}$ where $A\ \in \{X, T, Y\}$ (the set of observable variables).

With the generic DGP and the three concretization strategies in mind, I proceed to review and categorize the designs found in the literature.

4.3.2.2 Pure Designs for Monte Carlo Evaluation

Pure designs are pure because they rely exclusively on sampling functions of the type $f_{synthetic}$ - in the case of synthetic designs - or $f_{empirical}$ - in the case of empirical designs. Below, I outline each of the resultant design approaches and how they tend to appear in the literature.

Synthetic Monte Carlo Designs Synthetic designs rely on hand crafted functions to specify the four degrees of freedom which control the distribution over covariates, treatment assignment, untreated outcome and treatment effect. These designs are common in literature which introduces new estimators designed to tackle specific distributional settings as the manual specification allows careful control over the resultant joint distribution. For recent examples, see Künzel et al., 2018; Fredrik D. Johansson et al., 2018, Fredrik D Johansson et al., 2016. Designs which fit into this class are used extensively in the literature beyond this small sample of recent papers. J. D. Y. Kang and Schafer, 2007 introduce a benchmark based on four independent, normally distributed covariates, a linear outcome with Gaussian noise, a constant treatment effect and a linear expit for the propensity score. Imbalance is created by variable censoring of individuals based on assignment treatment. This benchmark is cited in over eight hundred works at the time of writing, the majority of which use the proposed DGP for evaluating causal inference estimators or use results from analysis of this kind by others. Kallus, 2016 introduces a similar benchmark which attempts to more accurately model non-linearities present in real data. The benchmark presented is that work is based on two covariates drawn from a uniform distribution, a non-linear propensity score expit and normally distributed outcomes with means derived from the covariates. This benchmark is cited in 10 other works, mostly those applying semi/nonparametric estimators. In both cases, the stated design principle is to create a DGP that is representative of the kinds of data found in practice.

The top left panel of Table 4 below represents the synthetic design using the function labels and generic DGP defined above. All four degrees of freedom are specified to be synthetic functions and the consequence is a DGP devoid of empirical grounding. The bottom left panel expresses the J. D. Kang and Schafer, 2007 DGP design as an assignment of the four base functions.

Empirical Monte Carlo Designs In contrast to synthetic designs, empirical designs specify the four functional degrees of freedom using real-world data. The challenge is how to do this in a way that still provides access to the treatment effect which is required to evaluate an estimator. Empirical designs overcome this challenge by using data randomized control trials, replacing the control group with non-randomized survey/administrative observations over the same covariates that were observed in the experiment Huber et al., 2013. The average effect from the randomized control trial provides the 'ground truth' target for the average effect estimand. Examples of this design for evaluation include Lalonde, 1986, Fraker and Maynard, 1987, Friedlander and Robins, 1995, Heckman and Todd, 1998, R. H. Dehejia and Wahba, 1999, Smith and Todd, 2005, and Flores and Mitnik, 2009.

There are three things about this design worth noting. First, the empirical specification of the functional degrees of freedom means the true distributional setting is not known. Second, and related, the ground truth is itself an estimate rather than a known value. This produces additional uncertainty in the results which come from this evaluation method. Finally, this design can, by default, on evaluate estimators for the average treatment effect. The conditional average treatment effect must be estimated from the data using the very methods which we would like to evaluate using a Monte Carlo benchmark. This means the evaluation of conditional average treatment estimators is impossible under empirical designs of the kind described here.

The top right panel of Table 4 below represents the empirical design using the function labels and generic DGP defined above. All four degrees of freedom are specified to empirical functions (although some of these are unknown). The consequence is a DGP grounded in empirical data but with an unknown joint distribution over X, T, and Y and an unknown values for the true treatment effect and potential outcomes. The bottom left panel expresses the Lalonde, 1986 DGP design as an assignment of the four base functions.

There is an obvious concern worth addressing in calling this a Monte Carlo design. The component functions are specified by data rather than sampling functions which means that the DGP cannot be used to generate multiple samples from the joint distribution over the variables. As a result, this method does not provide access to a sampling distribution for the estimand but rather only a single estimate. I believe the Monte Carlo classification is justified for two reasons despite this concern. Firstly, we can conceive of the empirical design as a single sample from some latent DGP which could - in theory - be used produce more samples. This design might be weakened by the small (single) sample size but this doesn't change the fundamental principle behind its operation. Second, the analysis below will introduce hybrid designs which share common components with the pure empirical design but do allow sampling. In this sense, the pure empirical design can be thought of the extreme (and somewhat degenerate) end of a Monte Carlo design spectrum.

