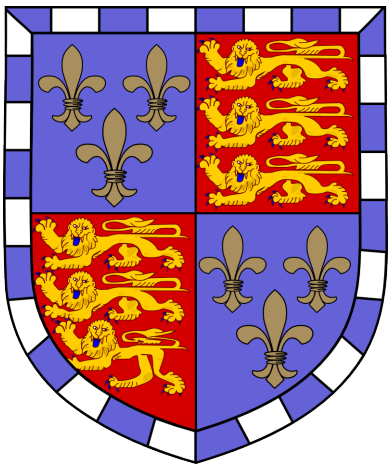
Topics in conditional causal inference



Yao Zhang

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College

Jan 2022

To my parents.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit for the Mathematics Degree Committee.

Chapter 2 is joint work with Qingyuan Zhao (University of Cambridge), and has been submitted for publication as Zhang and Zhao (2021) and Zhang and Zhao (2022). Chapter 3 is joint work with Jeroen Berrevoets (University of Cambridge) and Mihaela van der Schaar (University of Cambridge), and has been published in AISTATS 2022 as Zhang et al. (2022). Chapter 4 is joint work with Alexis Bellot (University of Cambridge) and Mihaela van der Schaar, and has been published in AISTATS 2020 as Zhang et al. (2020). Chapter 5 is joint work with Hyun-Suk Lee (Sejong University), William Zame (UCLA), Cong Shen (University of Virginia), Jang-Won Lee (Yonsei University) and Mihaela van der Schaar, and has been published in NeurIPS 2020 as Lee et al. (2020).

Yao Zhang
Jan 2022

Abstract

Topics in conditional causal inference **Yao Zhang**

With the growth of complex experimental designs and large-scale observational data, causal questions arising in applications are now more targeted and precise. For example, one might ask if the treatment is effective at a particular time point, or if the treatment is effective for a particular individual. To answer many questions of this kind, this thesis concerns conditional causal inference, generally referring to techniques of constructing or interpreting conditional distributions or expectations for inference about a causal effect of interest. This thesis consists of five chapters. In Chapter 1, we first review the classical potential outcomes framework and some basic causal inference methods relevant to this thesis, then provide a summary of the problems and methods studied in the following chapters.

In Chapter 2, we consider testing causal effects in complex experimental designs via *conditional randomization tests* (CRTs). The CRTs we define are randomization tests conditioning on a subset of treatment assignments tailored to the effect of interest. Because many potential outcomes are missing in complex designs, a single CRT is rarely powerful. We develop a general theory for constructing multiple jointly valid CRTs in arbitrary designs. Following this theory, we propose practical methods that can collect and combine statistical evidence in different parts of an experiment to test a global effect of interest. Under a general framework of CRTs, we connect and discuss randomization tests developed for different statistical problems in the literature, which may be of independent interest.

The following three chapters concern the problem of estimating *conditional average treatment effects* (CATEs). CATEs quantify individual-level treatment effects by conditioning on individual covariates. In Chapter 3, we consider estimating CATEs in the presence of high-dimensional covariates. We propose a neural network-based dimensionality reduction method that can transform high-dimensional covariates into a low-dimensional and informative representation. Neural network models are overparameterized and non-convex. We propose a sample-splitting and randomization method that enables the representation to be partially identifiable and converge consistently. In Chapter 4, we consider estimating CATEs in

the presence of imbalanced treated and control populations. We take a Bayesian method to measure the overlap between the two populations and rebalance the populations by minimizing the posterior variances of counterfactual outcomes. We propose a PAC-Bayes generalization bound to show that this method is beneficial and consistent in estimating CATEs. In Chapter 5, we introduce a recursive partitioning method that can convert any black-box CATE estimates into interpretable subgroups. Our method uses a distribution-free technique called conformal prediction to quantify the uncertainties in CATE estimates, then leverage the uncertainties to construct robust subgroups. It leads to more well-identified subgroups and fewer false discoveries due to random noise in the data.

All the methods proposed in this thesis are tested using multiple simulations or datasets. Overall, experimental results support our theories and demonstrate the advantages of our methods compared with some baseline methods.

Acknowledgements

First, I would like to thank my advisor, Mihaela van der Schaar, for her guidance and support in the last three years. I admire her unwavering enthusiasm, diverse research interests and ideas in machine learning and its applications in medicine. I thank her for giving me the freedom to pursue my own research ideas. I am deeply grateful for her insights, generosity and encouragement throughout my PhD. Her creativity and vision of research have and will always be a great source of inspiration for me.

Next, my thank goes to Qingyuan Zhao for his advice and our collaboration. My research has and will continue to benefit a lot from his comprehensive course on causal inference. Throughout many discussions we had over the last few years, his wisdom and insights in statistics have made a profound impact on my development as a researcher.

My thank also goes to my coauthors: Daniel Jarrett, Alexis Bellot, Hyun-Suk Lee, Jeroen Berrevoets, Zhaozhi Qian, Jonathan Crabbé at the University of Cambridge; James Jordon and Ioana Bica at the University of Oxford; Trent Kyono, Ahmed Alaa, Jinsung Yoon and William Zame at UCLA. It has been a great honour and pure enjoyment to work with them. I would also like to thank my colleagues Changhee Lee, Alihan Huyuk, Nick Maxfield, Alicia Curth, Fergus Imrie, Yuchao Qin, Boris van Breugel, Alex Chan, Nabeel Seedat and Sam Holt. My research has benefited from their invaluable feedback.

I thank my examiners Rajen Shah and Chengchun Shi for their careful reading, stimulating discussion and useful comments on this thesis.

Finally, I thank my parents for their love and support over many years. No words can express how grateful I am towards them. My deepest gratitude goes to Yunyun for her encouragement, trust and love. This thesis would not have been possible without her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is built upon the classical potential outcomes framework for causal inference. The framework is often attributed to the seminal work by Rubin (1974) who used potential outcomes to define the causal effect of an educational treatment. Rubin (1990) acknowledged that the framework can be dated back to Neyman (1923) who adopted the language of potential outcomes to study agricultural field experiments. Nevertheless, Rubin (1974)'s work changed the long-standing convention of analyzing observational data with the notation purely in terms of observed outcomes. Ever since then, the framework has been popularized and is now widely applied for causal inference in social and biomedical sciences. This chapter starts by introducing the framework and reviewing the basic methods that are relevant to our works in this thesis. It ends by summarizing the methodological contributions and mathematical notation in the following chapters.

1.1 Potential outcomes

Drawing causal conclusions is often an informal task in our daily life. For example, one might wonder if going to college gets us a better job, or if taking a painkiller reduces our toothache. In the potential outcomes framework, establishing causality is about estimating the effect of an *action* (e.g. a medical treatment, or an educational campaign) on the outcome of a *unit* (e.g. an individual, or a school). Consider a unit i, we define its treatment variable $A_i \in \mathcal{A} = \{0, 1\}$ ($A_i = 1$: treated, $A_i = 0$: control) and an outcome variable $Y_i \in \mathcal{Y}$. Let $Y_i(1)$ be the outcome variable that would have been observed if unit i is treated ($A_i = 1$), and $Y_i(0)$ be the outcome variable that would have been observed if unit i is control ($A_i = 0$). The random variables $Y_i(0)$ and $Y_i(1)$ are referred to as *potential outcomes* of unit i, which emphasize the fact that depending on whether unit i is treated or not, either of these two

outcome variables can be potentially observed. The definition of potential outcomes seems straightforward but makes two well-known assumptions in the literature.

Consider a finite population of N units, $i \in \{1, \dots\} = [N]$. Defining potential outcomes should take into account interference between the units. For example, carrying out an educational campaign in one school (unit) may also affect the other $N-1$ schools (units) in the same city. Then, every school i's outcome Y_i depends on the treatment variables of all the schools, $\boldsymbol{A} = (A_j : j \in [N])$ with A_j indicating if the campaign is given in school j; every school i has 2^N potential outcomes, $Y_i(\boldsymbol{a}), \forall \boldsymbol{a} \in \mathcal{A}^N = \{0,1\}^N$. To reduce the number of unobserved outcomes, we often make the following assumption.

Assumption 1 (No interference)**.** $Y_i(\boldsymbol{a}) = Y_i(a_i)$ for any $\boldsymbol{a} \in \mathcal{A}^N$ and $i \in [N]$.

Under no interference, every school i has only two potential outcomes, $Y_i(a_i), a_i = 0, 1$. Causal inference methods that allow for interference (Hudgens and Halloran, 2008; Rosenbaum, 2007; Tchetgen and VanderWeele, 2012) often assume a known and local interference structure between the units, then attempt to estimate the direct effect from a unit's treatment variable and the indirect effect from its neighbour units' treatment variables.

To make sure that the observed outcomes are consistent with the defined potential outcomes, the next assumption formalizes the connection between them.

Assumption 2 (Consistency)**.** $Y_i = Y_i(\boldsymbol{a})$ if $\boldsymbol{A} = \boldsymbol{a}$ for any $\boldsymbol{a} \in \mathcal{A}^N$ and $i \in [N]$.

The consistency assumption implies that if the treatment variable is continuous, we may have no observation of potential outcomes at some treatment levels in finite samples. A naive solution to this problem is by converting the continuous treatment into a binary one indicating if a unit is treated or not. For example, suppose that in a clinical trial, a group of patients have taken different doses of an experimental drug while another group of patients have taken a placebo. One may consider estimating the drug effect by comparing the groups in terms of average outcomes. This comparison assumes that the drug effect does not depend on the dose, i.e., the potential outcomes at any non-zero dose levels are the same, which is not part of the consistency assumption above. If the drug effect is only realized at large doses, this approach may underestimate or fail to detect the drug effect. A better estimation strategy is to leverage the smoothness of the drug effect at different dose levels. This can be done by applying causal inference methods adapted for continuous treatment variables (Gill and Robins, 2001; Hirano and Imbens, 2004; Imai and Van Dyk, 2004; Kennedy et al., 2017).

In the literature, authors often make the stable unit treatment value assumption (SUTVA) (Rubin, 1980b), which essentially refers to Assumptions 1 and 2 taken together.

Assumption 3 (SUTVA)**.** $Y_i = Y_i(a)$ if $A_i = a$ for any $a \in \mathcal{A}$ and $i \in [N]$.

Under SUTVA, causal inference problems become more tractable since we will observe either $Y_i(0)$ or $Y_i(1)$ for all $i \in [N]$ given any $\boldsymbol{A} \in \mathcal{A}^N$. Then we can formally call $Y_i(0)$ and $Y_i(1)$ potential outcomes, because are indeed potentially observed. In this thesis, we also call the observed potential outcome factual and the unobserved potential outcome counterfactual, respectively. Under SUTVA, the factual outcome is $Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A)Y_i(0)$, while the counterfactual outcome is $Y_i(1 - A_i) = (1 - A_i)Y_i(1) + A_i Y_i(0)$.

Given the treatment variables $\boldsymbol{A}$ and outcome variables $\boldsymbol{Y} = (Y_i : i \in [N])$, one may consider estimating the *individual treatment effect* of each unit i as the difference $\tau_i = Y_i(1) - Y_i(0)$. However, τ_i is never observed and non-estimable using $\boldsymbol{A}$ and $\boldsymbol{Y}$, which is referred to as the fundamental problem of causal inference (Holland, 1986) and motivates researchers to start by looking at constant or average effects at the population level.

1.2 Randomization inference

Echoing Freedman (2006)'s quote "Experiments should be analyzed as experiments, not as observational studies", we stress that causal inference methods developed for analyzing experimental and observational data have one fundamental difference in their setups. In experimental studies, the *treatment assignment mechanism* (i.e. the joint distribution of $\boldsymbol{A} = (A_1, \ldots, A_N)$) is known from the experimenter, but $A_i, \ldots, A_N$ are often not sampled independently. In observational studies, the treatment assignment mechanism is unknown so we often assume that $A_1, \ldots, A_N$ are independent and identically distributed (i.i.d) samples from an unknown distribution. When we analyze experimental data with a model, we should be mindful of whether the model assumptions are satisfied or not. For example, the treatment variable A_i can be correlated with the error variable in a linear regression model (Freedman, 2008a,b). Next, we review a model-free inference method called *Fisher randomization test* (FRT). It highlights the fact that causal or statistical inference on experimental data can be based on randomization and nothing more than randomization.

1.2.1 Fisher randomization test

Given $\boldsymbol{A}$ and $\boldsymbol{Y}$ from a treated-versus-control experiment, Fisher (1935) proposes to test the *sharp null hypothesis* of constant treatment effect on all the units, that is

$$H_{\text{Fisher}} : Y_i(1) = Y_i(0) + \Delta, \forall i \in [N].$$

Under SUTVA, we let $\boldsymbol{Y}(\boldsymbol{a}) = (Y_i(a_i) : i \in [N])$. The key observation is that under Fisher's null H_{Fisher}, we can observe or impute all the potential outcomes

$$\boldsymbol{Y}(\boldsymbol{a}) = \boldsymbol{Y} + \Delta(\boldsymbol{a} - \boldsymbol{A}),\ \forall \boldsymbol{a} \in \mathcal{A}^N.$$

Let Ω be the support of the assigment mechanism $\mathbb{P}(\boldsymbol{A})$. The experimenter randomizes $\boldsymbol{A}$ without knowing any potential outcomes $\boldsymbol{Y}(\boldsymbol{a}), \boldsymbol{a} \in \Omega$, so the following assumption holds.

Assumption 4 (Randomization)**.** $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{Y}(\boldsymbol{a})$ for any $\boldsymbol{a} \in \Omega$.

Using nothing but the act of physical randomization in the experiment, the classical Fisher randomization test (FRT) computes the p-value as

$$P(\boldsymbol{A}, \boldsymbol{Y}) = \sum_{\boldsymbol{a}^* \in \Omega} \mathbb{P}(\boldsymbol{a}^*) 1\{T(\boldsymbol{a}^*, \boldsymbol{Y}) \geq T(\boldsymbol{A}, \boldsymbol{Y})\},$$

where the function $T : \mathcal{A}^N \times \mathcal{Y}^N \to \mathbb{R}$ is a chosen test statistics, e.g., the Wilcoxon's rank-sum statistic $T(\boldsymbol{A}, \boldsymbol{Y}) = \sum_{i=1}^N \sum_{j=i+1}^N 1\{A_i > A_j\} \cdot 1\{Y_i > Y_j\}$. FRT compares the statistics under the observed assignment $\boldsymbol{A}$ with the statistics under any $\boldsymbol{a}^* \in \Omega$, which is essentially all the assignments that can possibly happen in the experiment provided that we can rerun the experiment for many times. If the observed assignment has a larger statistics than $(1-\alpha)\%$ of the assignments $\boldsymbol{a}^* \in \Omega$, the p-value $P(\boldsymbol{A}, \boldsymbol{Y})$ is smaller than α hence we reject the Fisher's null H_{Fisher}. It is well-known that FRT controls the type I error in finite samples such that $\mathbb{P}\{P(\boldsymbol{A}, \boldsymbol{Y}) \leq \alpha\} \leq \alpha$, under SUTVA, Assumption 4 (Randomization)[1] and H_{Fisher}.

1.2.2 Discussions

Different from Fisher's idea, Neyman (1923)'s analysis of randomized experiments is through testing the sample-average treatment effect (SATE), $N^{-1}\sum_{i=1}^N Y_i(1) - N^{-1}\sum_{i=1}^N Y_i(0)$. Neyman's test for the null hypothesis $H_{\text{Neyman}} : \text{SATE} = \Delta$, is based on the asymptotic analysis of the SATE estimator,

$$\hat{\tau}_{\text{Neyman}} = \frac{\sum_{i=1}^N A_i Y_i}{\sum_{i=1}^N A_i} - \frac{\sum_{j=1}^N (1-A_j) Y_j}{\sum_{j=1}^N (1-A_j)},$$

under the distribution $\mathbb{P}(\boldsymbol{A})$ and the assumption that $Y_i(0), Y_i(1), i \in [N]$ are i.i.d samples from an unknown bivariate distribution. The central limit theorem ensures that $\hat{\tau}_{\text{Neyman}} \xrightarrow{d} \mathcal{N}(\text{SATE}, \mathbb{V}_{\text{Neyman}})$. Using $\hat{\tau}_{\text{Neyman}}$ and a sample-variance estimator $\hat{\mathbb{V}}_{\text{Neyman}}$,

[1] By the independence in the randomization assumption, we can marginalize out the potential outcomes in $\mathbb{P}\{P(\boldsymbol{A}, \boldsymbol{Y}) \leq \alpha \mid \boldsymbol{Y}(\boldsymbol{a}), \forall \boldsymbol{a} \in \Omega\} \leq \alpha$ and obtain the unconditional inequity $\mathbb{P}\{P(\boldsymbol{A}, \boldsymbol{Y}) \leq \alpha\} \leq \alpha$.

we can implement a Student's t-test for Neyman's null H_{Neyman}. We next give an up-to-date discussion of the pros and cons of FRT (in comparison to Neyman's test).