	Synthetic Monte Carlo	**Empirical Monte Carlo**
Concretized DGP	$\rho_{synthetic} \leftarrow f_{synthetic}$ $\Omega_{synthetic} \leftarrow f_{synthetic}$ $\Phi_{synthetic} \leftarrow f_{synthetic}$ $\tau_{synthetic} \leftarrow f_{synthetic}$ $X \leftarrow \rho_{synthetic}\left(\right)$ $T \leftarrow \Omega_{synthetic}\left(X\right)$ $Y_0 \leftarrow \Phi_{synthetic}\left(X\right)$ $TE \leftarrow \tau_{synthetic}\left(X\right)$ $Y_1 \ \leftarrow Y_0 + TE$ $Y = T \times Y_1 + (1 - T) \times Y_0$	$\rho_{empirical} \leftarrow f_{emp} = X_{empirical}$ $\Omega_{empirical} \leftarrow f_{emp} = T_{empirical}$ $\Phi_{empirical} \leftarrow f_{emp} =?$ $\tau_{empirical} \leftarrow f_{emp} =?$ $X \leftarrow X_{empirical}$ $T \leftarrow T_{empirical}$ $Y_0 \leftarrow ?$ $TE \leftarrow \ ?$ $Y_1 \ \leftarrow ?$ $Y = Y_{empirical}$
Example Concretization	Kang and Schafer (2007) $\rho \leftarrow N\left(X\|0,1\right)$ $\Omega \ \leftarrow expit(-X_1 + 0.5X_2$ $... - 0.25X_3 - 0.1X_4)$ $\Phi \leftarrow N(Y\|210 + 27.4X_1$ $... + 13.7\left(X_2 + X_3 + X_4\right), 1)$ $\tau \leftarrow \ C \ \in R$	Lalonde (1986) $\rho \leftarrow \{X_{experiment}, \ X_{survey}\} \sim ?$ $\Omega \leftarrow \ T_{empirical} \ \sim ?$ $\Phi \leftarrow \ ?$ $\tau \leftarrow \ ?$
Joint Distribution over Observables	$P\left(X, Y, T\right) \sim \rho \times \Omega \times \Phi$	$P\left(X, Y, T\right) \sim \ ?$

Table 4: Synthetic vs Empirical Monte Carlo Designs expressed as concretized versions of the generic DGP along with examples framed in terms of assignments to the four functional degrees of freedom.

Comparing the Designs The framing proposed above clarifies how the theoretical trade-off between specific and general validity manifests in the pure design strategies. Synthetic designs provide granular control of a well-defined distributional setting because all of the component degrees of freedom are set by hand with functions of known form and parameterization. This enables easy testing of estimators across a range of settings, promoting general validity. But, the hand-crafted nature of the function specification raises concerns about the correspondence of the DGP to reality, thereby weakening specific validity. Empirical designs are realistic and correspond to a valid evaluation (with some caveats) of a specific distributional setting (specific validity) but do not allow for easy control of the distributional setting, weakening general validity.

This comparison between the designs, grounded in concretization strategies, provides some insight but is still too abstract. It is unclear what is meant by the ideas of realism and control which appear throughout the paragraph above. Below, I synthesize the work in Paxton et al., 2001, Huber et al., 2013 and Dorie et al., 2019 to compare the properties of synthetic and empirical designs that contribute to specific and general validity. Many of the points which appear below are implicit or explicit in the analysis above but this complete enumeration will prove useful later when considering hybrid designs which mix the properties of the pure designs.

Synthetic designs tend to have weak specific validity for three reasons which stem from the manual specification of functions:

- Unrealistic covariate distributions: the number, type (continuous, categorical, binary) and joint distribution (over the covariates) is often poorly calibrated to the data researchers encounter in practice. It is unlikely that covariates in the real-world are drawn from a multivariate normal distribution.

- Unrealistic treatment and outcome functions: the complexity of the functional forms - in terms of interactions, non-linearity etc - is often poorly calibrated to mechanisms likely to be found in the real world. It is unlikely that the treatment policy in the real world is a logistic function with a linear expit.

- The potential for specification bias: because covariates and treatment/outcome functions are selected by hand there is the potential for intentional or unintentional bias in the design such that the DGP favors the method being evaluated. This implies DGPs with similar properties would or could lead to worse results. Consider that many researchers may change the DGP specification until their method produces the results they *expect* to see from it under the assumption that their DGP specification, and not their method, is at fault for unexpected results. This would induce the kind of bias described here.

However, synthetic designs have three characteristics which counter-balance those above when considering specific validity:

- The DGP can be guaranteed to meet the assumptions required for observational causal inference, specifically ignorability and overlap. This avoids false negatives in which an otherwise valid estimator produces poor results as a result of violated assumptions rather than inherent flaws.

- The DGP can be used to produce many samples and, therefore, to ensure convergence to a true sampling distributions for the estimand. Further, this provides access to a distribution over performance which is useful for quantifying best and worst case performance.

- The DGP provides access to the ground truth of the individual treatment effect and average treatment effect. This precludes the need for additional estimation of the ground truth and means that a wider range of estimators can be evaluated.

Moving on to consider general validity, synthetic designs have two properties which make for strong general validity:

- The joint distribution over covariates, treatment and outcomes is well-defined and known. This means the performance of the estimator is relative to a known distributional setting.

- The distributional setting is controllable such that the estimator can be tested over a range of settings.

Empirical designs are effectively polar opposite to synthetic designs when examined in terms of the eight properties above. In order of the points above: They possess inherently realistic covariate distributions and outcome/treatment functions and, by extension, unbiased specification (given that the researcher has no control). However, they do not guarantee that causal inference assumptions are met, do not allow repeat sampling and do not provide access to the ground truth of the average or individual treatment effect without estimation. Further, they use DGPs with an unknown distributional setting and do not allow for researcher control.

Table 5 provides a summary of the eight properties outlined above, which of the two validity axes they affect, the functional degrees of freedom from the generic DGP that affect the possession of the property, and whether each of the two pure designs possesses the (positive) property.

Property	**Validity Axis**	**Relevant functional degrees of freedom**	**Synthetic Designs**	**Empirical Designs**
Realistic Covariate Distribution	Specific	ρ	🟥	🟩
Realistic Treatment/Outcome Functions	Specific	$\Omega,\ \Phi, \tau$	🟥	🟩
Guaranteed Unbiased Specification	Specific	$\rho,\ \Omega,\ \Phi, \tau$	🟥	🟩
Access to Ground Truth	Specific	Φ, τ	🟩	🟥
Guaranteed to Obey Causal Assumptions	Specific	$\Omega,\ \Phi, \tau$	🟩	🟥
Allows Repeat Sampling	Specific	$\rho,\ \Omega,\ \Phi, \tau$	🟩	🟥
Known Distributional Setting	General	$\rho,\ \Omega,\ \Phi, \tau$	🟩	🟥
Controllable Distributional Setting	General	$\rho,\ \Omega,\ \Phi, \tau$	🟩	🟥

Table 5: The eight properties of Monte Carlo designs which affect specific and general validity of evaluation along with the functional degrees of freedom which impact this property

The analysis above establishes a more granular understanding of the properties of the two pure designs which produce the differences in their specific and general validity. It is clear that these two designs are polar opposites but this does not necessarily mean that they are suboptimal in terms of the trade-off between specific and general validity. If one could not do better in one without sacrificing the other, then these two designs would represent a (discrete) optimal frontier with researchers being forced to choose one or the other depending on their needs. Fortunately, this is not the case.

4.3.2.3 Hybrid Designs for Monte Carlo Evaluation

Hybrid designs combine the concretization strategies from synthetic and empirical designs to produce a Monte Carlo evaluation method with more of the desirable properties enumerated above. The unifying principle behind hybrid designs is the maximal use of data to inform specification of the DGP - to improve realism and specific validity - while maintaining the control required to allow general validity.

The earliest examples of thought in the direction of blending synthetic and empirical designs appear in Abadie and Imbens, 2002 and Diamond et al., 2012. Both papers use DGPs designed to mimic the data in Lalonde, 1986 but both use synthetic, simplified covariate sampling and hand-crafted treatment/outcome functions which are 'inspired' by the data but not directly informed by it.

Huber et al., 2013 formalize this idea and introduce what they call *Empirical Monte Carlo* (EMC). The context for the introduction of their method is instructive. Their paper attempts to resolve a disagreement between Frölich, 2004 and Busso, Dinardo, and Mccrary, 2014, both of whom evaluate the same set of propensity-score based estimators using synthetic Monte Carlo but find quite different results. Huber et al., 2013 points out that both papers use DGPS with poor specific validity as a result of arbitrary synthetic design. This inspires the design of a method which uses a "DGP not entirely [built] on relations specified by the researcher, but [one which] exploits real data as much as possible instead, [using] observed outcomes and covariates instead of simulated ones as well as an observed selection process" . The authors propose the following design: They define a population of observations as a dataset covering all German nationals with social insurance in the years between 1990 and 2006 (millions of observations). The treatment is participation in any employment assistance program during the observed period and the outcome is the employment status in the 36 months following participation. EMC then involves four steps:

1) A propensity score model is estimated based on the entire population using a logistic regression with a linear expit.
2) A random sample of the non-treated individuals is drawn from the population. This means the observed outcome - with an unknown functional relation to the covariates - serves as the value Y_0 .
3) Each individual is assigned a treatment status based on a draw from a Bernoulli distribution parameterized by the propensity score for the individual under the model from step 1.
4) Each individual is subjected to a 'placebo' treatment to construct Y_1. This is treatment with a constant, zero treatment effect.
5) Steps 2 through 4 are repeated for multiple samples.