FRT is distribution-free, i.e., nonparametric. Neither does it make any assumption on the potential outcomes $Y_i(0)$ and $Y_i(1)$, $\forall i \in [N]$, nor does it assume the individuals are drawn from an imaginary distribution. FRT's validity is not based on a large-sample normal approximation as in Neyman's test, so FRT is applicable even when $\hat{\tau}_{\text{Neyman}}$ fails to be asymptotically normal. FRT is flexible in the use of test statistics. For example, when we have covariates information in a study, we can let the statistics be a parameter from a (semi-)parametric model. When there are outliers in the data, we can use the rank statistics rather than the difference-in-means statistics. Statistics that can discriminate between the observed assignment and the other assignments will potentially improve the test power.

FRT can be inverted to construct a finite-sample valid confidence interval of a treatment effect if the effect is constant and additive; see Section 2.9, Imbens and Rubin (2015, Chapters 5.7) or Ernst (2004, Section 3.4) for more details. Assuming the effect parameter is constant and additive is often part of the assumption of a linear parametric model. The confidence interval from FRT is valid under a weaker assumption by only requiring the treatment effect to be constant and additive. Even compared with a semiparametric partially linear model, the validity of FRT is finite-sample hence more attractive to practitioners. One of our simulations in the next chapter demonstrates this advantage of our proposed conditional randomization tests in comparison to linear mixed-effects models; see Section 2.5.2 for more details.

FRT is often impossible to be implemented exactly due to the large number of assignments in Ω. In practice, we often compute the p-value using a Monte Carlo approximation with a large number of assignments randomly drawn from $\mathbb{P}(\boldsymbol{A})$. This idea is first proposed by Dwass (1957). FRT loses its exactness when using Monte Carlo, but it is often close enough to be considered as an exact test (Lunneborg, 2000). Jockel (1986) derives finite-sample bounds for the power of the Monte Carlo version of FRT, which is useful to determine the required simulation effort for approximating the original FRT. He also shows a lower bound for the asymptotic efficiency loss in a FRT as a function of the number of random assignments, which indicates that the efficiency loss decreases as the number of assignments increases.

The subtle difference between Fisher's and Neyman's nulls have confused both theoretical and practical statisticians. In theory, Fisher's null implies Neyman's null. But in simulations, Ding (2017) demonstrates an intriguing paradox that a rejection of Neyman's null does not imply a rejection of Fisher's null in many realistic situations. This paradox is explained by Loh et al. (2017) through the fact that Neyman's test is anti-conservative (i.e. the rejection probability of a α-level Neyman's test is larger than α) under Fisher's null in finite samples. Thus, the paradox does not exist in large samples. By some careful constructions (e.g.

prepivoting) of the test statistics, Cohen and Fogarty (2021); Fogarty (2021); Zhao and Ding (2021) show that FRT is asymptotically valid for testing Neyman's null, and Caughey et al. (2021) show that FRT can be applied to test null hypotheses of bounded treatment effects and quantiles of individual treatment effects. These recent developments encourage broader applications of FRTs beyond the scope of constant treatment effects.

1.3 Treatment effects estimation

We next review the methods for estimating (conditional) average treatment effects from observational data. Treatment effects are defined under the distribution of potential outcomes. We start by introducing the assumptions for identifying treatment effects, i.e., expressing the treatment effects as functionals of some observable distributions.

Besides from the observed vectors $\boldsymbol{A}$ and $\boldsymbol{Y}$ mentioned above, we now assume that we also observe a set of covariates $X_i = (X_{i1}, \ldots, X_{id}) \in \mathcal{X}$ for every unit i. Let $\boldsymbol{X} = (X_i : i = [N])$. Let $\boldsymbol{O} = (\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}) = \{O_i = (X_i, A_i, Y_i)\}_{i=1}^N$ be our observed data. We assume that the units are random samples from an infinite population called *super-population*. In order words, $(X_i, Y_i(0), Y_i(1)), i \in [N]$, are i.i.d samples from an unknown distribution $\mathbb{P}[X, Y(0), Y(1)]$.

The average treatment effect (ATE) at the super-population level is defined as

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]. \tag{1.1}$$

The conditional average treatment effect (CATE), also known as *individualized treatment effects*, is defined as

$$\text{CATE}(X) = \mathbb{E}[Y(1) - Y(0) \mid X], \tag{1.2}$$

which individualizes the treatment effect by conditioning on a unit's covariates.

Both ATE and CATE are defined under $\mathbb{P}[X, Y(0), Y(1)]$, which is not our observed data distribution. A general treatment assignment mechanism in observational studies is given by

$$\mathbb{P}(\boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{0}), \boldsymbol{Y}(\boldsymbol{1})).$$

One might wonder why $\boldsymbol{A}$ depends on the partially observed $\boldsymbol{Y}(\boldsymbol{0})$ and $\boldsymbol{Y}(\boldsymbol{1})$. The answer is that there may exist unobserved covariates that are confounders of A, $Y(0)$ and $Y(1)$ but not included in X. We often assume that there is no unmeasured confounder, i.e., the counterfactual outcomes are strongly ignorable (Rosenbaum and Rubin, 1983) such $\mathbb{P}(\boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{0}), \boldsymbol{Y}(\boldsymbol{1})) = \mathbb{P}(\boldsymbol{A} \mid \boldsymbol{X})$. In general, we make the following assumption.

Assumption 5 (Unconfoundedness)**.** $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{Y}(\boldsymbol{a}) \mid \boldsymbol{X}$ for any $\boldsymbol{a} \in \mathcal{A}^N$.

The unconfoundedness assumption is untestable. By contrast, randomized experiments have a known assignment mechanism $\mathbb{P}(\boldsymbol{A})$ which satisfies Assumption 4 automatically. This is one of the main reasons why randomized experiments remain the gold standard of causal inference. In observational studies, $\boldsymbol{A}$ is randomized by an unknown conditional distribution. Researchers may concern that the unconfoundedness assumption fails to hold in practice, especially when a known confounder of $\boldsymbol{Y}(\boldsymbol{0})$, $\boldsymbol{Y}(\boldsymbol{1})$ and $\boldsymbol{A}$ is not included in X. For example, age is an important confounder that determines both the illness ($\boldsymbol{Y}(\boldsymbol{0})$ and $\boldsymbol{Y}(\boldsymbol{1})$) of individuals infected with covid and the assignment of covid vaccines ($\boldsymbol{A}$). Then the assumption fails to hold if the age variable is not included in X.

In this thesis, we mainly study causal inference in either experimental studies or observational studies under the unconfoundedness assumption. Here we refer to two classes of methods that can deal with unmeasured confounding under various assumptions. First, sensitivity analysis methods (Cornfield et al., 1959; Imbens, 2003; Rosenbaum, 1987; VanderWeele and Ding, 2017) can validate how robust a treatment effect estimate is by violating the assumption under some postulated models of unobserved confounders. The second class of methods attempts to remove the effects from unobserved confounders with the help of some additional variables, e.g., instrumental variable (Baiocchi et al., 2014), synthetic control (Abadie et al., 2010) and negative controls (Lipsitch et al., 2010). A more unifying framework called proximal causal inference is proposed by Tchetgen et al. (2020) recently.

To estimate the expected outcome in ATE or CATE, we also need the following assumption.

Assumption 6 (Positivity)**.** $\mathbb{P}(A = a \mid X = x) \in (0, 1), \forall a \in \mathcal{A}$ and $x \in \mathcal{X}$.

Under Assumptions 3 (SUTVA), 5 (Unconfoundedness) and 6 (Positivity), we can write

$$\mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y(a) \mid X = x, A = a] = \mathbb{E}[Y \mid X = x, A = a], \text{ and} \tag{1.3}$$

$$\mathbb{E}[Y(a)] = \mathbb{E}\big[\mathbb{E}(Y \mid X = x, A = a)\big]. \tag{1.4}$$

The positivity assumption ensures that the expectation $\mathbb{E}[Y(a) \mid X = x, A = a]$ is well defined and the distribution of $X \mid A = a$ has the same support for all $a \in \mathcal{A}$. Finally, we assume that $O_i = (X_i, A_i, Y_i), i \in [N]$, are i.i.d samples from an *observed data distribution* $\mathbb{P}(O) = \mathbb{P}(Y \mid X, A)\mathbb{P}(A \mid X)\mathbb{P}(X)$, where $\mathbb{P}(Y \mid X, A)$ is connected with the *data distribution* $\mathbb{P}[X, Y(0), Y(1)]$ above by $Y = AY(1) + (1 - A)Y(0)$ under Assumption 3 (SUTVA).

1.3.1 Average treatment effect

Outcome regression (OR) estimator. Following (1.4), we can write the ATE in (1.1) as

$$\text{ATE} = \mathbb{E}\left[\mu_1(X) - \mu_0(X)\right], \tag{1.5}$$

where $\mu_a(X) = \mathbb{E}(Y \mid X = x, A = a), a = 0, 1$. Suppose that we estimate $\mu_a(X)$ by a regression model $\hat{\mu}_a(\cdot)$ fitted to the data $\{(X_i, Y_i) : A_i = a\}$. The outcome regression (OR) estimator of ATE is given by

$$\hat{\tau}_{\text{OR}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_1(X_i) - \frac{1}{N}\sum_{j=1}^{N}\hat{\mu}_0(X_j).$$

Inverse probability weighted (IPW) estimator. We define the *propensity score* as the probability $e(X) = \mathbb{P}(A = 1 \mid X) = \mathbb{E}(A \mid X)$. We can write $\mathbb{E}\left[Y(1)\right]$ as

$$\mathbb{E}\left[\frac{Y(1)}{e(X)}\mathbb{E}(A \mid X)\right] = \mathbb{E}\left[\frac{Y(1)}{e(X)}\mathbb{E}(A \mid Y(1), X)\right] = \mathbb{E}\left[\mathbb{E}\left(\frac{AY(1)}{e(X)} \mid Y(1), X\right)\right] = \mathbb{E}\left[\frac{AY}{e(X)}\right].$$

The first equality is achieved by Assumption 5 (Unconfoundedness). Since A is a Bernoulli random variable conditional on X, $A^2 = A$ and $A(1-A) = 0$. The last equality is due to the fact that $AY(1) = A^2Y(1) + A(1-A)Y(0) = A[AY(1) + (1-A)Y(0)] = AY$ under Assumption 3 (SUTVA). Then, it follows from the last equation that

$$\text{ATE} = \mathbb{E}\left[\frac{AY}{e(X)} - \frac{(1-A)Y}{1-e(X)}\right]. \tag{1.6}$$

Suppose we estimate $e(\cdot)$ by a logistic regression model $\hat{e}(\cdot)$ fitted to the data $\{(X_i, A_i)\}_{i=1}^N$. The inverse probability weighting (IPW) estimator (Horvitz and Thompson, 1952) is given by

$$\hat{\tau}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\frac{A_i}{\hat{e}(X_i)}Y_i - \frac{1}{N}\sum_{j=1}^{N}\frac{1-A_j}{1-\hat{e}(X_j)}Y_j.$$

Augmented inverse probability-weighted (AIPW) estimator. Robins et al. (1994) propose the well-known AIPW estimator which combines the outcome regression and propensity score models to estimate ATE. We next introduce AIPW and its advantages in the language of semiparametric theory (Kennedy, 2016; Tsiatis, 2006).

Suppose that we view ATE as a functional $\psi : \mathcal{P} \to \mathbb{R}$, where $\mathcal{P}$ is a general class of distributions, i.e., the set of all possible observed data distributions. Suppose that any $\mathbb{P} \in \mathcal{P}$ has a density function p_O. We define a path through $\mathbb{P}$ as one-dimensional submodel that passes through $\mathbb{P}$ at $\epsilon = 0$ in the direction of a zero-mean function s s.t. $\|s\|_2 < C$ and

$\epsilon < 1/C$ for some constant $C > 0$. The submodel $\mathbb{P}_\epsilon$ has a density $p_{O,\epsilon}(o) = p_O(o)[1 + \epsilon s(o)]$ for any $o \in \mathcal{O} = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. The tangent space is the set of mean-zero functions s for any paths through $\mathbb{P}$. For nonparametric models, the tangent space is the Hilbert space of zero-mean functions. Suppose that the ATE functional $\psi(\mathbb{P})$ is path-wise differentiable:

$$\dot{\psi}_{\mathbb{P}}(s) := \frac{d}{d\epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = \int [\phi_{\mathbb{P}}(o) - \psi(\mathbb{P})]\, s(o) p_O(o) do, \tag{1.7}$$

where $s(o) = \frac{d}{d\epsilon} \log p_{O,\epsilon}(o)\big|_{\epsilon=0}$ and $\phi_{\mathbb{P}}(o) - \psi(\mathbb{P})$ is the unique riesz representer of $\dot{\psi}_{\mathbb{P}}(\cdot)$. In semiparametric theory, we call $\phi_{\mathbb{P}}(o)$ the uncentered efficient influence function (EIF) of $\psi(\mathbb{P})$. We obtain $\phi_{\mathbb{P}}$ by taking a derivative of $\psi(\mathbb{P}_\epsilon)$ with respect to ϵ and rewriting it as an inner product with $s(o)$; see Levy (2019, Section 4.2) for the detailed derivation. We call $\eta = (\mu_0, \mu_1, e)$ the nuisance parameter (model) and $\hat{\eta} = (\hat{\mu}_0, \hat{\mu}_1, \hat{e})$ the nuisance estimator. The uncentered EIF $\phi_{\mathbb{P}}(O)$ depends on $\mathbb{P}$ through η so we denote it by

$$\phi_\eta(O) = \frac{A}{e(X)}[Y - \mu_1(X)] + \mu_1(X) - \frac{1-A}{1-e(X)}[Y - \mu_0(X)] - \mu_0(X),$$

which satisfies that

$$\text{ATE} = \psi(\mathbb{P}) = \mathbb{P}\phi_\eta := \mathbb{E}\left[\phi_\eta(O)\right]. \tag{1.8}$$

Let $\mathbb{P}_N$ be the empirical measure of the samples $O_i, \ldots, O_N$, and $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - \mathbb{P})$. The AIPW estimator is given by

$$\hat{\tau}_{\text{AIPW}} = \hat{\psi}(\mathbb{P}_N) = \mathbb{P}_N\phi_{\hat{\eta}} = \frac{1}{N}\sum_{i=1}^{N}\phi_{\hat{\eta}}(O_i).$$

Consider the following decomposition,

$$\sqrt{N}\left[\hat{\psi}(\mathbb{P}_N) - \psi(\mathbb{P})\right] = \mathbb{G}_N\{\phi_{\hat{\eta}} - \phi_\eta\} + \mathbb{G}_N\phi_\eta + \sqrt{N}\left(\mathbb{P}\phi_{\hat{\eta}} - \mathbb{P}\phi_\eta\right).$$

The first term $\mathbb{G}_N\{\phi_{\hat{\eta}} - \phi_\eta\} = o_{\mathbb{P}}(1)$ if η and $\hat{\eta}$ are in a Donsker class of functions (Van der Vaart, 2000, Lemma 19.24) and $\hat{\eta}$ is a consistent estimator of η. We can avoid the Donsker assumption by using separate samples for constructing $\hat{\eta}$ and the empirical measure $\mathbb{P}_N$. Sample splitting opens the door for complex machine learning models to estimate ATE. The asymptotic efficiency loss from sample splitting can be remedied by cross-fitting (Chernozhukov et al., 2018). The second term $\mathbb{G}_N\phi_\eta = o_{\mathbb{P}}(1)$ by the central limit theorem. We want the bias $\mathbb{P}\phi_{\hat{\eta}} - \mathbb{P}\phi_\eta = o_{\mathbb{P}}(1/\sqrt{N})$ so that the third term also equals to $o_{\mathbb{P}}(1)$. Then $\hat{\psi}(\mathbb{P}_N)$ is consistent and asymptotically normal (CAN) estimator of $\psi(\mathbb{P})$ such that

$$\sqrt{N}\left[\hat{\tau}_{\text{AIPW}} - \psi(\mathbb{P})\right] \xrightarrow{d} \mathcal{N}\left(0, \text{Var}[\phi_\eta]\right). \tag{1.9}$$

For obtaining a CAN estimator, the squared bias of AIPW enjoys a multiplicative property in its upper bound,

$$(\mathbb{P}\phi_{\hat{\eta}} - \mathbb{P}\phi_{\eta})^2 \leq C\mathbb{E}\left\{[\hat{e}(X) - e(X)]^2\right\} \cdot \max_{a\in\{0,1\}} \mathbb{E}\left\{[\hat{\mu}_a(X) - \mu_a(X)]^2\right\}, \tag{1.10}$$

for some universal constant $C > 0$. The upper bound implies that $\mathbb{P}\phi_{\hat{\eta}} - \mathbb{P}\phi_{\eta} = o_{\mathbb{P}}(1/\sqrt{N})$ even when $(\hat{\mu}_0, \hat{\mu}_1)$ and $\hat{e}$ converge at slower nonparametric rates. This is not the case for the OR and IPW estimators. By applying the same decomposition, we can see that their squared bias is upper bounded by the mean squared error of either $(\hat{\mu}_0, \hat{\mu}_1)$ or $\hat{e}$. The upper bound (1.10) also shows that the bias of AIPW is 0 as long as $\hat{e} = e$ or $\hat{\mu}_a = \mu_a$ for $a = 0, 1$. This result is well known as the *doubly robust* property of AIPW in the literature.