Table 6 expresses this design using the generic DGP and functional building blocks from above.

Empirical Monte Carlo - Huber et al (2013)

$\rho \leftarrow f_{empirical}$
$\Omega \leftarrow f_{estimate}$
$\Phi \leftarrow f_{empirical}$
$\tau \leftarrow f_{synthetic}$
$X \leftarrow X_{empirical}$
$T \leftarrow \text{LogisticFit}\left(T_{empirical} \sim X_{empirical}\right)$
$Y_0 \leftarrow Y_{empirical}$
$TE \leftarrow 0$
$Y_1 \ \leftarrow Y_0 + TE$
$Y = T \times Y_1 + (1 - T) \times Y_0$

Table 6: Concretization of the generic DGP under the design of Huber, Lechner, and Wunsch, 2013

Analyzing the assignment of the functional degrees of freedom from Table 6 in light of Table 5 makes analyzing the validity properties of this design relatively straightforward. These are listed below with a summary given in Table 9. It is clear from these properties that hybrid designs are promising - the use of data provides realism while the simulated components of the treatment and outcome provide access to ground truth and ensure assumptions are obeyed. However, as is evident from the analysis below, the EMC design provides only partial conformance to some of the desirable properties of validity.

- **Realistic Covariate Distribution - Yes:** The use of observed covariates means the covariate distribution is realistic (in the specific domain of labour data).

- **Realistic Treatment/Outcome Functions - Partial:** The use of the observed outcome mechanism means this too is realistic. However, the treatment assignment and treatment effect are both hand crafted and too simplistic.

- **Guaranteed unbiased specification - No:** The hand selection of the two functions above means there is potential for researcher bias (which holds even if more complex functions are chosen).

- **Access to ground truth - Yes:** the simulated treatment means the ground truth is known at the individual and average effect level.

- **Guaranteed to meet causal assumptions - Yes:** the treatment assignment mechanism is known to include only the observed covariates so we have selection on observables and therefore ignorability. The authors also include an offset in the treatment assignment expit to ensure that the probability of treatment is bounded away from 0 and 1, meaning there is overlap. So both important causal assumptions are met.

- **Allows repeat sampling - Yes:** the large population of observations allows repeat samping as in a synthetic Monte Carlo design:

- **Known distributional setting - Weak:** the joint distribution over the covariates and the outcome is unknown. The treatment assignment mechanism is known.

- **Controllable distributional setting - Weak:** the treatment assignment mechanism can be altered but the joint distribution over the covariates and the outcome mechanism is fixed by the data.

Knaus et al., 2018 build on EMC. They use a very similar dataset as Huber et al., 2013 - but covering Swiss administrative data on labour and employment rather than German. They also share the same logistic estimation for the treatment assignment. But, unlike Huber et al, they specify a non-zero, non-constant treatment effect mechanism. This is specified synthetically rather than estimated because "estimation may favor [functionally] similar estimators under investigation" . It is unclear why the authors are concerned about this problem in outcome mechanism but are ok with fixed functional form estimation in the treatment assignment mechanism. This paper is designed to test methods which infer individual treatment effects, so the authors specify the treatment effect mechanism with the goal of "making disambiguating selection and treatment effect heterogeneity hard" . To this end, they include the propensity score in the synthetic treatment effect, apply a highly non-linear transform, and add random noise. The treatment effect from this function is added to the observed outcome (much like in EMC, only non-treated individuals are selected into samples so the observed outcome equals the untreated potential outcome Y_0). In addition to the synthetic treatment effect, the authors parameterize the population sampler, treatment assignment mechanism and treatment effect function to generate 24 datasets with different sample sizes, treatment effect magnitude, treatment effect heterogeneity, and counts of treated vs control individuals.

Knaus et al (2018)
$\rho \leftarrow f_{empirical}$
$\Omega \leftarrow f_{estimate}$
$\Phi \leftarrow f_{empirical}$
$\tau \leftarrow f_{synthetic}$
$X \leftarrow X_{empirical}$
$T \leftarrow \text{LogisticFit}\left(T_{empirical} \sim X_{empirical}\right)$
$Y_0 \leftarrow Y_{empirical}$
$TE \leftarrow f_{synthetic}$
$Y_1 \ \leftarrow Y_0 + TE$
$Y = T \times Y_1 + (1 - T) \times Y_0$

Table 7: Concretization of the generic DGP under the design of Knaus et al., 2018

The concretized DGP is given in Table 7. The makes the similarity with Huber et al obvious. But the changes do represent a moderate improvement over EMC when examining the validity properties. The outcome mechanism is arguably more realistic as a result of the complex synthetic treatment effect function combined with the empirical control outcome. But, the synthetic component of the specification opens this design to the same criticisms of realism and potential bias as any other synthetic design.

The primary contribution of these authors is the parameterization of the functions to generate datasets with different, well-defined distributional settings. This allows the evaluation of estimators over a partial subspace of all possible distributional settings. These changes are reflected in the 5th column of Table 9 which upgrades the known and controllable distributional setting properties from weak under EMC to partial under Knaus et al.

The two approaches analyzed above use an empirical covariate population and empirical outcome with an estimated treatment assignment mechanism. There is a loosely analogous set of designs which use an empirical treatment and specify a synthetic outcome mechanism. Hill, 2011 develops a method to test their new causal estimator based on non-parametric Bayesian regression. The proposed method uses experimental data from the Infant Health and Development Program (IHDP) - a "randomized experiment which targeted low birth-weight, premature infants and provided the treatment group with high-quality individual care" Brooks-Gunn, Klebanov, and Liaw, 1991. Per Hill, 2011, using an experimental dataset guarantees overlap (and balance) in the covariate distributions and simulating the outcome using only the observed covariates ensures ignorability. The outcome mechanism (the outcome function Φ and treatment effect function τ) are specified by hand - the author chooses two conditions, one linear outcome and constant treatment effect and one non-linear outcome with heterogenous treatment effect. Starting from the position of balanced covariate distribution allows the author to induce imbalance by non-randomly censoring individuals from the treat/control group. This is the first instance of an intentional specification of the level of balance.

The concretization of the generic DGP proposed by Hill, 2011 is summarized in Tbale 8. Ultimately, as a result of the parallel between simulating outcome and treatment, this method has very similar validity properties to those analyzed above. Unlike the methods in Huber et al., 2013 and Knaus et al., 2018, the simulated (untreated) outcome provides greater control over the distribution of outcomes and the experimental datasets allows for the easy control of balance (although a similar censoring method could be applied to non-experimental data). This means the method has arguably better definition and control of the distributional setting but this control is still worse than under purely synthetic treatment assignment and outcome mechanisms.

Hill (2011)

$\rho \leftarrow f_{empirical}$
$\Omega \leftarrow f_{empirical}$
$\Phi \leftarrow f_{synthetic}$
$\tau \leftarrow f_{synthetic}$
$X \leftarrow X_{empirical}$
$T \leftarrow T_{empirical}$
$Y_0 \leftarrow f_{synthetic}$
$TE \leftarrow f_{synthetic}$
$Y_1 \leftarrow Y_0 + TE$
$Y = T \times Y_1 + (1 - T) \times Y_0$

Table 8: Concretization of the generic DGP under the design of Hill, 2011

Two papers iterate on the design proposed by Hill, 2011 but change the specification of the outcome mechanism (the outcome and treatment effect functions) in order to improve the realism of the synthetic outcomes and reduce the potential for bias. Wendling et al., 2018 uses four experimental datasets from the medical field but, instead of manually specifying the outcome and treatment effect functions, estimates these using neural networks. This technically avoids the need to specify a functional form because neural networks can represent an arbitrary set of functional forms. This is a move back toward the estimation used in EMC with $\Phi \leftarrow f_{estimated}$ and $\tau \leftarrow f_{estimated}$. While this may more accurately model the true outcome function, the estimation method could still favor specific estimators - for example those based on neural networks which will more naturally recover the same estimated function from the same data. It also gives up on control over the outcome aspect of the distributional setting because it is not clear what level of interaction/non-linearity is represented by the estimated functions nor how to change these functions to change the level of these aspects of the distribution. The first point results in improved realism of the outcome/treatment but no change to the Guaranteed Unbiased Specification property. The second point results in worsened scores on Known and Controllable distributional setting - see Table 9.

Kern et al., 2016 pursue a different approach - they specify the outcome and treatment effect functions manually but use a “saturated parametric model” with some desired level of nonlinearity and interaction. This means that, for example, all covariates may appear in both linear and quadratic form as well as in every possible (pairwise) interaction with lower/higher powers and interaction levels possible. The motivation for this is to create a principled approach for building complex functions with desired properties but without allowing the function to be hand crafted with arbitrary form and/or modified to produce desired results. This moves towards a design that allows for control over the outcome mechanism while also partially mitigating specification bias. Note that the the full saturation is itself fairly arbitrary (although it is closed to respecification in order to achieve desired results). On balance, the result of this change is a design which appears to have some degree of positive across-the-board ‘performance’ when assessed on the properties of validity - see Table 9. This is not a claim to formal superiority but rather an indication of a promising design direction.