AIPW is also known as *semiparametric efficient*, which means that it is more efficient than any CAN estimators. This property of AIPW is due to the fact that the uncentered EIF ϕ_η in (1.9) has a smaller variance than any other uncentered influence function. Roughly speaking, deriving the EIF in (1.7) is similar to deriving the score function for a maximum likelihood estimator in a parametric model. However, the nuisance parameter η in a semiparametric model is infinite-dimensional. So we can no longer define the score function analogously. From parametric models to semiparametric models, influence functions play a crucial role in generalizing the asymptotic efficiency theory, e.g., the Cramer-Rao lower bound. We refer the readers to a more formal discussion about the semiparametric efficiency theory in (Bickel et al., 1993; Tsiatis, 2006; Van der Vaart, 2000).

In summary, the AIPW estimator is *doubly robust and semiparametric efficient*. It is a very influential result in causal inference and related fields. It has been found that there are various ways to obtain a doubly robust estimator (Bang and Robins, 2005; Cheng et al., 2020a; Chernozhukov et al., 2018; Robins et al., 2007; Van Der Laan and Rubin, 2006; Zhao and Percival, 2017). The idea of deriving doubly robust estimators based on EIFs has been extended to estimate other important causal parameters such as continuous treatment effects (Kennedy et al., 2017), time-varying treatment rules (Zhang et al., 2013) and so on.

1.3.2 Conditional average treatment effect

Following (1.3), we can write the CATE function in (1.2) as

$$\tau(X) := \text{CATE}(X) = \mathbb{E}[Y \mid X, A = 1] - \mathbb{E}[Y \mid X, A = 0]. \tag{1.11}$$

The CATE function is infinite-dimensional and not pathwise-differentiable. The one dimensional ATE parameter can be viewed a function of the density value $p_O(o)$ for any $o \in \mathcal{O}$,

assuming that $p_O(o) > 0, \forall o \in \mathcal{O}$. Any observation $O_i \in \mathcal{O}$ provides information about the ATE parameter. As a comparison, only the observations nearby the point x provides information about CATE(x).

Like any localized regression model (e.g. k-nearest neighbors (Altman, 1992), or Nadaraya-Watson (Nadaraya, 1964; Watson, 1964)), conditioning on (nearby) observations not exactly at x biases the CATE estimate at x. Similarly, the size of bias depends on the smoothness of the CATE function τ. In nonparametric regression, if we know the conditional mean $f(x) = \mathbb{E}[Y \mid X = x]$ is a β-Hölder continuous function (Tsybakov, 2008, Definition 1.2) (β is known), it is possible to construct a valid confidence interval for $f(x)$ at a particular x using the minimax optimal root mean squared error of the estimator $\hat{f}$ (Györfi et al., 2002; Low, 1997). The validity we discuss here is different from the honesty of confidence regions or balls proposed by Li (1989), which intends to cover the error $\sqrt{N}\|\hat{f} - f\|_2$ with high probability. If only a lower bound β_0 is known for the true β, constructing a valid confidence interval for $f(x)$ at a particular x turns out to be impossible (Genovese and Wasserman, 2008; Low, 1997). However, achieving nearly marginal coverage guarantee for $f(x)$ at most of the points $x \in \mathcal{X}$ is possible if $\beta \leq 2\beta_0$ (Cai et al., 2014; Hall and Horowitz, 2013).

These fundamental results in nonparametric regression also apply to CATE estimation. In what follows, we will mainly discuss how to achieve a better rate of convergence in CATE estimation. The key observation is that the CATE function τ is often smoother than μ_0 and μ_1 because μ_0 and μ_1 are two similar functions and τ is given by their difference in (1.11). Leveraging the smoothness of τ has led to some recent progress in CATE estimation.

The identification formulas for ATE above prepare us to estimate CATEs. In (1.5), (1.6) or (1.8), ATE $= \mathbb{E}[\mathbb{E}[f_\eta(O) \mid X]]$ for some function $f_\eta(O)$ depending on (part of) the nuisance parameter $\eta = (\mu_0, \mu_1, e)$. Since ATE $= \mathbb{E}[\tau(X)]$ and $\tau(X) = \mathbb{E}[f_\eta(O) \mid X]$, a CATE estimator can be given by regressing $f_{\hat{\eta}}(O_i)$ on X_i for some unit $i \in [N]$.

We now redefine the nuisance parameter η to formulate a more general CATE estimation framework. Let $\hat{\tau}_1$ be an estimator of

$$\tau_1(X) = \mathbb{E}\{Y - \mu_0(X) \mid X, A = 1\},$$

obtained by regressing $Y_i - \hat{\mu}_0(X)$ onto X_i for $i \in [N]$ with $A_i = 1$. Let $\hat{\tau}_0$ be an estimator of

$$\tau_0(X) = \mathbb{E}\{\mu_1(X) - Y_i \mid X, A = 0\},$$

obtained by regressing $\hat{\mu}_1(X) - Y_i$ onto X_i for $i \in [N]$ with $A_i = 0$. Let $\eta = (\mu_0, \mu_1, \tau_0, \tau_1, e)$ and $\hat{\eta} = (\hat{\mu}_0, \hat{\mu}_1, \hat{\tau}_0, \hat{\tau}_1, \hat{e})$. A general workflow to estimate τ consists of three steps:

Table 1.1 A summary of CATE learners: the CATE estimator $\hat{\tau}$ is given by the minimizer h^* of the loss function (1.12) defined by a pseudo-outcome $f_{\hat{\eta}}(O)$ and a regressor $R_{\hat{\eta},h}(O)$.

Method	Sample-splitting	Pseudo-outcome $f_{\hat{\eta}}(O)$	Regressor $R_{\hat{\eta},h}(O)$
T-learner	No	$\hat{\mu}_1(X) - \hat{\mu}_0(X)$ from (1.5)	$h(X)$
X-learner	No	$\hat{e}(X)\hat{\tau}_0(X) + [1 - \hat{e}(X)]\hat{\tau}_1(X)$	$h(X)$
IPW-learner	Yes	$AY/\hat{e}(X) - (1-A)Y/[1-\hat{e}(X)]$ from (1.6)	$h(X)$
DR-learner	Yes	$\phi_{\hat{\eta}}(O)$ from (1.8)	$h(X)$
R-learner	Yes	$Y - \hat{\mu}_0(X)$	$[A - \hat{e}(X)]h(X)$

(i) Split the samples $[N]$ into $\mathcal{M}_1$ and $\mathcal{M}_2$.[2]

(ii) Estimate η by $\hat{\eta} = (\hat{\mu}_0, \hat{\mu}_1, \hat{\tau}_0, \hat{\tau}_1, \hat{e})$ on the subsamples $\mathcal{M}_1$;

(iii) Construct the estimator $\hat{\tau}$ on the subsamples $\mathcal{M}_2$ as the minimizer

$$\hat{h} \in \arg\min_h \frac{1}{|\mathcal{M}_2|} \sum_{m \in \mathcal{M}_2} \left[f_{\hat{\eta}}(O_m) - R_{\hat{\eta},h}(O_m) \right]^2 . \tag{1.12}$$

In terms of the pseudo-outcome $f_{\hat{\eta}}$ and regressor $R_{\hat{\eta},h}$ in (1.12), Table 1.1 summarizes the CATE estimators (i.e. learners) in the literature, which includes the T-learner (Künzel et al., 2019), X-learner (Künzel et al., 2019), IPW-learner (Kennedy, 2020; Knaus et al., 2021), DR-learner (Kennedy, 2020) and R-learner (Nie and Wager, 2021).

In the T- and X-learners, $R_{\hat{\eta},h}(O)$ and $f_{\hat{\eta}}(O)$ only depend on X, then the step of pseudo-outcome regression in (1.12) simply gives $\hat{\tau} = h^* = f_{\hat{\eta}}$. In the other learners, $R_{\hat{\eta},h}$ and $f_{\hat{\eta}}$ also depend on A and Y so sample-splitting is required to prevent overfitting in pseudo-outcome regression. Fortunately, the efficiency loss from sample splitting can be remedied by cross-fitting (Chernozhukov et al., 2018): for every $i \in [N]$, generate the pseudo-outcome $f_{\hat{\eta}_{-i}}(O_i)$ using the nuisance estimator $\hat{\eta}_{-i}$ fitted to the samples $[N] \setminus \{i\}$, then let pseudo-outcome regression base on the outcomes $f_{\hat{\eta}_{-i}}(O_i), i \in [N]$. More implementation details of the learners can be found in the review article (Jacob, 2021).

Suppose that τ (i.e. τ_0 and τ_1) is γ-Hölder continuous, μ_0 and μ_1 are β-Hölder continuous, e is λ-Hölder continuous. If τ is smoother than μ_0, μ_1 and e (i.e. $\gamma > \beta, \lambda$), the oracle learner for CATE is given by regressing $Y_i(1) - Y_i(0)$ onto X_i for all $i \in [N]$, provided that $Y_i(0)$ and $Y_i(1), \forall i \in [N]$, are known. The oracle learner estimates τ with mean squared error $O(N^{-\gamma})$. In Table 1.1, each pseudo-outcome serves as a proxy of the unobserved $Y_i(1) - Y_i(0)$ so inherit the error from $\hat{\eta}$. Perhaps surprisingly, even when $\gamma > \beta, \lambda$, some learners can still achieve

[2]The T- and X-learners in Table 1.1 do not require sample-splitting. We let $\mathcal{M}_1 = \mathcal{M}_2 = [N]$.

the oracle error under some assumptions on β, λ and γ. For example, we know from Table 1.1 that the DR-learner's pseudo-outcome is given by the uncentered EIF $\phi_{\hat{\eta}}(O)$ in (1.8) which depends on $\hat{\mu}_0, \hat{\mu}_1$ and $\hat{e}$. The learner needs to further split the samples in constructing $\hat{\mu}_0, \hat{\mu}_1$ and $\hat{e}$ (Kennedy, 2020). However, by cross-fitting and ignoring the constant difference from the further split, it estimates τ with mean squared error $O(N^{-\beta}N^{-\lambda} + N^{-\gamma})$. If $\beta + \lambda > \gamma$, it achieves the oracle error $O(N^{-\gamma})$. Another example is the R-Learner (Nie and Wager, 2021), a nonparametric kernel regression extension of Robinson's transformation (Robinson, 1988). It can also achieve the oracle error $O(N^{-\gamma})$ if $\beta, \lambda > \gamma/2$.

To conclude, we note that besides leveraging the smoothness of CATEs, there is growing interest in developing statistical and machine learning models to estimate heterogeneous treatment effects, mostly CATEs. We summary some notable advances in the literature, which includes lasso (Imai and Ratkovic, 2013), boosting (Powers et al., 2018), regression trees (Athey and Imbens, 2016; Hahn et al., 2020; Hill, 2011; Su et al., 2009) random forests (Wager and Athey, 2018), Gaussian processes (Alaa and Schaar, 2018; Alaa and van der Schaar, 2017) and neural networks (Johansson et al., 2016; Kallus, 2020; Yao et al., 2018). Using these models can further improve the finite-sample performance of CATE learners.

1.4 Summary of chapters

The remainder of this thesis consists of four chapters, contributing to the above-discussed research topics, Fisher randomization tests (FRTs) (Section 1.2) and conditional average treatment effects (CATEs) (Section 1.3), respectively.

With the growing application of causal inference, randomized experiments are increasingly carried out with a complex design on a domain such as a period of time or a social network. These designs are often motivated by estimating a special kind of causal effect or improving the trial flexibility. For example, a network design tests the spill-over effect of a social campaign; a stepped-wedge design allows participants to start the treatment at different time points of the study. Despite the good incentives, they give rise to two general statistical problems in practice. First, they may violate Assumption 1 (No interference) in the potential outcomes framework and render standard causal inference methods based on Assumption 3 (SUTVA) unusable. Second, the designs may scatter the effect evidence over the domain, which requires tools to collect the evidence and combine them effectively. In Chapter 2, we propose a theory of multiple conditional randomization tests (CRTs) that addresses the two challenges simultaneously. CRTs preserve the advantages of the classical FRTs discussed in Section 1.2.2. CRTs reinforce the versatility of FRTs by introducing a conditioning mechanism, which allows researchers to test granular causal hypotheses in a complex experiment, e.g.,

if a treatment is only effective at a particular time point, or if a social network campaign only affects some users but not their connected friends. Further, our theory establishes the conditions for constructing multiple jointly valid CRTs in any arbitrary complex design. This leads to new practical methods which can combine statistical evidence over the domain to test the global effect of interest. The advantages of our method (e.g. better power and weaker model assumption) are validated extensively through simulations.

As discussed at the beginning of Section 1.3.2, CATE estimation is essentially a non-parametric regression problem, thereby overcoming the curse of dimensionality is crucial for the performance of CATE learners. Linear dimensionality reduction methods (e.g. PCA) can transform the observed covariates into a low dimensional representation. But the learnt representation may fail to preserve the predictive information of the outcome and treatment variables if the linear model is misspecified. On the other hand, applying nonlinear dimensionality reduction methods in machine learning (e.g. deep autoencoder) may lead to a non-identifiable and inconsistent representation. In Chapter 3, we resolve the non-identifiability and inconsistency issue in a partially randomized energy-based model (EBM), which is essentially an exponential family model parameterized by a deep neural network. Theoretically, we show that the representation in our model is partially identifiable up to some universal constants. We also show that by using our noise contrastive training strategy, the learnt representation will converge consistently with an increasing sample size. Experiments on simulated and real data confirm the convergence, as well as show that CATE learners based on our representations perform better than on the raw covariates or the representations obtained from other dimensionality reduction methods.

In observational studies, treatment assignment mechanisms often create a low overlap region between the treated and control populations. Estimating CATEs in the low overlap region is particularly challenging. A popular machine learning solution is to find a matching representation of the covariates by minimizing a distributional distance between the two populations. This approach is inevitably biased because the matching representation discards the information predictive of the treatment variable, which is also the key information to predict the outcome variable. In Chapter 4, we take a different approach to measure the population mismatch. Our approach is based on the posterior variance of two Bayesian outcome regression models $\hat{\mu}_0$ and $\hat{\mu}_1$. Intuitively, $\hat{\mu}_1$ has a large posterior variance in the region containing many control units, i.e., where $\hat{\mu}_1$ is fitted to very few treated units, while $\hat{\mu}_0$ has a large posterior variance in the region containing many treated units, i.e., where $\hat{\mu}_0$ is fitted to very few control units. We learn an adaptive matching representation of the covariates by minimizing the posterior variances of $\hat{\mu}_0$ and $\hat{\mu}_1$ in predicting the counterfactual outcomes. The idea of our method is orthogonal to what we discuss in Section 1.3.2 about leveraging the smoothness of CATEs. We propose a PAC-Bayes generalization bound to show

that our method is effective and consistent in estimating CATEs. Experiments demonstrate a superior performance of our method compared with a variety of baseline methods.

Despite good accuracy, CATE estimates from black-box models are often non-interpretable and untrustworthy. To combat this challenge, Chapter 5 proposes a method that can convert any black-box CATE estimates into interpretable subgroups. The method starts by converting the point effect estimates into confidence intervals of CATEs, using a distribution-free technique called conformal prediction. Then a proposed recursive partitioning procedure is applied to divide the covariate space into subsets by minimizing the interval overlap between units across subsets. In this way, our method can potentially identify subgroups of units with significantly different effect sizes. Compared with CATE estimates from black-box models, the subgroups are much more interpretable in terms of the threshold values that partition the covariate space. Experiments on (semi-)synthetic datasets show that our method can identify subgroups with higher effect homogeneity within subgroups and higher heterogeneity across subgroups compared with various subgroup analysis methods.

1.5 Notation

Throughout this thesis, we always assume we have a study of N units (or individuals). In Chapter 2, we denote the structural treatment variables in complex experimental designs by $Z_i, \ldots, Z_N$ to distinguish with the binary treatment variables $A_1, \ldots, A_N$ in the standard treated-versus-control studies. In Chapters 3 to 5, we assume every unit i has two potential outcomes $Y_i(0)$ and $Y_i(1)$, a set of covariates $X_i \in \mathcal{X}$, a treatment variable $A_i \in \mathcal{A} = \{0, 1\}$ and an observed outcome $Y_i \in \mathcal{Y}$. The assumptions we make to CATEs from observational data are 3 (SUTVA), 5 (Unconfoundedness) and 6 (Positivity). We combine these assumptions into one below and will refer to this assumption when it is required.

Assumption 7 (SUTVA, Unconfoundedness and Positivity)**.** For any $i \in [N]$ and $a \in \{0, 1\}$, $Y_i = Y_i(a)$ if $A_i = a$. For any $i \in [N]$, the distribution of $X_i, A_i, Y_i(0)$ and $Y_i(1)$ satisfies that $Y_i(0), Y_i(1) \perp\!\!\!\perp A_i | X_i$ and $\mathbb{P}(A_i = a | X_i = x) \in (0, 1), \forall x \in \mathcal{X}$ and $a \in \{0, 1\}$.

Chapter 2

Multiple conditional randomization tests

We propose a general framework for (multiple) conditional randomization tests that incorporate several important ideas in the recent literature. We establish a general sufficient condition on the construction of multiple conditional randomization tests under which their p-values are "independent", in the sense that their joint distribution stochastically dominates the product of uniform distributions under the null. The versatility of our framework is further illustrated by a method developed for testing lagged treatment effects in stepped-wedge randomized trials. A weighted Z-score test is further proposed to maximize the power when the tests are combined. We compare the efficiency and robustness of our methods with the commonly used mixed-effects models using simulated experiments and real trial data. Conceptually, we argue that randomization should be understood as the mode of inference precisely based on randomization. We show that under a change of perspective, many existing statistical methods, including permutation tests for (conditional) independence and conformal prediction, are special cases of our general conditional randomization test.

2.1 Introduction

Randomization is one of the oldest and most important topics in statistics and experimental designs (Fisher, 1926, 1935). Randomization tests were proposed by Fisher (1935, Section 21) to substitute the t-test when normality is not true and to restore randomization as "the physical basis of the validity of the test". This idea was immediately extended by Pitman (1937), Welch (1937), Wilcoxon (1945), and Kempthorne (1952), among many others.

Randomization tests are appealing because they are exact and do not rely on distributional assumptions. Due to this reason, they are often advocated with rank-based statistics under the name of *nonparametric tests* (Lehmann, 1975). This has led to a wide held belief that randomization tests are synonymous with permutation tests, which emphasize instead on the algorithm rather than the basis of inference. The role of physical randomization has also become obscure in practice. For example, at the time of writing, randomization tests are being described on the Wikipedia page on "Resampling (statistics)" with the bootstrap, subsampling, and cross-validation procedures. In the Cambridge dictionary of statistics (Everitt and Skrondal, 2002), randomization tests are defined as "procedures for determining statistical significance directly from data without recourse to some particular sampling distribution" without referring to the physical act of randomization.

More recently, there has been a rejuvenated interest in randomization tests in several areas of statistics, including testing associations in genomics (Bates et al., 2020; Efron et al., 2001), testing conditional independence (Berrett et al., 2020; Candès et al., 2018), conformal inference for machine learning methods (Lei et al., 2013; Tibshirani et al., 2019; Vovk et al., 2005), analysis of complex experimental designs (Ji et al., 2017; Morgan and Rubin, 2012), evidence factors for observational studies (Karmakar et al., 2019; Rosenbaum, 2010, 2017), and causal inference with interference (Athey et al., 2018; Basse et al., 2019). Because randomization tests are distribution-free, they offer an easy way out of the difficulties (or even impossibilities) in theoretically deriving the sampling distribution, a frequent task in modern statistical applications. One of the main goals of this chapter is to provide a unified framework for (conditional) randomization tests and try to push some existing ideas to their natural boundaries. This framework should subsume several general ideas and concepts that have appeared in the literature: explicit conditioning on the counterfactual or potential outcomes of the experiment (Rosenbaum, 1984; Rubin, 1980a); algebraic structure of permutation tests (Lehmann and Romano, 2006; Rosenbaum, 2017; Southworth et al., 2009); randomization model versus population model (Ernst, 2004; Lehmann, 1975); post-experiment conditioning and randomization (Basse et al., 2019; Bates et al., 2020; Hennessy et al., 2016); using exchangeability to obtain distribution-free predictive intervals (Vovk et al., 2005).

Some of these ideas were introduced recently, but many were re-introduced from the earlier literature. For example, counterfactual outcomes were used in one of the earliest works on randomization tests by Welch (1937). Conditioning is not a new technique in statistical inference (e.g. Fisher's exact test for 2×2 tables (Fisher, 1925)), nor is a randomized test (e.g., the Neyman-Pearson lemma for discrete distributions, where decision at the critical value needs to be randomized to make the test exact). However, we believe that these ideas have not been fully explored in the context of randomization inference, perhaps partly due to the fact that they originated from some very different applications.

Randomization inference should be precisely understood as its name suggests: it is a mode of statistical inference that is based on randomization and nothing more than randomization. To make the nature of randomization clear, we argue that it is helpful to consider counterfactual versions of the data, even if they cannot be immediately conceptualized for the problem at hand. The introduction of potential/counterfactual outcomes allows us to trichotomize the randomness in data as

(i) Randomness in nature that is involved in all the counterfactual variables;

(ii) Randomness that is introduced by the experimenter through physical acts (e.g., drawing balls from an urn or using a pseudo-random number generator on a computer);

(iii) Randomness that is optionally introduced by the analyst.

Using this trichotomy, a randomization test can be understood as a null hypothesis significance test that conditions on the potential outcomes and obtains the sampling distribution (often called the *randomization distribution*) by random variations from the second and third sources. In other words, a randomization test is based solely on the randomness introduced by humans and thus provides a coherent logic of scientific induction as envisioned by Fisher (1956).

The above trichotomy is hardly new, nor is the definition of randomization distribution. As mentioned above, conditioning on the potential outcomes is almost always implicit. The difference between randomization before and after the experiment has also been well recognized by Basu (1980). However, these conceptual and methodological ideas often only appear as neat tricks that solve some specific problems or witty points in a philosophical debate about statistics. We argue that if these ideas are put together in a single rigorous framework, a great deal can be learned: the structure of randomization tests becomes clearer, one can understand the strengths and limitations of randomization tests much better, and some confusions and misunderstandings in the literature can be settled.

As an example, a common belief is that randomization tests rely on certain kinds of group structure or exchangeability (Lehmann and Romano, 2006; Rosenbaum, 2017; Southworth et al., 2009), which is perhaps why some texts treat randomization tests and permutation tests as synonyms. We will show that such algebraic structures are not necessary. This point becomes almost immediately clear once randomization inference is understood as the mode of inference based on randomization. That being said, in conditional randomization tests some invariance properties are needed to ensure that conditioning is well defined. This point can be easily overlooked from an algorithmic point of view and has led to mistakes.

Another frequent point of confusion is that randomization tests are applicable in two different models, one involving physical randomization and one involving exchangeability.

The former is usually referred to as the *randomization model* and the latter is referred to as the *population model* (Ernst, 2004; Lehmann, 1975). Although such a classification may be useful for pedagogic purposes, we argue that these two models are "two sides of the same coin". A randomization test not only tests the null hypothesis (relation between the potential outcomes), but also tests the assumption that the treatment is independent of the potential outcomes (Rubin, 1980a). The two models are thus unified: the first hypothesis is satisfied by definition in the population model (so independence is tested), while the second hypothesis is automatically satisfied by the physical act of randomization in the randomization model (so the null hypothesis on the potential outcomes is tested).

Our main theoretical contribution is a very general sufficient condition in Theorem 2 below that ensures the p-values from multiple conditional randomization tests are "independent", in the sense that their joint distribution stochastically dominates the multivariate uniform distribution under the null. This is important because not only is the type I error of each test under control, the individual tests can also be combined by standard methods such as Fisher's combination method to test the global null. Our sufficient condition generalizes the knit product structure described by Rosenbaum (2017), which requires that the randomization tests are constructed in a sequential manner. Our condition places fewer constraints on the construction of randomization tests and allows common techniques like sample splitting.

We briefly introduce some notations used in this chapter. We use calligraphic letters for other sets, boldface letters for vectors, upper-case letters for random quantities, and lower-case letters for fixed quantities. We use $\mathbf{1}_l$ to denote a length l vector of ones and $\mathbf{0}_l$ a length l vector of zeros. The subscript l is often omitted if the vector's dimension is clear from the context. We use a single integer in a pair of square brackets as a shorthand notation for the indexing set from 1: $[N] = \{1, \ldots, N\}$. We use set-valued subscript to denote a sub-vector; for example, $\boldsymbol{Y}_{\{1,3,5\}} = (Y_1, Y_3, Y_5)$.

The rest of this chapter is structured as follows. Section 2.2 builds a general framework for constructing a conditional randomization test (CRT) by pooling ideas scattered in the literature. Section 2.3 introduces and proves the main theorem of this chapter that gives a sufficient condition for "independent" CRTs. Because the theory of CRTs in Sections 2.2 and 2.3 is quite abstract, some simple practical methods and techniques are provided in Sections 2.2.5 and 2.3.3. Section 2.4 develops a multiple-CRTs-based method and proposes an efficient p-value combination method for testing lagged effects in stepped-wedge randomized controlled trials. Section 2.5 validates the robustness and improved power of our methods by a set of numerical experiments and real trial data. Section 2.6 illustrates our theory by discussing a variety of randomization tests or related methods in the literature. Section 2.7 concludes the chapter with a summary discussion.

2.2 A single conditional randomization test

We start with a general construction of randomization tests for the existence of treatment effect in randomized experiments. This requires us to adopt a causal inference perspective and use the potential outcomes language (Neyman, 1923; Rubin, 1974). Many randomization tests in the literature do not explicitly involve potential outcomes, but we argue in Section 2.6 that they can be viewed as special cases of our general construction.

2.2.1 Potential outcomes framework

Consider an experiment on N units in which a treatment variable $\boldsymbol{Z} \in \mathcal{Z}$ is randomized. We use boldface $\boldsymbol{Z}$ to emphasize that the treatment $\boldsymbol{Z}$ is usually multivariate. Most experiments assume that $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$ collects a common attribute of the experimental units (e.g., whether a drug is administered). However, the dimension and nature of the treatment variable $\boldsymbol{Z}$ is not very important.[1] All that is required by the general theory below is that

(i) $\boldsymbol{Z}$ is randomized in an exogenous way by the experimenter (e.g., by tossing coins or using a random number generator);

(ii) The distribution of $\boldsymbol{Z}$ is known (this is often called the *treatment assignment mechanism*);

(iii) One can reasonably define or conceptualize the potential outcomes of the experimental units under different treatment assignments.

To formalize these requirements, we adopt the potential outcomes (also called the Neyman-Rubin or counterfactual) framework for causal inference (Imbens and Rubin, 2015; Neyman, 1923; Rubin, 1974). In this framework, unit i has a vector of real-valued *potential outcomes* $(Y_i(\boldsymbol{z}) \mid \boldsymbol{z} \in \mathcal{Z})$. We assume the *observed outcome* (or *factual outcome*) for unit i is given by $Y_i = Y_i(\boldsymbol{Z})$, where $\boldsymbol{Z}$ is the realized treatment assignment. This is often referred to as the *consistency assumption* in the causal inference literature. When the treatment $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$ is an N-vector, it is often reasonable to assume that there is *no interference* in the sense that $Y_i(\boldsymbol{z})$ only depends on $\boldsymbol{z}$ through z_i.[2] However, Our theory does not rely

[1] For example, consider an experiment where we randomize how the units interact. More specifically, suppose the units only interact with their neighbours in an undirected network and we randomly choose which pairs of units are connected. The theory in this chapter can be used for such an experiment, but the dimension of the treatment $\boldsymbol{Z}$ is $N(N-1)/2$ rather than N.

[2] The well-known *stable unit treatment value assumption* (SUTVA) assumes both no interference and consistency (Rubin, 1980a).

on this assumption; rather, we treat no interference as part of the sharp null hypothesis introduced in Section 2.2.2 below.

It is convenient to introduce some vector notation for the potential and realized outcomes. Let $\boldsymbol{Y}(\boldsymbol{z}) = (Y_1(\boldsymbol{z}), \ldots, Y_N(\boldsymbol{z})) \in \mathcal{Y} \subseteq \mathbb{R}^N$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_N) \in \mathcal{Y}$. Furthermore, let $\boldsymbol{W} = (\boldsymbol{Y}(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{Z}) \in \mathcal{W}$ collect all the potential outcomes (which are random variables defined on the same probability space as $\boldsymbol{Z}$). We will call $\boldsymbol{W}$ the *potential outcomes schedule*, following Freedman (2009)'s terminology.[3] This is also known as the *science table* in the literature (Rubin, 2005). It may be helpful to view potential outcomes $\boldsymbol{Y}(\boldsymbol{z})$ as a (vector-valued) function that maps $\mathcal{Z}$ to $\mathcal{Y}$; in this sense, $\mathcal{W}$ consists of all the functions from $\mathcal{Z}$ to $\mathcal{Y}$. We assume that the experiment is randomized in the following sense.

Assumption 8 (Randomized experiment)**.** $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ and the density function $\pi(\cdot)$ of $\boldsymbol{Z}$ (with respect to some reference measure on $\mathcal{Z}$) is known and positive everywhere.

We write the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{W}$ in Assumption 8 as $\boldsymbol{Z} \mid \boldsymbol{W} \sim \pi(\cdot)$. This assumption formalizes the requirement that $\boldsymbol{Z}$ is randomized in an exogenous way. Intuitively, the potential outcomes schedule $\boldsymbol{W}$ is determined by the nature of experimental units. Since $\boldsymbol{Z}$ is randomized by the experimenter, it is reasonable to assume that $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$.[4] In many experiments, the treatment is randomized according to some other observed covariates $\boldsymbol{X}$ (e.g., characteristics of the units or some observed network structure on the units). This can be dealt with by assuming $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{X}$. Notice that in this case the treatment assignment mechanism π may depend on $\boldsymbol{X}$. To simplify the exposition, unless otherwise mentioned we will simply treat $\boldsymbol{X}$ as fixed, so $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ is still true (in the conditional probability space with $\boldsymbol{X}$ fixed at the observed value).

Our theory allows an arbitrary support $\mathcal{Z}$ for the treatment variable, but we will start with a finite $\mathcal{Z}$ (e.g., $\mathcal{Z} = \{0, 1\}^N$) to be consistent with most of the existing causal inference literature (Athey et al., 2018; Rosenbaum, 1984). In the discrete case, we always use the counting measure on $\mathcal{Z}$ as the reference measure. This is convenient because all subsets of $\mathcal{Z}$ are measurable. Section 2.2.4 extends the theory to an arbitrary $\mathcal{Z}$.

2.2.2 Partially sharp null hypothesis

Treatment effects are defined as differences between potential outcomes, but only one potential outcome $Y_i = Y_i(\boldsymbol{Z})$ is observed for every unit i under the consistency assumption. A

[3] Freedman actually called this *response schedule*.