	Pure Designs		Hybrid Designs				
Property	**Synthetic Designs**	**Empirical Designs**	**Huber et al (2013)**	**Knaus et al (2018)**	**Hill (2011)**	**Wendling (2019)**	**Kern et al (2016)**
Realistic Covariate Distribution	No	Yes	Yes	Yes	Yes	Yes	Yes
Realistic Treatment/Outcome Functions	No	Yes	Partial	Partial	Partial	Partial [2]	Partial
Guaranteed Unbiased Specification	No	Yes	No	No	No	No	Partial
Access to Ground Truth	Yes	No	Yes	Yes	Yes	Yes	Yes
Guaranteed to Obey Causal Assumptions	Yes	No	Yes	Yes	Yes	Yes	Yes
Allows Repeat Sampling	Yes	No	Yes	Yes	Yes	Yes	Yes
Known Distributional Setting	Yes	No	Weak	Partial	Partial	Weak	Partial
Controllable Distributional Setting	Yes	No	Weak	Partial	Partial	Weak	Partial

Table 9: Comparison of validity properties across hybrid designs in the literature.

The methods explored up until this point share a few key properties. They all use empirical covariate data from which samples are drawn. In all cases, either the treatment assignment or outcome mechanism is simulated to provide access to ground truth and allow for control over the distributional setting. Some methods, like EMC, trade off control for realism but newer approaches propose designs which allow for some degree of realism and control - either by approximation with non-parametric functions Wendling et al., 2018 or by the principled specification of functions from simpler components Kern et al., 2016. The ideal method would improve on the methods above by improving the realism of the treatment assignment/outcome mechanism while also increasing control over the distributional setting.

Dorie et al., 2019 propose a method which moves toward this ideal, building on the

[2] The semi-parametric neural network estimation of outcome function and empirical treatment assignment makes this approach arguably more realistic, to the extent that one believes that semi-parametric methods are an a priori better method of function estimation.

ideas of Hill, 2011 by simulating functions on top of experimental data and Kern et al., 2016 by providing a principled way to construct functions while maintaining control over desired properties. The authors explicitly set out to create a “testing ground” to help applied researchers choose between different causal estimators - a goal which is shared by this paper. There are two notable design contributions in this paper. First, the authors motivate for the simulation of both the outcome and treatment assignment mechanisms. This allows for control over the full range of distribution settings including “balance, overlap, and nonlinearity” . Second, the authors mitigate the (potential) decrease in realism by employing a meta-sampling approach: The functions for untreated outcome, treatment effect and treatment assignment are sampled randomly from the space of “generalized additive functions” by (stochastically) selecting terms from a set of covariate transform building blocks - polynomial powers, discontinuous-valued ‘steps’, discontinuous-gradient ‘kinks’ - and combining these through the (stochastic) use of the operators addition, exponentiation, and multiplication. By parameterizing the distribution over the base terms and combination operators, the (distribution over) the functional properties of the sampled functions can be modified. This provides a principled way to construct functions with desired properties - as in Kern et al., 2016 - but allows testing over a large number of different functions which are more likely to capture performance on realistic DGPs by, one, sampling from space large enough to contain realistic DGPs and, two, averaging results over many samples from this space to maximize accuracy and remove idiosyncrasies. Combined, this makes for strong realism. Further, the authors parameterize the function selection mechanism to allow control over six “knobs” which cover most of the evaluation space from the previous section. The level of each knob is measured through the use of well-defined metrics. This allows the relatively precise specification of a distributional setting of a simulation in terms of targeted metric values and a corresponding sample of functions. This allows for close control over the distributional setting without giving method evaluators control over the actual functional forms.

The six knobs are:

1) The degree of nonlinearity in treatment assignment/outcome
2) The treatment effect heterogeneity
3) The treatment effect magnitude
4) The percentage treated
5) The degree of treat/control group covariate overlap
6) The degree of alignment between treatment and control mechanism

The design of the evaluation method is presented in Table 10 below. Note that there are effectively two Monte Carlo processes happening in this design. The first is a sample of DGPs from the space of DGPs defined by the function sampling parameters. The second is a sample of the data defined by the sampled DGP. The authors do not run repeat sampling at the data level, instead conflating both sampling processes into a single sample run repeatedly. As pointed out by the authors, this should have no effect on the convergence of the results to the true sampling distribution of the targeted estimand.

Dorie et al (2019)

$\rho \leftarrow f_{empirical}$
$\Omega \leftarrow f_{synthetic} \ \sim P\left(...\right)$
$\Phi \leftarrow f_{synthetic} \ \sim P\left(...\right)$
$\tau \leftarrow f_{synthetic} \ \sim P\left(...\right)$
$X \leftarrow X_{empirical}$
$T \leftarrow f_{synthetic}$
$Y_0 \leftarrow f_{synthetic}$
$TE \leftarrow f_{synthetic}$
$Y_1 \ \leftarrow Y_0 + TE$
$Y = T \times Y_1 + (1 - T) \times Y_0$

Table 10: Concretization of the generic DGP under the design of Dorie, Hill, Shalit, Scott, and Cervone, 2019

Based on the properties above, it appears that the design proposed by Dorie et al., 2019 has all of the desired validity properties - although the degree of validity may be contingent on design details like the size of the function space. Table 11 compares the validity properties of the three methods based on experimental data and principled function construction.