[4] From the perspective of measure-theoretic probability theory, we can view $\boldsymbol{Z}$ and $\boldsymbol{W}$ as random variables on two different sample spaces. The former is determined by the experimenter while the latter is generated by nature. We then consider the product space equipped with the product probability measure $\mathbb{P}$, so $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ under $\mathbb{P}$ by definition.

causal hypothesis defines a number of relationships between the potential outcomes. Each relationship allows us to impute some of the potential outcomes if another potential outcome is observed. We can summarize these relationships using a set-valued mapping.

Definition 1. A (partially) sharp null hypothesis H defines an *imputability mapping*

$$\begin{aligned}\mathcal{H} : \mathcal{Z} \times \mathcal{Z} &\to 2^{[N]}, \\ (\boldsymbol{z}, \boldsymbol{z}^*) &\mapsto \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*),\end{aligned}$$

where $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ is the largest subset of $[N]$ such that $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{z},\boldsymbol{z}^*)}(\boldsymbol{z}^*)$ is imputable from $\boldsymbol{Y}(\boldsymbol{z})$ under H.

Given the observed assignment $\boldsymbol{Z}$ and another assignment $\boldsymbol{z}^* \in \mathcal{Z}$, $\mathcal{H}(\boldsymbol{Z}, \boldsymbol{z}^*)$ informs us the subset of units whose outcome $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{Z},\boldsymbol{z}^*)}(\boldsymbol{z}^*)$ is imputable from the observed outcome $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{Z})$. A randomization test for H is implemented by comparing the test statistics under the realized $\boldsymbol{Z}$ and all other $\boldsymbol{z}^* \in \mathcal{Z}$. Thus, the mapping $\mathcal{H}(\boldsymbol{Z}, \boldsymbol{z}^*)$ essentially tells us the largest subset of units we can use in a randomization test that considers comparing $\boldsymbol{Z}$ with $\boldsymbol{z}^*$. Following Definition 1, we call a hypothesis H *sharp* if $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = [N]$ for any $\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$. Otherwise, we call H *partially sharp*. Some examples are provided to illustrate the definitions as follows.

Example 1 (Simple experiment with two treatment arms)**.** Suppose the treatment for each unit is binary, i.e., $Z_i = 0$ (control) or 1 (treated). Consider the constant treatment effect hypothesis $H : Y_i(1) = Y_i(0) + \tau, \forall i \in [N]$. With no interference and consistency, we can impute any potential outcome by $Y_i(z_i^*) = Y_i + (z_i^* - Z_i)\tau$. Thus, $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = [N]$ for all $\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z} = \{0, 1\}^N$ and the hypothesis H is sharp.

Example 2 (Interference)**.** Consider a randomized experiment in which units interfere with others in the same cluster (Hudgens and Halloran, 2008). For every unit i, Z_i can take four different values 0, 1, 2 and 3, denoting control, neighbour-treated, itself-treated, neighbour- and itself-treated respectively. Consider a hypothesis $H : Y_i(1) = Y_i(0) + \tau, \forall i \in [N]$. If $Z_i = 2$ or 3, the potential outcomes $Y_i(0)$ and $Y_i(1)$ are not imputable from Y_i under H. For any i such that $z_i \notin \{0, 1\}$ and $z_i^* \in \{0, 1\}$, we have $i \notin \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$. Thus, $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ is a strict subset of $[N]$ in general, implying that H is partially sharp. As a comparison, a fully sharp hypothesis is $H : Y_i(3) = Y_i(2) = Y_i(1) = Y_i(0) + \tau, \forall i \in [N]$, which assumes that the treatment effect is the same for any treatment statuses of a unit and its neighbour.

Example 3 (Stepped-wedge trial (Brown and Lilford, 2006) or staggered adoption (Abraham and Sun, 2018; Athey and Imbens, 2018))**.** Consider a sequential experiment that randomizes each unit's treatment starting time $Z_i \in [T] \cup \{\infty\}$; $Z_i = \infty$ denotes never-treated. Suppose we only observe an outcome at the end of the experiment, $Y_{iT}, \forall i \in [N]$. Consider the

hypothesis $H : Y_{iT}(1) = Y_{iT}(\infty) + \tau, \forall i \in [N]$, i.e., treating at time 1 has a constant effect τ. This hypothesis is partially sharp because $Y_i(z_i^*)$ is not imputable from $Y(z_i)$ if $z_i \in \{2, 3, \ldots, T\}$ and $z_i^* \in \{1, \infty\}$. As a comparison, a sharp hypothesis is given by $H : Y_{iT}(1) = \ldots = Y_{iT}(T) = Y_{iT}(\infty) + \tau, \forall i \in [N]$, which further assumes that the treatment starting time does not alter the treatment effect.

2.2.3 Conditional randomization tests for discrete treatments

A common feature of the complex designs (e.g. Examples 2 and 3) is that the units have more than two potential outcomes, but the partially sharp hypotheses do not involve all of them. In this case, $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ depends on $\boldsymbol{z}$ and $\boldsymbol{z}^*$ in a non-trivial way, and a randomization test can no longer compare the observed test statistics with the entire randomization distribution (i.e. all the test statistics under other assignments). As an alternative, conditioning offers a principled way to address non-imputable potential outcomes arising from testing partially sharp hypotheses. We confine ourselves to a smaller set of treatment assignments by partitioning the assignment space $\mathcal{Z}$ into subsets $\mathcal{S}_m, m \in [M]$. We will construct a randomization distribution for each subset $\mathcal{S}_m$, which is given by the test statistics under all assignments $\boldsymbol{z}^* \in \mathcal{S}_m$. If the observed assignment $\boldsymbol{Z} \in \mathcal{S}_m$, we would only compare the observed test statistic with the randomization distribution on $\mathcal{S}_m$.

Definition 2. A *conditional randomization test* (CRT) for a discrete treatment $\boldsymbol{Z}$ is defined by

(i) A *partition* $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^M$ of $\mathcal{Z}$ such that $\mathcal{S}_1, \ldots, \mathcal{S}_M$ are disjoint subsets of $\mathcal{Z}$ satisfying $\mathcal{Z} = \bigcup_{m=1}^M \mathcal{S}_m$; and

(ii) A collection of *test statistics* $(T_m(\cdot, \cdot))_{m=1}^M$, where $T_m : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$ is a real-valued function that computes a test statistic for each realization of treatment assignment $\boldsymbol{Z}$ given the potential outcomes schedule $\boldsymbol{W}$.

Any partition $\mathcal{R}$ defines an equivalent relation $\equiv_{\mathcal{R}}$ (and vice versa), so $\mathcal{S}_1, \ldots, \mathcal{S}_M$ are simply the equivalence classes generated by $\equiv_{\mathcal{R}}$. With an abuse of notation, we let $\mathcal{S}_{\boldsymbol{z}} \in \mathcal{R}$ denote the equivalence class containing $\boldsymbol{z}$. For any $\boldsymbol{z} \in \mathcal{S}_m$, we thus have $\mathcal{S}_{\boldsymbol{z}} = \mathcal{S}_m$ and $T_{\boldsymbol{z}}(\cdot, \cdot) = T_m(\cdot, \cdot)$. This notation is convenient because the p-value of the CRT defined below conditions on $\boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{z}}$ when we observe $\boldsymbol{Z} = \boldsymbol{z}$. The following property follows immediately from the fact that $\equiv_{\mathcal{R}}$ is an equivalence relation:

Lemma 1 (Invariance of conditioning sets and test statistics)**.** *For any $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}$, we have $\boldsymbol{z} \in \mathcal{S}_{\boldsymbol{z}}$, $\mathcal{S}_{\boldsymbol{z}^*} = \mathcal{S}_{\boldsymbol{z}}$ and $T_{\boldsymbol{z}^*}(\cdot, \cdot) = T_{\boldsymbol{z}}(\cdot, \cdot)$.*

Definition 3. The *p-value* of the CRT in Definition 2 is given by

$$P(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{W}) \leq T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}, \boldsymbol{W}\}, \tag{2.1}$$

where $\boldsymbol{Z}^*$ is an independent copy of $\boldsymbol{Z}$ conditional on $\boldsymbol{W}$.

Because $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ (Assumption 8), the independence copy $\boldsymbol{Z}^* \sim \pi(\cdot)$ and is independent of $\boldsymbol{Z}$ and $\boldsymbol{W}$. In (2.1), we use the notation $\mathbb{P}^*$ to emphasize that the probability is taken over the randomness of $\boldsymbol{Z}^*$.

The invariance property in Lemma 1 is important because it ensures that when computing the p-value, the same conditioning set is used for all treatment assignments within it. By using the equivalence relation $\equiv_{\mathcal{R}}$ defined by the partition $\mathcal{R}$, we can rewrite (2.1) as

$$P(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{W}) \leq T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z}^* \equiv_{\mathcal{R}} \boldsymbol{Z}, \boldsymbol{W}\}.$$

When $\mathcal{S}_{\boldsymbol{z}} = \mathcal{Z}$ for all $\boldsymbol{z} \in \mathcal{Z}$ (so $\boldsymbol{z} \equiv_{\mathcal{R}} \boldsymbol{z}^*$ for all $\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$), this reduces to an unconditional randomization test.

Notice that $T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{W})$ generally depends on some unobserved potential outcomes in $\boldsymbol{W}$. Thus (2.1) may not be computable if the null hypothesis does not make enough restrictions on $\boldsymbol{W}$. By using the imputability mapping $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ in Definition 1, this is formalized in the next definition.

Definition 4. Consider a CRT defined by the partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^M$ and test statistics $(T_m(\cdot,\cdot))_{m=1}^M$. We say the test statistic $T_{\boldsymbol{z}}(\cdot,\cdot)$ is *imputable* under a partially sharp null hypothesis H if for all $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}$, $T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{W})$ only depends on the potential outcomes schedule $\boldsymbol{W} = (\boldsymbol{Y}(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{Z})$ through its imputable part $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{z},\boldsymbol{z}^*)}(\boldsymbol{z}^*)$.

The proof of the next result can be found in Section 2.8.1.

Lemma 2. *Suppose Assumption 8 is satisfied and $T_{\boldsymbol{z}}(\cdot,\cdot)$ is imputable for all $\boldsymbol{z} \in \mathcal{Z}$. Then the p-value $P(\boldsymbol{Z}, \boldsymbol{W})$ only depends on $\boldsymbol{Z}$ and $\boldsymbol{Y}$.*

Definition 5. Under the assumptions in Lemma 2, we say the p-value is *computable* under H and denote it by $P(\boldsymbol{Z}, \boldsymbol{Y})$ with an abuse of notation.

Given a computable p-value, the CRT then rejects the null hypothesis H at significance level $\alpha \in [0,1]$ if $P(\boldsymbol{Z}, \boldsymbol{Y}) \leq \alpha$. The next theorem establishes the validity of this test.

Theorem 1. *Consider a CRT defined by the partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^M$ and test statistics $(T_m(\cdot,\cdot))_{m=1}^M$. Then the p-value $P(\boldsymbol{Z}, \boldsymbol{W})$ stochastically dominates the uniform distribution*

on $[0,1]$ *in the following sense:*

$$\mathbb{P}\{P(\boldsymbol{Z},\boldsymbol{W}) \leq \alpha \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}, \boldsymbol{W}\} \leq \alpha, \quad \forall \alpha \in [0,1], \boldsymbol{z} \in \mathcal{Z}. \tag{2.2}$$

In consequence, given Assumption 8 and a partially sharp null hypothesis H*, if* $P(\boldsymbol{Z},\boldsymbol{W})$ *is computable, then*

$$\mathbb{P}\{P(\boldsymbol{Z},\boldsymbol{Y}) \leq \alpha\} \leq \alpha, \quad \forall \alpha \in [0,1]. \tag{2.3}$$

Proof. We first write the p-value (2.1) as a probability integral transform. For any fixed $\boldsymbol{z} \in \mathcal{Z}$, let $F_{\boldsymbol{z}}(\cdot;\boldsymbol{W})$ denote the distribution function of $T_{\boldsymbol{z}}(\boldsymbol{Z},\boldsymbol{W})$ given $\boldsymbol{W}$ and $\boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}$. Given $\boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}$ (so $\mathcal{S}_{\boldsymbol{Z}} = \mathcal{S}_{\boldsymbol{z}}$ and $T_{\boldsymbol{Z}} = T_{\boldsymbol{z}}$ by Lemma 1), the p-value can be written as

$$P(\boldsymbol{Z},\boldsymbol{W}) = F_{\boldsymbol{z}}(T_{\boldsymbol{z}}(\boldsymbol{Z},\boldsymbol{W});\boldsymbol{W}).$$

To prove (2.2), we can simply use the following probabilistic result: let T be a random variable and $F(t) = \mathbb{P}(T \leq t)$ be its distribution function, then $\mathbb{P}(F(T) \leq \alpha) \leq \alpha$ for all $0 \leq \alpha \leq 1$. See Section 2.8.2 for a proof.

If the p-value is computable, we have $P(\boldsymbol{Z},\boldsymbol{W}) = P(\boldsymbol{Z},\boldsymbol{Y})$ by Lemma 2. By the law of total probability, for any $\alpha \in [0,1]$,

$$\begin{aligned}\mathbb{P}\{P(\boldsymbol{Z},\boldsymbol{Y}) \leq \alpha \mid \boldsymbol{W}\} &= \sum_{m=1}^{M} \mathbb{P}\{P(\boldsymbol{Z},\boldsymbol{Y}) \leq \alpha \mid \boldsymbol{Z} \in \mathcal{S}_m, \boldsymbol{W}\}\mathbb{P}(\boldsymbol{Z} \in \mathcal{S}_m \mid \boldsymbol{W}) \\ &\leq \sum_{m=1}^{M} \alpha\,\mathbb{P}(\boldsymbol{Z} \in \mathcal{S}_m \mid \boldsymbol{W}) = \alpha.\end{aligned}$$

Marginalizing over the potential outcomes schedule $\boldsymbol{W}$, we obtain (2.3). □

2.2.4 Extension to continuous treatments

Although most randomized experiments only involve discrete treatment variables, the results in Section 2.2.3 can be extended to the case of continuous treatments using measure-theoretic probability theory. This allows us to not only consider experiments involving continuous dosage, but also recast permutation tests as CRTs later in Section 2.6.1.

Suppose the treatment assignment $\boldsymbol{Z}$ and potential outcomes schedule $\boldsymbol{W}$ are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. That is, suppose $(\boldsymbol{Z},\boldsymbol{W})$ is a measurable function from the sample space Ω to the product space $\mathcal{Z} \times \mathcal{W}$. As in the discrete setting, we assume $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ (Assumption 8). We modify Definition 2 to allow the CRT to be defined by a countable partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^{\infty}$ of $\mathcal{Z}$ where $\mathcal{S}_1, \mathcal{S}_2, \ldots$ are measurable subsets, and $(T_m(\cdot,\cdot))_{m=1}^{\infty}$ is

a countable sequence of measurable functions (test statistics). We require $\int_{\mathcal{S}_m} \pi(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} > 0$ for all m to avoid conditioning on zero probability events, although this (and countability of the partition) can be relaxed if suitable conditional density function can be defined on zero probability events. As in the discrete setting, we abduct the notation and denote $\mathcal{S}_{\boldsymbol{z}} = \mathcal{S}_m$ and $T_{\boldsymbol{z}}(\cdot,\cdot) = T_m(\cdot,\cdot)$ for all $\boldsymbol{z} \in \mathcal{S}_m$. Similarly, the p-value $P(\boldsymbol{Z}, \boldsymbol{W})$ is still given by (2.1). This is well defined and $P(\boldsymbol{Z}, \boldsymbol{W})$ is indeed measurable because $(T_m(\cdot,\cdot))_{m=1}^{\infty}$ are measurable.