	Pure Designs		**Hybrid Designs**		
Property	**Synthetic Designs**	**Empirical Designs**	**Hill (2011)**	**Kern et al (2016)**	**Dorie et al (2019)**
Realistic Covariate Distribution	No	Yes	Yes	Yes	Yes
Realistic Treatment/Outcome Functions	No	Yes	Partial	Partial	Yes [3]
Guaranteed Unbiased Specification	No	Yes	No	Partial	Yes
Access to Ground Truth	Yes	No	Yes	Yes	Yes
Guaranteed to Obey Causal Assumptions	Yes	No	Yes	Yes	Yes
Allows Repeat Sampling	Yes	No	Yes	Yes	Yes
Known Distributional Setting	Yes	No	Partial	Partial	Yes
Controllable Distributional Setting	Yes	No	Partial	Partial	Yes

Table 11: Comparison of validity properties across designs based on experimental data and principled function construction.

4.4 Conclusion

Hybrid Monte Carlo evaluation designs - which combine empirical data and simulated functional forms - provide the foundation for evaluation methods with both specific and general validity. Not all hybrid designs are optimal. Many of the proposed designs - like those by Huber et al., 2013, Knaus et al., 2018, Hill, 2011, Wendling et al., 2018, Kern et al., 2016 - do allow for some degree of specific and general validity but are still subject to criticism on both axes. Dorie et al., 2019 provide a sampling-based approach which appears to establish a near optimal middle ground in the design space by allowing careful control over the distributional while using leveraging the Monte Carlo sampling of designs to improve realism and mitigate bias.

[3] Provided the space of functions which can be sampled is large enough to contain realistic functions and enough samples are taken to allow for effective averaging.

Chapter 5

Synthesized Hybrid Benchmarking Design

PENDING: this chapter will outline a design which builds on the methods reviewed in the previous chapter but taking the best (compatible) parts of each.

Part IV

Introducing Maccabee

Chapter 6

Design Overview

PENDING: this chapter will outline the technical design of the Maccabee package which implements the procedural design from Chapter 5. For now, I have included a data generation demonstration notebook (originally submitted with my individual deliverable) which shows the DGP sampling process and verifies that the user specified parameters have the desired effect on the distributional setting of the sampled data.
Some things to consider including
- Aggregation in the benchmarking section - Abstract Syntax Trees for equation construction - Parallelism - Good OOP practices throughout.

Bibliography

Abadie, A., & Imbens, G. W. [G W]. (2002). Large Sample Properties of Matching Estimators for Average Treatment Effects. *October*, *0136789*(October), 146–146. doi:10.3386/t0283

Abadie, A., & Imbens, G. W. [Guido W.]. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, *74*(1), 235–267. doi:10.1111/j.1468-0262.2006.00655.x

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7353–7360. doi:10.1073/pnas.1510489113

Athey, S., & Imbens, G. W. [Guido W.]. (2017). The state of applied econometrics: Causality and policy evaluation. In *Journal of economic perspectives* (Vol. 31, *2*, pp. 3–32). doi:10.1257/jep.31.2.3

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *80*(4), 597–623. doi:10.1111/rssb.12268

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1179–1203. doi:10.1214/18-AOS1709

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. doi:10.1080/00273171.2011.568786

Bishop, C. M. (2006). Pattern recognition and machine learning.

Brooks-Gunn, J., Klebanov, P. K., & Liaw, F. r. (1991). The learning, physical, and emotional environment of the home in the context of poverty: The infant health and development program. *Children and Youth Services Review*, *17*(1-2), 251–276. doi:10.1016/0190-7409(95)00011-Z

Broomberg, J. (2017). *Deep Causal Inference.*

Busso, M., Dinardo, J., & Mccrary, J. (2014). NEW EVIDENCE ON THE FINITE SAMPLE PROPERTIES OF PROPENSITY SCORE REWEIGHTING AND MATCHING ESTIMATORS. doi:10.1162/REST{_}a{_}00431

Calder, K. (1953). *Statistical Inference.* Retrieved from https://www.asc.ohio-state.edu/calder.13/stat528/Lectures/lecture21_6slides.PDF

Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*(4), 1199–1209. doi:10.1111/biom.12676

Dehejia, R. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics*, *125*, 355–364. doi:10.1016/j.jeconom.2004.04.012

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062. doi:10.1080/01621459.1999.10473858

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. doi:10.1162/003465302317331982

Diamond, A., Sekhon, J. S., Abadie, A., Brady, H., Caughey, D., Dehejia, R., ... Todd, P. (2012). *Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies*. Retrieved from http://sekhon.berkeley.edu/,

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition 1. *Statistical Science*, *34*(1), 43–68. doi:10.1214/18-STS667

Flores, C., & Mitnik, O. (2009). Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from Experimental Data.