A benefit of this measure-theoretic formulation is that it provides a more concise way of stating (2.2). Let $\mathcal{G}$ be the σ-algebra generated by the conditioning events in (2.1):

$$\mathcal{G} = \sigma\left(\{\boldsymbol{Z} \in \mathcal{S}_m\}_{m=1}^{\infty}\right).$$

Because $\{\mathcal{S}_m\}_{m=1}^{\infty}$ is a partition of $\mathcal{Z}$, $\mathcal{G}$ consists of all countable unions of $\{\mathcal{S}_m\}_{m=1}^{\infty}$. This allows us to rewrite (2.2) as

$$\mathbb{P}\left(P(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha \mid \mathcal{G}, \boldsymbol{W}\right) \leq \alpha, \quad \forall \alpha \in [0,1]. \tag{2.4}$$

The conditional probability on the left-hand side of (2.4) is a random variable (function from $\mathcal{Z} \times \mathcal{W}$ to $[0,1]$) and is well defined by the Radon-Nikodym theorem as $P(\boldsymbol{Z}, \boldsymbol{W})$ is measurable. In fact, because $\mathcal{G}$ is generated by a countable partition, we may write

$$\mathbb{P}\left(P(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha \mid \mathcal{G}, \boldsymbol{W}\right) = \sum_{m=1}^{\infty} 1_{\{\boldsymbol{Z} \in \mathcal{S}_m\}}\, \mathbb{P}\left(P(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha \mid \boldsymbol{Z} \in \mathcal{S}_m, \boldsymbol{W}\right).$$

This measure-theoretic formulation not only is more general but also allows us to state the assumptions for multiple conditional randomization tests in Section 2.3 more easily. We will adopt this formulation in what follows.

2.2.5 Practical methods

The theory above is quite abstract and does not offer any guidance on how to construct computable and powerful CRTs by partitioning the assignment space $\mathcal{Z}$. Next, we summarize some practical techniques that have been proposed in the literature.

Conditioning on a function of the treatment

First, to construct invariant conditioning sets, it is common to condition on a function of $\boldsymbol{Z}$ in a CRT (Hennessy et al., 2016). This idea is formalized by the following result.

Proposition 1. *Any function $g : \mathcal{Z} \to \mathbb{N}$ defines a countable collection of invariant conditioning sets $\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z}^* \in \mathcal{Z} : g(\boldsymbol{z}^*) = g(\boldsymbol{z})\}$.*

Proof. This follows immediately by the following equivalence relation: we define $\boldsymbol{z}^* \equiv_{\mathcal{R}} \boldsymbol{z}$ if $g(\boldsymbol{z}^*) = g(\boldsymbol{z})$ holds. □

In the measure-theoretic view in Section 2.2.4, this means that we simply condition on the σ-algebra $\mathcal{G}$ generated by the random variable $g(\boldsymbol{Z})$. We require that the image of $g(\cdot)$ is countable to avoid conditioning on zero probability events.[5]

Focal units

In practice, the test statistic of a CRT usually only depends on the potential outcomes corresponding to $\boldsymbol{z}^*$ and takes the form $T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{W}) = T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{Y}(\boldsymbol{z}^*))$. The fundamental problem then is that only a sub-vector $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{z},\boldsymbol{z}^*)}(\boldsymbol{z}^*)$ of $\boldsymbol{Y}(\boldsymbol{z}^*)$ is imputable under H. To solve this problem, a natural idea is to only use the imputable potential outcomes. This is formalized in the next result.

Proposition 2. *Given any partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^{\infty}$ of $\mathcal{Z}$, let $\mathcal{H}_m = \bigcap_{\boldsymbol{z},\boldsymbol{z}^* \in \mathcal{S}^m} \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$. Then, under Assumption 8, the partition $\mathcal{R}$ and test statistics $(T_m(\boldsymbol{z}, \boldsymbol{Y}_{\mathcal{H}_m}(\boldsymbol{z})))_{m=1}^{\infty}$ define a computable p-value.*

Although Proposition 2 provides a general way of constructing imputable test statistics, the CRT is powerless if $\mathcal{H}_m$ is an empty set. More generally, the power of the CRT depends on the size of $\mathcal{S}_m$ and $\mathcal{H}_m$ and there is a trade-off: with a coarser $\mathcal{R}$, the CRT is able to utilize a larger subset $\mathcal{S}_m$ of treatment assignments but a smaller subset $\mathcal{H}_m$ of experimental units. In many problems, choosing a good partition $\mathcal{R}$ is not a trivial problem. This is particularly challenging when the problem involves interference, as the imputability mapping $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ can be quite complex. In such cases, it may be helpful to impose some structure on the imputability mapping.

Definition 6. A partially sharp null hypothesis H is said to have a *level-set structure* if there exist *exposure functions* $D_i : \mathcal{Z} \to \mathcal{D}, i = 1, \ldots, N$, such that $\mathcal{D}$ is countable and

$$\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) = D_i(\boldsymbol{z}^*)\}. \tag{2.5}$$

[5] In principle, the theory should extend to more general $g(\cdot)$ as long as the conditional distributions are well defined, see e.g., Pollard (2002, Chapter 4).

In other words, the imputability mapping is defined by the level sets of the exposure functions. To the best of our knowledge, Definition 6 was first proposed by Athey et al. (2018), but the concept of exposure mapping can be traced back to some earlier articles (Aronow and Samii, 2017; Manski, 2013; Ugander et al., 2013).

An immediate consequence of the level-set structure is that $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ is symmetric, i.e., $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \mathcal{H}(\boldsymbol{z}^*, \boldsymbol{z})$ for all $\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$. Moreover, by using the level-set structure, we can write $\mathcal{H}_m$ in Proposition 2 as

$$\mathcal{H}_m = \bigcap_{\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{S}_m} \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) \text{ is a constant over } \boldsymbol{z} \in \mathcal{S}_m\}. \tag{2.6}$$

This provides a reasonable way to choose the test statistic (experimental units) once a partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^{\infty}$ is given. However, it is often more practical to proceed in the other direction and choose the experimental units first. Aronow (2012) and Athey et al. (2018) proposed to choose a partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^{\infty}$ such that $\mathcal{H}_m$ is equal to a fixed subset of "focal units", $\mathcal{I} \subseteq [N]$, for all m. Given any $\mathcal{I} \subseteq [N]$, the conditioning set is given by all the treatment assignments such that all the units in $\mathcal{I}$ receive the same exposure. That is,

$$\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z}^* \in \mathcal{Z} : \mathcal{I} \subseteq \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)\} = \{\boldsymbol{z}^* \in \mathcal{Z} : \boldsymbol{D}_{\mathcal{I}}(\boldsymbol{z}^*) = \boldsymbol{D}_{\mathcal{I}}(\boldsymbol{z})\}, \tag{2.7}$$

where $\boldsymbol{D}_{\mathcal{I}}(\cdot) = (D_i(\cdot) : i \in \mathcal{I})$. From the right-hand side of (2.7), it is easy to see that $\{\mathcal{S}_{\boldsymbol{z}} : \boldsymbol{z} \in \mathcal{Z}\}$ satisfies Lemma 1 and thus forms a partition of $\mathcal{Z}$. Furthermore, $\{\mathcal{S}_{\boldsymbol{z}} : \boldsymbol{z} \in \mathcal{Z}\}$ is countable because $\mathcal{S}_{\boldsymbol{z}}$ is determined by $\boldsymbol{D}_{\mathcal{I}}(\boldsymbol{z})$, a subset of the countable set $\mathcal{D}^{\mathcal{I}}$.

The next proposition summarizes the method proposed by Athey et al. (2018)[6] and immediately follows from our discussion above.

Proposition 3. *Given a null hypothesis H with a level-set structure in Definition 6, and a set of focal units $\mathcal{I} \subseteq [N]$. Under Assumption 8, the partition $\mathcal{R} = \{\mathcal{S}_{\boldsymbol{z}} : \boldsymbol{z} \in \mathcal{Z}\}$ as defined in (2.7) and any test statistic $T(\boldsymbol{z}, \boldsymbol{Y}_{\mathcal{I}}(\boldsymbol{z}))$ induce a computable p-value.*

Bipartite graph representation

Puelz et al. (2021) provided an alternative way to use the level-set structure. They consider imputability mapping of the form (suppose $0 \in \mathcal{D}$)[7]

$$\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) = D_i(\boldsymbol{z}^*) = 0\}, \tag{2.8}$$

[6] They used the same test statistic in all conditioning events, which is reflected in Proposition 3. Our theory can further allow the test statistic $T_{\boldsymbol{Z}}(\boldsymbol{z}, \boldsymbol{Y}_{\mathcal{I}}(\boldsymbol{z}))$ to depend on $\boldsymbol{Z}$ through $\boldsymbol{D}_{\mathcal{I}}(\boldsymbol{Z})$.

[7] The "null exposure graph" in (Puelz et al., 2021) actually allows $D_i(\boldsymbol{z})$ and $D_i(\boldsymbol{z}^*)$ to belong to a prespecified subset of $\mathcal{D}$. This can be allowed in our setup by redefining the exposure functions.

which is slightly more restrictive than (2.5). The "conditional focal units" $\mathcal{H}_m$ in (2.6) can then be written as

$$\mathcal{H}_m = \{i \in [N] : D_i(\boldsymbol{z}) = 0, \forall \boldsymbol{z} \in \mathcal{S}_m\}. \tag{2.9}$$

Their key insight is that imputability mapping of the above form can be represented as a bipartite graph with vertex set $\mathcal{V} = [N] \cup \mathcal{Z}$ and edge set

$$\mathcal{E} = \{(i, \boldsymbol{z}) \in [N] \times \mathcal{Z} : D_i(\boldsymbol{z}) = 0\}.$$

Puelz et al. (2021) referred to this as the *null exposure graph*. Then by using (2.9), we have

Proposition 4. *$\mathcal{V}_m = \mathcal{H}_m \cup \mathcal{S}_m$ and $\mathcal{E}_m = \{(i, \boldsymbol{z}) \in \mathcal{H}_m \times \mathcal{S}_m\}$ form a biclique (i.e., a complete bipartite subgraph) in the null exposure graph.*

Therefore, the challenging problem of finding a good partition of $\mathcal{Z}$ is reduced to finding a collection of large bicliques $\{(\mathcal{V}_m, \mathcal{E}_m)\}_{m=1}^M$ in the graph such that $\{\mathcal{S}_m\}_{m=1}^M$ partitions $\mathcal{Z}$ (this was called a *biclique decomposition* in (Puelz et al., 2021)). They further described an algorithm to find a biclique decomposition by greedily removing treatment assignments in the largest biclique (some approximate algorithm is needed to find the largest biclique as it is generally an NP-hard problem).

Randomized CRTs

When using the method of focal units or bipartite cliques, the power of the CRT often heavily depends on the set of focal units or the bipartite decomposition we use. The CRT is allowed to use the treatment assignments in the conditioning set $\mathcal{S}_{\boldsymbol{Z}}$ that depends on the realized assignment $\boldsymbol{Z}$. In general, we may have several reasonable ways to partition $\mathcal{Z}$ and each partition may have higher power for different realizations of $\boldsymbol{Z}$. In this circumstance, a natural idea is to post-randomize the test.

Consider a collection of CRTs defined by $\mathcal{R}(v) = \{\mathcal{S}_m(v)\}_{m=1}^\infty$ and $(T_m(\cdot, \cdot; v))_{m=1}^\infty$ that are indexed by $v \in \mathcal{V}$ where $\mathcal{V}$ is countable. Each v defines a p-value

$$P(\boldsymbol{Z}, \boldsymbol{W}; v) = \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{W}; v) \le T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{W}; v) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}(v), \boldsymbol{W}\},$$

where $\boldsymbol{Z}^*$ is an independent copy of $\boldsymbol{Z}$. Since this defines a CRT for each fixed v, the theory in Section 2.2.3 immediately applies. Similar to Definition 5, we say $P(\boldsymbol{Z}, \boldsymbol{W}; v)$ is computable if it is a function of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ and write it as $P(\boldsymbol{Z}, \boldsymbol{Y}; v)$. In a randomized CRT, the analyst

can take V as a random variable that is independent of $\boldsymbol{Z}$ and $\boldsymbol{W}$.[8] The next result shows that we can safely test a partially sharp null hypothesis by using a randomly drawn CRT.

Corollary 1. *Under the setting above, the randomized CRT is valid in the following sense*

$$\mathbb{P}\left\{P(\boldsymbol{Z}, \boldsymbol{W}; V) \leq \alpha \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}(V), \boldsymbol{W}, V\right\} \leq \alpha, \quad \forall \alpha \in [0,1], \boldsymbol{z} \in \mathcal{Z}.$$

Proof. This immediately follows from Theorem 1. □

We can further generalize Proposition 1 and Corollary 1 to allow conditioning on a random variable $G = g(\boldsymbol{Z}, V)$ that depends on both the randomness introduced by the experimenter in $\boldsymbol{Z}$ and also the randomness introduced by the analyst in V. As above, suppose $V \perp\!\!\!\perp (\boldsymbol{Z}, \boldsymbol{W})$ and G has a countable support (to avoid conditioning on zero probability events; see footnote 5). Because G is generated by $\boldsymbol{Z}$, the conditional distribution of G given $\boldsymbol{Z}$ is known. Let $\pi(\cdot \mid g)$ be the density function of $\boldsymbol{Z}$ given $G = g$ that can be obtained from Bayes' formula:

$$\pi(\boldsymbol{z} \mid g) = \frac{\mathbb{P}(G = g \mid \boldsymbol{Z} = \boldsymbol{z})\pi(\boldsymbol{z})}{\int \mathbb{P}(G = g \mid \boldsymbol{Z} = \boldsymbol{z})\pi(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}}.$$

Let $T_g(\cdot, \boldsymbol{W})$ be the test statistic that is now indexed by g in the support of G. Then the randomized p-value is defined as

$$P(\boldsymbol{Z}, \boldsymbol{W}; G) = \mathbb{P}^*\left\{T_G(\boldsymbol{Z}^*, \boldsymbol{W}) \leq T_G(\boldsymbol{Z}, \boldsymbol{W}) \mid G, \boldsymbol{W}\right\},$$

where the probability is taken over $\boldsymbol{Z}^* \mid G, \boldsymbol{W} \overset{d}{=} \boldsymbol{Z} \mid G, \boldsymbol{W}$. In other words, $\boldsymbol{Z}^* \sim \pi(\cdot \mid G)$ and the randomized p-value can be written as

$$P(\boldsymbol{Z}, \boldsymbol{W}; G) = \int 1_{\{T_G(\boldsymbol{z}^*, \boldsymbol{W}) \leq T_G(\boldsymbol{Z}, \boldsymbol{W})\}} \pi(\boldsymbol{z}^* \mid G)\,\mathrm{d}\mu(\boldsymbol{z}^*),$$

where μ is the reference measure on $\mathcal{Z}$. Similar to above, we say $P(\boldsymbol{Z}, \boldsymbol{W}; g)$ is computable if it is a function of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ and write it as $P(\boldsymbol{Z}, \boldsymbol{Y}; g)$.

Corollary 2. *Under the setting above, the randomized CRT is valid in the following sense*

$$\mathbb{P}\left\{P(\boldsymbol{Z}, \boldsymbol{W}; G) \leq \alpha \mid \boldsymbol{W}, G\right\} \leq \alpha, \quad \forall \alpha \in [0,1].$$

Proof. This follows from the observation that once G is given, this is simply a unconditional randomization test. The p-value is computed using $\boldsymbol{Z}^* \sim \pi(\cdot \mid G)$, so given G, we can write $P(\boldsymbol{Z}, \boldsymbol{W}; G)$ as a probability integral transform as in the proof of Theorem 1. □

[8] This may require enlarging the sample space as in footnote 4, as the randomness in V is introduced by the analyst of the experiment, which may be different from the experimenter who only randomizes $\boldsymbol{Z}$. In what follows, we use $\mathbb{P}$ to denote the product probability measure on $(\boldsymbol{Z}, \boldsymbol{W}, V)$ when the CRT is randomized.

Corollary 2 is essentially the same as (Basse et al., 2019, Theorem 1), although we do not require imputability of the test statistic. In other words, imputability only affects whether the p-value can be computed using the observed data and is not necessary for the validity of the p-value conditional on potential outcome schedule $\boldsymbol{W}$.

2.3 Multiple conditional randomization tests

In the previous section, we have shown how to construct valid and computable CRTs for partially sharp null hypotheses arising in complex experimental designs. Another key feature of complex designs is that they offer scattered evidence of causation over the experimental domain. However, combining the evidence (i.e. testing the intersection of partially sharp null hypotheses) in a single CRT is not straightforward or requires careful consideration of the design. This is because different partially sharp null hypotheses are based on different potential outcomes, so testing each of them requires a different conditioning set of assignments to ensure the test statistics is imputable. The difficulty of running a single CRT for multiple hypotheses will be further illustrated in the next section by stepped-wedge trials (previewed in Example 3) which spread the evidence of causation across time.

In this section, we introduce a theory concerning how to construct multiple jointly valid CRTs. Building upon the general setup in the last section, our theory places the condition only on the construction of randomization tests but not on the underlying design. Thus, it is potentially applicable to any design with arbitrary treatment variable, assignment mechanism and unit interference. The jointly valid CRTs we describe later can be combined by standard methods such as Fisher's method (Fisher, 1925) and can be easily extended to simultaneous testing using the closed testing procedure (Marcus et al., 1976).

2.3.1 Main theorem

Consider K conditional randomization tests, defined by partitions $\mathcal{R}^{(k)} = \left\{\mathcal{S}_m^{(k)}\right\}_{m=1}^{\infty}$ and test statistics $(T_m^{(k)}(\cdot,\cdot))_{m=1}^{\infty}$, for K possibly different hypotheses $H^{(k)}$, $k = 1, \ldots, K$. We denote the corresponding p-values (2.1) as $P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}), \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W})$.

For any subset of tests $\mathcal{J} \subseteq [K]$, we define the *union*, *refinement* and *coarsening* of the conditioning sets as

$$\mathcal{R}^{\mathcal{J}} = \bigcup_{k\in\mathcal{J}} \mathcal{R}^{(k)}, \quad \underline{\mathcal{R}}^{\mathcal{J}} = \left\{ \bigcap_{j\in\mathcal{J}} \mathcal{S}_{\boldsymbol{z}}^{(j)} : \boldsymbol{z} \in \mathcal{Z} \right\}, \quad \text{and} \quad \overline{\mathcal{R}}^{\mathcal{J}} = \left\{ \bigcup_{j\in\mathcal{J}} \mathcal{S}_{\boldsymbol{z}}^{(j)} : \boldsymbol{z} \in \mathcal{Z} \right\}.$$

As in Section 2.2.4, let $\mathcal{G}^{(k)} = \sigma(\{\boldsymbol{Z} \in \mathcal{S}\} : \mathcal{S} \in \mathcal{R}^{(k)})$ be the σ-algebra generated by the conditioning events for test k. Let $\mathcal{G}^{\mathcal{J}}$ be the σ-algebra generated by the sets in $\mathcal{R}^{\mathcal{J}}$, so $\mathcal{G}^{\mathcal{J}} = \sigma(\{\boldsymbol{Z} \in \mathcal{S}\} : \mathcal{S} \in \mathcal{R}^{\mathcal{J}})$. Similarly, we define the σ-algebras $\underline{\mathcal{G}}^{\mathcal{J}} = \sigma(\{\boldsymbol{Z} \in \mathcal{S}\} : \mathcal{S} \in \underline{\mathcal{R}}^{\mathcal{J}})$ and $\overline{\mathcal{G}}^{\mathcal{J}} = \sigma(\{\boldsymbol{Z} \in \mathcal{S}\} : \mathcal{S} \in \overline{\mathcal{R}}^{\mathcal{J}})$.

Theorem 2. *Suppose the following two conditions are satisfied for all $j, k \in [K]$, $j \neq k$:*

$$\underline{\mathcal{R}}^{\{j,k\}} \subseteq \mathcal{R}^{\{j,k\}}, \tag{2.10}$$

$$T_{\boldsymbol{Z}}^{(j)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp T_{\boldsymbol{Z}}^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \underline{\mathcal{G}}^{\{j,k\}}, \boldsymbol{W}. \tag{2.11}$$

Then we have

$$\begin{aligned} \mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(1)}, \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(K)} \mid \overline{\mathcal{G}}^{[K]}, \boldsymbol{W}\right\} \leq \prod_{k=1}^{K} \alpha^{(k)}, \\ \forall \alpha^{(1)}, \ldots, \alpha^{(K)} \in [0,1]. \end{aligned} \tag{2.12}$$

In consequence, given Assumption 8 and that the null hypotheses $H^{(1)}, \ldots, H^{(K)}$ are satisfied, if the CRTs are computable, then

$$\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{Y}) \leq \alpha^{(1)}, \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{Y}) \leq \alpha^{(K)}\right\} \leq \prod_{k=1}^{K} \alpha^{(k)}, \quad \forall \alpha^{(1)}, \ldots, \alpha^{(K)} \in [0,1]. \tag{2.13}$$

Before sketching a proof of this theorem, we first give a simple illustration of this general result. When $K = 2$, condition (2.10) amounts to assuming a nested structure between the two partitions: $\mathcal{S}_{\boldsymbol{z}}^{(1)} \subseteq \mathcal{S}_{\boldsymbol{z}}^{(2)}$ or $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(2)}$ for all $\boldsymbol{z} \in \mathcal{Z}$. In other words, $\mathcal{S}_{\boldsymbol{z}}^{(1)} \cap \mathcal{S}_{\boldsymbol{z}}^{(2)} = \mathcal{S}_{\boldsymbol{z}}^{(1)}$ or $\mathcal{S}_{\boldsymbol{z}}^{(2)}$ for all $\boldsymbol{z} \in \mathcal{Z}$. Notice that this condition allows $\mathcal{S}_{\boldsymbol{z}}^{(1)} \subseteq \mathcal{S}_{\boldsymbol{z}}^{(2)}$ for some $\boldsymbol{z}$ and $\mathcal{S}_{\boldsymbol{z}^*}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}^*}^{(2)}$ for another $\boldsymbol{z}^* \neq \boldsymbol{z}$. Furthermore, when $K = 2$, condition (2.11) is equivalent to assuming

$$T_{\boldsymbol{z}}^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp T_{\boldsymbol{z}}^{(2)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}^{(1)} \cap \mathcal{S}_{\boldsymbol{z}}^{(2)}, \boldsymbol{W}, \quad \forall \boldsymbol{z} \in \mathcal{Z}.$$

Finally, when $K = 2$, the main conclusion (2.12) is equivalent to

$$\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(1)}, P^{(2)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha_2 \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}^{(1)} \cup \mathcal{S}_{\boldsymbol{z}}^{(2)}, \boldsymbol{W}\right\} \leq \alpha^{(1)}\alpha^{(2)}$$

for all $\boldsymbol{z} \in \mathcal{Z}$ and $\alpha^{(1)}, \alpha^{(2)} \in [0,1]$. See Section 2.3.3 for a simple proof of the $K = 2$ case under slightly stronger conditions than (2.10) and (2.11), which sheds some light on the proof for the general case.

2.3.2 Proof outline

We now outline a proof of Theorem 2. To this end, we need to consider the structure of the conditioning sets given by (2.10). We start with the following observation (the proof can be found in Section 2.8.3).

Lemma 3. *Suppose (2.10) is satisfied. Then for any $\mathcal{J} \subseteq [K]$ and $\mathcal{S}, \mathcal{S}' \in \mathcal{R}^{\mathcal{J}}$, the sets $\mathcal{S}$ and $\mathcal{S}'$ are either disjoint or nested, that is,*

$$\mathcal{S} \cap \mathcal{S}' \in \{\varnothing, \mathcal{S}, \mathcal{S}'\}.$$

Furthermore, we have $\underline{\mathcal{R}}^{[K]} \subseteq \mathcal{R}^{[K]}$ and $\overline{\mathcal{R}}^{[K]} \subseteq \mathcal{R}^{[K]}$.

The sets in $\mathcal{R}^{[K]}$ can be partially ordered by set inclusion. This induces a graphical structure on $\mathcal{R}^{[K]}$:

Definition 7. The *Hasse diagram* for $\mathcal{R}^{[K]} = \{\mathcal{S}_{\boldsymbol{z}}^{(k)} : \boldsymbol{z} \in \mathcal{Z}, k \in [K]\}$ is a graph where each node in the graph is a set in $\mathcal{R}^{[K]}$ and a directed edge $\mathcal{S} \to \mathcal{S}'$ exists between two distinct nodes $\mathcal{S}, \mathcal{S}' \in \mathcal{R}^{[K]}$ if $\mathcal{S} \supset \mathcal{S}'$ and there is no $\mathcal{S}'' \in \mathcal{R}^{[K]}$ such that $\mathcal{S} \supset \mathcal{S}'' \supset \mathcal{S}'$.

It is straightforward to show that all edges in the Hasse diagram for $\mathcal{R}^{[K]}$ are directed and this graph has no cycles. Thus, the Hasse diagram is a directed acyclic graph. For any node $\mathcal{S} \in \mathcal{R}^{[K]}$ in this graph, we can further define its parent set as $\mathrm{pa}(\mathcal{S}) = \{\mathcal{S}' \in \mathcal{R}^{[K]} : \mathcal{S}' \to \mathcal{S}\}$, child set as $\mathrm{ch}(\mathcal{S}) = \{\mathcal{S}' \in \mathcal{R}^{[K]} : \mathcal{S} \to \mathcal{S}'\}$, ancestor set as $\mathrm{an}(\mathcal{S}) = \{\mathcal{S}' \in \mathcal{R}^{[K]} : \mathcal{S}' \supset \mathcal{S}\}$, and descendant set as $\mathrm{de}(\mathcal{S}) = \{\mathcal{S}' \in \mathcal{R}^{[K]} : \mathcal{S} \supset \mathcal{S}'\}$.

Notice that one conditioning set $\mathcal{S} \in \mathcal{R}^{[K]}$ can be used in multiple CRTs. To fully characterize this structure, we introduce an additional notation.

Definition 8. For any $\mathcal{S} \in \mathcal{R}^{[K]}$, let $\mathcal{K}(\mathcal{S}) = \{k \in [K] : \mathcal{S} \in \mathcal{R}^{(k)}\}$ be the collection of indices such that $\mathcal{S}$ is a conditioning set in the corresponding test. Furthermore, for any collection of conditioning sets $\mathcal{R} \subseteq \mathcal{R}^{[K]}$, denote $\mathcal{K}(\mathcal{R}) = \cup_{\mathcal{S} \in \mathcal{R}} \mathcal{K}(\mathcal{S})$.

Lemma 4. *Suppose (2.10) is satisfied. Then for any $\mathcal{S} \in \mathcal{R}^{[K]}$, we have*

(i) If $ch(\mathcal{S}) \neq \varnothing$, then $ch(\mathcal{S})$ is a partition of $\mathcal{S}$;

(ii) $\{\mathcal{K}(an(\mathcal{S})), \mathcal{K}(\mathcal{S}), \mathcal{K}(de(\mathcal{S}))\}$ forms a partition of $[K]$.

(iii) For any $\mathcal{S}' \in ch(\mathcal{S})$, $\mathcal{K}(an(\mathcal{S}')) = \mathcal{K}(an(\mathcal{S}) \cup \{\mathcal{S}\})$ and $\mathcal{K}(\{\mathcal{S}'\} \cup de(\mathcal{S}')) = \mathcal{K}(de(\mathcal{S}))$.

Using Lemma 4, we can prove the following key lemma that establishes the conditional independence between two p-values. The proof of Lemmas 4 and 5 can be found in Sections 2.8.4 and 2.8.5.

Lemma 5. *Suppose* (2.10) *and* (2.11) *are satisfied. Then for any* $\mathcal{S} \in \mathcal{R}^{[K]}$, $j \in \mathcal{K}(\mathcal{S})$, *and* $k \in \mathcal{K}(an(\mathcal{S}) \cup \{\mathcal{S}\}) \setminus \{j\}$, *we have*

$$P^{(j)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}.$$

Finally, we state a result based on the above Hasse diagram that is more general than Theorem 2.

Lemma 6. *Given conditions* (2.10) *and* (2.11), *we have, for any* $\mathcal{S} \in \mathcal{R}^{[K]}$,

$$\begin{aligned} &\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(k)}, \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(K)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \\ \leq\ &\mathbb{P}\left\{P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(k)} \text{ for } k \in \mathcal{K}(an(\mathcal{S})) \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\{\mathcal{S}\} \cup de(\mathcal{S}))} \alpha_j. \end{aligned} \tag{2.14}$$

Theorem 2 almost immediately follows from Lemma 6. The proof of Lemma 6 and Theorem 2 can be found in Sections 2.8.6 and 2.8.7, respectively.

2.3.3 Practical methods

Theorem 2 provides a sufficient condition for jointly sufficient valid CRTs: the partitions $\mathcal{R}^{(k)}, k = 1, \ldots, K$, need to satisfy the nested relation characterized by (2.10), and the test statistics need to satisfy the conditional independence in (2.11). However, these two conditions are quite abstract. To illustrate the versatility of the theorem, we provide some practical techniques to construct CRTs that satisfy both conditions. To simplify the exposition, below we assume that the test statistics satisfy $T^{(k)}(\cdot, \cdot) = T_m^{(k)}(\cdot, \cdot)$ for all m and k.

Independent treatment variables

We begin by noting that (2.10) is immediately satisfied if all the randomization tests are unconditional (i.e., $\mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{Z}$ for all $\boldsymbol{z} \in \mathcal{Z}$ and $k \in [K]$). Further, the conditional independence (2.11) may be satisfied if the problem structure allows the treatment $\boldsymbol{Z}$ to be decomposed into independent components. This strategy is summarized in the proposition below.

Proposition 5. *The conditions* (2.10) *and* (2.11) *are satisfied for all* $j, k \in [K], j \neq k$ *if*

(i) $\mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{Z}$ *for all* k *and* $\boldsymbol{z}$; *and*

(ii) $T^{(k)}(\boldsymbol{Z}, \boldsymbol{W})$ *only depends on* $\boldsymbol{Z}$ *through* $\boldsymbol{Z}^{(k)} = h^{(k)}(\boldsymbol{Z})$ *for all* k *and* $\boldsymbol{Z}^{(j)} \perp\!\!\!\perp \boldsymbol{Z}^{(k)}$ *for all* $j \neq k$.

This result can be easily extended to the case if the same partition is used for all the tests, i.e., $\mathcal{R}^{(1)} = \cdots = \mathcal{R}^{(K)}$.

Sequential CRTs

Our next strategy is to construct a sequence of CRTs, where each CRT conditions on the randomness utilized by the preceding CRTs. Mathematically, we assume that the CRTs are constructed in a way such that $\mathcal{G}^{(1)} \subseteq \cdots \subseteq \mathcal{G}^{(K)}$. By definition, this is equivalent to using a sequence of nested partitions so that $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \cdots \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K)}$ for all $\boldsymbol{z}$. Under this construction, it is easy to see that $\underline{\mathcal{R}}^{\{j,k\}} = \mathcal{R}^{(\max(j,k))}$ and (2.10) is immediately satisfied. Moreover, (2.11) is automatically satisfied if $T^{(j)}(\boldsymbol{z}, \boldsymbol{W})$ does not depend on $\boldsymbol{z}$ when $\boldsymbol{z} \in \mathcal{S}_m^{(k)}$ for any m and $k > j$.[9] To see this, the condition independence

$$T^{(j)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp T^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z} \in \mathcal{S}_m^{(k)}, \boldsymbol{W},$$

is immediately satisfied because $T^{(j)}(\boldsymbol{Z}, \boldsymbol{W})$ is just a constant given $\boldsymbol{Z} \in \mathcal{S}_m^{(k)}$.

To summarize, we have the following result.

Proposition 6. *The conditions (2.10) and (2.11) are satisfied for all $j, k \in [K], j \neq k$ if*

(i) $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \cdots \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K)}$ for all $\boldsymbol{z} \in \mathcal{Z}$; and

(ii) $T^{(j)}(\boldsymbol{z}, \boldsymbol{W})$ does not depend on $\boldsymbol{z}$ when $\boldsymbol{z} \in \mathcal{S}_m^{(k)}$ for all m and $k > j$.

Under the conditions in Proposition 6, the proof of Theorem 2 can be greatly simplified. To illustrate this, we consider the simplest case that $K = 2$. Let $\psi^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) = 1_{\{P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(k)}\}}$ for $k = 1, 2$ be the test functions. By Proposition 6(i), $\mathcal{G}^{(1)} \subseteq \mathcal{G}^{(2)}$. By Proposition 6(ii) and the definition of the p-value (2.1), $P^{(1)}(\boldsymbol{Z}, \boldsymbol{w})$ and thus $\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})$ are $\mathcal{G}^{(2)}$-measurable for any fixed $\boldsymbol{w}$. Then by the law of iterated expectation, for any $\boldsymbol{w} \in \mathcal{W}$,

$$\begin{aligned}
&\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{w}) \leq \alpha^{(1)}, P^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \leq \alpha^{(2)} \mid \mathcal{G}^{(1)}\right\} \\
&= \mathbb{E}\left\{\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})\psi^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(1)}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})\psi^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(2)}\right] \mid \mathcal{G}^{(1)}\right\} \\
&= \mathbb{E}\left\{\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})\mathbb{E}\left[\psi^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(2)}\right] \mid \mathcal{G}^{(1)}\right\} \\
&\leq \alpha^{(1)}\alpha^{(2)}.
\end{aligned}$$

[9] In other words, $T^{(j)}(\boldsymbol{z}, \boldsymbol{W})$ only depends on $\boldsymbol{z}$ through the indicator function for the event $\{\boldsymbol{z} \in \mathcal{S}_m^{(k)}\}$. In our measure-theoretic language, this means that $T^{(\min(j,k))}(\boldsymbol{Z}, \boldsymbol{W})$ is $\mathcal{G}^{(\max(j,k))}$-measurable given $\boldsymbol{W}$.