Fraker, T., & Maynard, R. (1987). The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *The Journal of Human Resources*, *22*(2), 194. doi:10.2307/145902

Friedlander, D., & Robins, P. K. (1995). Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *American Economic Review*, *85*(4), 923–37.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. In *Review of economics and statistics* (Vol. 86, *1*, pp. 77–90). doi:10.1162/003465304323023697

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*. doi:10.1093/biomet/57.1.97

Heckman, J. J., & Todd, P. (1998). *Matching As An Econometric Evaluation Estimator*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. doi:10.1198/jcgs.2010.08162

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2005). A Comparison of Experimental and Observational Data Analyses. (pp. 49–60). doi:10.1002/0470090456.ch5

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. doi:10.1111/1468-0262.00442

Holland, P. W. (1986). *Statistics and Causal Inference* (tech. rep. No. 396).

Horvitz, D. G., & Thompson, D. J. (1952). *A Generalization of Sampling Without Replacement From a Finite Universe* (tech. rep. No. 260).

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*, 1–21. doi:10.1016/j.jeconom.2012.11.006

Imai, K., King, G., & Stuart, E. A. (2008). *Misunderstandings between experimentalists and observationalists about causal inference* (tech. rep. No. 2).

Imbens, G. W. [Guido W.], & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. doi:10.1257/jel.47.1.5

Johansson, F. D. [Fredrik D.], Kallus, N., Shalit, U., & Sontag, D. (2018). Learning Weighted Representations for Generalization Across Designs. Retrieved from http://arxiv.org/abs/1802.08598

Johansson, F. D. [Fredrik D], Shalit, U., & Sontag, D. (2016). *Learning Representations for Counterfactual Inference*.

Kallus, N. (2016). A Framework for Optimal Matching for Causal Inference. Retrieved from http://arxiv.org/abs/1606.05188

Kang, J. D. Y., & Schafer, J. L. [Joseph L]. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data 1. *Statistical Science*, *22*(4), 523–539. doi:10.1214/07-STS227

Kang, J. D., & Schafer, J. L. [Joseph L.]. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539. doi:10.1214/07-STS227

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127. doi:10.1080/19345747.2015.1060282

Knaus, M. C., Lechner, M., Strittmatter, A., Graham, B., Santos, A., Schuler, A., . . . Zimmert, M. (2018). *Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence.*

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10). doi:10.1073/pnas.1804597116

Künzel, S. R., Stadie, B. C., Vemuri, N., Ramakrishnan, V., Sekhon, J. S., & Abbeel, P. (2018). Transfer Learning for Estimating Causal Effects using Neural Networks. Retrieved from http://arxiv.org/abs/1808.07804

Lalonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *Journal of Chemical Information and Modeling*, *76*(4), 604–620. doi:10.1017/CBO9781107415324.004

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. doi:10.1038/nature14539

Li, S., & Fu, Y. (2017). Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in neural information processing systems* (Vol. 2017-Decem, pp. 930–940).

Lu, M., Sadiq, S., Feaster, D. J., & Ishwaran, H. (2018). Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *Journal of Computational and Graphical Statistics*, *27*(1), 209–219. doi:10.1080/10618600.2017.1356325

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. doi:10.1016/S0889-8588(05)70309-9

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling*, *8*(2), 287–312. doi:10.1207/S15328007SEM0802{_}7

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. doi:10.1214/09-SS057

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, *39*(2), 1180–1210. doi:10.1214/10-aos864

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129. doi:10.1080/01621459.1995.10476494

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi:10.1093/biomet/70.1.41

Rothwell, P. M. (2006). Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLoS Clinical Trials*, *1*(1), e9. doi:10.1371/journal.pctr.0010009

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585. doi:10.1080/01621459.2000.10474233

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models: Rejoinder. *Journal of the American Statistical Association*, *94*(448), 1135. doi:10.2307/2669930

Schölkopf, B. (2019). Causality for Machine Learning. Retrieved from http://arxiv.org/abs/1911.10500

Schwab, P., Linhardt, L., & Karlen, W. (2018). Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. Retrieved from http://arxiv.org/abs/1810.00656

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1343. doi:10.1198/016214508000000733

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1-2), 305–353.

Su, X., Edu, C., Wang, H., Nickerson, D. M., & Li, B. (2009). *Subgroup Analysis via Recursive Partitioning Chih-Ling Tsai.*

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business and Economic Statistics*, *34*(4), 661–672. doi:10.1080/07350015.2016.1172013

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532. doi:10.1080/01621459.2014.951443

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. doi:10.1080/01621459.2017.1319839

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, *37*(23), 3309–3324. doi:10.1002/sim.7820