By using the Hasse diagram for the conditioning sets, our proof of Theorem 2 essentially extends this argument to the more general setting.

Randomized CRTs

One can also randomize the CRTs as in Section 2.2.5. We illustrate this idea here using a proposal by Bates et al. (2020). In that problem, the test statistics are of the form $T^{(k)}(\boldsymbol{Z}^{(k)}, \boldsymbol{W})$ (as in Proposition 5) and there exists a random variable U such that

$$\boldsymbol{Z}^{(1)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \boldsymbol{Z}^{(K)} \mid U.$$

The challenge is that U is unobserved, although the joint distribution of $(U, \boldsymbol{Z})$ is known. The key idea of Bates et al. (2020) is to construct a post-treatment random variable $G = g(\boldsymbol{Z}, V)$ (V is randomized by the analyst), where the function $g(\cdot, \cdot)$ and the distribution of V are chosen such that G has the same conditional distribution as U given $\boldsymbol{Z}$. Then we have

$$\boldsymbol{Z}^{(1)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \boldsymbol{Z}^{(K)} \mid G. \tag{2.15}$$

Conditional on G, (2.10) is satisfied because the tests are unconditional and (2.11) is also immediately satisfied due to (2.15). To summarize, we have the following result.

Proposition 7. *Suppose the test statistics are constructed in a randomized way as in Corollary 2. Then conditional on G, conditions (2.10) and (2.11) are satisfied for all $j, k \in [K], j \neq k$ if*

(i) $\mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{Z}$ for all k and $\boldsymbol{z}$; and

(ii) $T^{(k)}(\boldsymbol{Z}, \boldsymbol{W})$ only depends on $\boldsymbol{Z}$ through $\boldsymbol{Z}^{(k)} = h^{(k)}(\boldsymbol{Z})$ for all k and $\boldsymbol{Z}^{(j)} \perp\!\!\!\perp \boldsymbol{Z}^{(k)} \mid G$ for all $j \neq k$.

In consequence, (2.12) is satisfied conditional on G.

Proposition 7 is basically the post-randomized version of Proposition 5. The nontrivial idea is that when conditional independence between treatment variables requires conditioning on some unobserved variable U, we can instead generate another variable G that follow the known distribution of U given Z and treat it as given. This proposal might seem magical at first, but notice that the power of the test will depend crucially on G and on how well it resembles U. If G is not similar to U, the post-randomized CRTs may have little power.

2.4 Testing lagged treatment effect in stepped-wedge randomized trials

We next demonstrate the versatility of our theory using the stepped-wedge randomized controlled trials, a widely used experimental design in medical and policy research. Mixed-effects models are commonly used to analyze stepped-wedge randomized trials (Hemming et al., 2018; Hussey and Hughes, 2007; Li et al., 2021). However, it is well recognized that the statistical inference tends to be biased (such as inflated type I error and poor confidence coverage) when the model is misspecified (Ji et al., 2017; Thompson et al., 2017). In light of this, unconditional randomization tests have been proposed to test null hypotheses of no treatment effect whatsoever by several authors (Hughes et al., 2020; Ji et al., 2017; Thompson et al., 2018; Wang and De Gruttola, 2017). However, this approach cannot be used to test fine-grained hypotheses such as those involving lagged treatment effects. Following our theory, we develop a method (Algorithm 1) to construct nearly independent CRTs for lagged treatment effects in Section 2.4.2. We also consider methods to combine the tests in Section 2.4.3.

2.4.1 Hypotheses for lagged treatment effects

The stepped-wedge randomized controlled trial is a monotonic cross-over design in which all units start out in the control and then cross over to the treatment at staggered times (Group et al., 1987; Hussey and Hughes, 2007). Figure 2.1 shows the treatment schedule for a typical stepped-wedge trial; the name "stepped-wedge" refers to the wedge shape given by the treated groups over subsequent crossover points. This design is also known as *staggered adoption* in econometrics (Abraham and Sun, 2018; Athey and Imbens, 2018). Many stepped-wedge trials are cluster-randomized, i.e., the units in the same cluster always have the same treatment status. For simplicity, here we assume that the cross-over times are randomized at the unit level. The same method below applies to cluster-randomized trials design by considering aggregated cluster-level outcomes (Middleton and Aronow, 2015; Thompson et al., 2018).

Consider a stepped-wedge trial on N units over some evenly spaced time grid $[T] = \{1, \cdots, T\}$. Let $N_1, \ldots, N_T$ be positive integers that satisfy $\sum_{t=1}^{T} N_t = N$. The cross-overs can be represented by a binary matrix $\boldsymbol{Z} \in \{0,1\}^{N\times T}$, where $Z_{it} = 1$ indicates that unit i crosses over from control to treatment at time t. We assume that the treatment assignment mechanism is given by the uniform distribution over

$$\mathcal{Z} = \left\{ \boldsymbol{z} \in \{0,1\}^{N\times T} : \boldsymbol{z}\mathbf{1} = \mathbf{1}, \mathbf{1}^\top \boldsymbol{z} = (N_1, \cdots, N_T) \right\}. \tag{2.16}$$

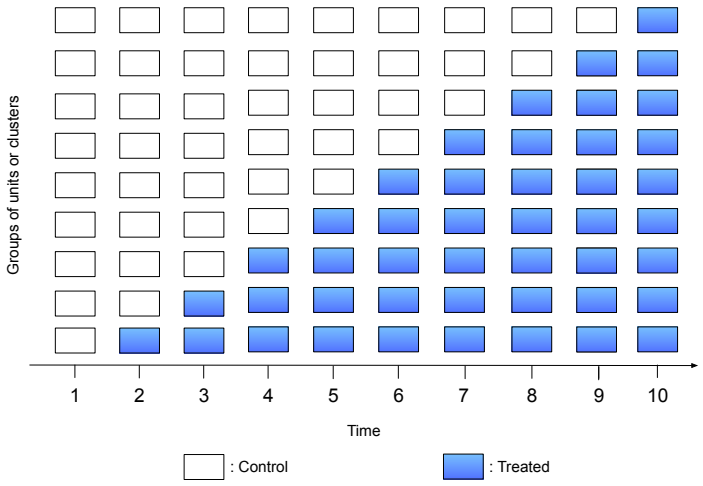


Figure 2.1 Stepped-wedge randomized trials. A group of 9 units or clusters switch from control to treatment at each time point and then remain exposed to the treatment. This figure shows the treatment status of the units right before the treatment assignment at each time point.

The set $\mathcal{Z}$ contains all binary matrices with exactly one element of 1 in each row and N_t elements of 1 in the t-th column, so $|\mathcal{Z}| = N!/(N_1! \cdots N_T!)$.

We can view the uniform distribution over $\mathcal{Z}$ as a sequentially randomized experiment. This key observation allows us to invoke the theory in Section 2.3. Let $\boldsymbol{Z}_t$ be the t-th column of $\boldsymbol{Z}$. It is not difficult to show that the uniform distribution over $\boldsymbol{Z}$ over $\mathcal{Z}$ is equivalent to randomly assigning treatment to N_t control units at time $t = 1, \ldots, T$. Denote $\boldsymbol{z}_{[t]} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t) \in \{0,1\}^{N \times t}$ and $N_{[t]} = \sum_{s=1}^{t} N_s$. More precisely, we may decompose the uniform distribution $\pi(\boldsymbol{z})$ over $\boldsymbol{z} \in \mathcal{Z}$ as

$$\frac{N_1! \cdots N_T!}{N!} = \pi(\boldsymbol{z}) = \pi_1(\boldsymbol{z}_1)\pi_2(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \cdots \pi_T(\boldsymbol{z}_T \mid \boldsymbol{z}_{[T-1]}),$$

where

$$\pi_t(\boldsymbol{z}_t \mid \boldsymbol{z}_{[t-1]}) = \frac{N_t!(N - N_{[t]})!}{(N - N_{[t-1]})!} \text{ for all } \boldsymbol{z}_t \in \{0,1\}^N \text{ such that } (\mathbf{1}_N - \boldsymbol{z}_{[t-1]}\mathbf{1}_{t-1})^T \boldsymbol{z}_t = N_t.$$

After the cross-overs at time t, the experimenter then measures the outcomes of all units, which we denote as $\boldsymbol{Y}_t \in \mathbb{R}^N$. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$ denote all the realized outcomes and $\boldsymbol{Y}(\boldsymbol{z})$ denote the collection of potential outcomes under treatment assignment $\boldsymbol{z} \in \mathcal{Z}$. Like before, we maintain the consistency assumption that $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{Z})$.

Recall that a null hypothesis H is fully sharp if its imputability mapping $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = [N]$ for any $\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$ (Section 2.2.2). An example of the fully sharp null hypothesis in this

setting is that the treatment has no effect whatsoever, i.e.,

$$H : \boldsymbol{Y}(\boldsymbol{z}) = Y(\boldsymbol{z}^*), \quad \forall \boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}, i \in [N], \text{ and } t \in [T]. \tag{2.17}$$

This hypothesis can be tested by an unconditional randomization test in a straightforward manner. Because $\boldsymbol{Z}$ is completely randomized, this is a permutation test and has been previously studied in the literature (Hughes et al., 2020; Ji et al., 2017; Wang and De Gruttola, 2017). The test statistic is usually obtained by fitting some linear mixed-effects model. However, this sharp null hypothesis is rather restrictive. In particular, the staggered treatment assignment offers an opportunity to investigate how soon the treatment takes effect, which is ignored in (2.17). This is what we will consider next.

To make the problem more tractable, we assume no interference and no anticipation effect (Athey and Imbens, 2018) in the following sense:

Assumption 9. For all $i \in [N]$ and $t \in [T]$, $Y_{it}(\boldsymbol{z})$ only depends on $\boldsymbol{z}$ through $\boldsymbol{z}_{i,[t]}$, the treatment history of unit i.

The hypothesis we consider assumes that the lag l treatment effect is equal to a constant τ_l (l is a fixed integer between 0 and $T-1$). More precisely, consider the following sequence of partially sharp null hypotheses,

$$H^{(t)} : Y_{i,t+l}((\mathbf{0}_{t-1}, 1, \mathbf{0}_l)) - Y_{i,t+l}(\mathbf{0}_{t+l}) = \tau_l, \ \forall i \in [N], \tag{2.18}$$

for $t = 1, \ldots, T-l$. The hypothesis $H^{(t)}$ states that the treatment that cross-overs at time t has a constant effect τ_l on the outcome at time $t+l$ for all the units. We are interested in testing the intersection of $H^{(1)}, \ldots, H^{(T-l)}$. Note that this intersection is not the same as the fully sharp hypothesis in (2.17) as it only concerns the lag-l effect.

2.4.2 Multiple conditional randomization tests for testing lagged treatment effects

We first note that $H^{(t)}$ is partially sharp. In fact, if we denote the imputability mapping of $H^{(t)}$ as $\mathcal{H}^{(t)}(\boldsymbol{z}, \boldsymbol{z}^*)$ (see Definition 1), then unit has imputable potential outcome (i.e., $i \in \mathcal{H}^{(t)}(\boldsymbol{z}, \boldsymbol{z}^*)$) if and only if i crosses over either at time t or after $t+l$ under both assignments $\boldsymbol{z}$ and $\boldsymbol{z}^*$. Therefore,

$$\mathcal{H}^{(t)}(\boldsymbol{z}, \boldsymbol{z}^*) = \left\{ i \in [N] : \boldsymbol{z}_{i,[t+l]}, \boldsymbol{z}^*_{i,[t+l]} \in \{(\mathbf{0}_{t-1}, 1, \mathbf{0}_l), \mathbf{0}_{t+l}\} \right\}.$$

This motivates us to only use the units that have imputable potential outcomes. We let the CRT for $H^{(t)}$ use the conditioning set

$$\mathcal{S}_{\boldsymbol{Z}}^{(t)} = \{\boldsymbol{z}^* \in \mathcal{Z} : g^{(t)}(\boldsymbol{z}^*) = g^{(t)}(\boldsymbol{Z})\},$$

where $g^{(t)}(\boldsymbol{Z})$ is a subset of units that cross over at time t or after $t+l$,

$$g^{(t)}(\boldsymbol{Z}) = \left\{i \in [N] : \boldsymbol{Z}_{i,[t+l]} = (\mathbf{0}_{t-1}, 1, \mathbf{0}_l) \text{ or } \mathbf{0}_{t+l}\right\}. \tag{2.19}$$

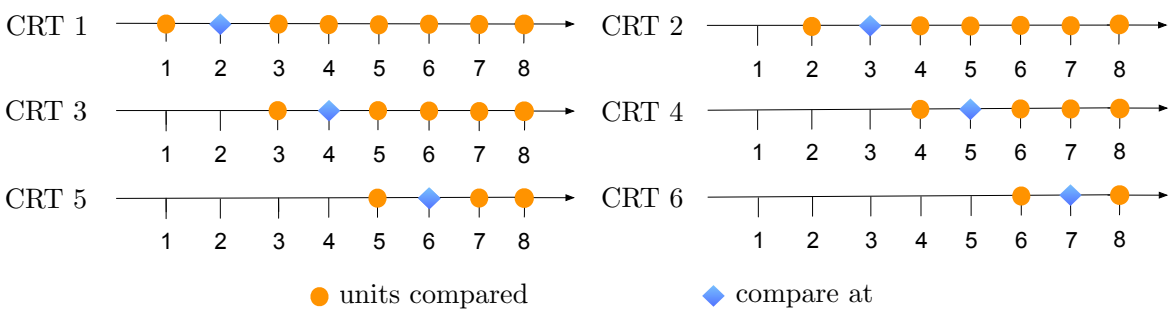
and use Proposition 1.

The conditioning sets $\mathcal{S}_{\boldsymbol{z}}^{(t)}, \boldsymbol{z} \in \mathcal{Z}$, form a partition $\mathcal{R}^{(k)} = \{\mathcal{S}_{\boldsymbol{z}}^{(t)} : \boldsymbol{z} \in \mathcal{Z}\}$ of $\mathcal{Z}$ based on the value of $g^{(t)}$. We may then compute the p-value (2.1) since both outcomes in (2.18) are either observed or imputable. Since the treatment assignment is completely randomized, the CRT is essentially a permutation test with $\mathcal{S}_{\boldsymbol{Z}}$ consisting of all the possible permutations of units that cross over at time t and after $t+l$.

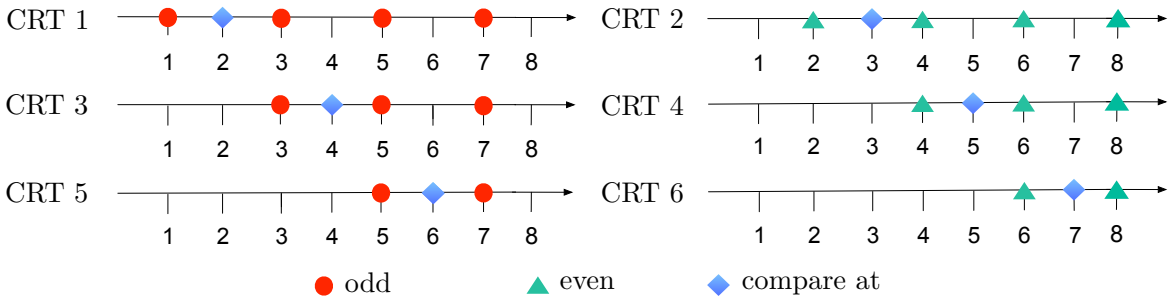
However, a careful examination reveals that the conditioning sets $\mathcal{S}_{\boldsymbol{Z}}^{(1)}, \ldots, \mathcal{S}_{\boldsymbol{Z}}^{(T-l)}$ fail to satisfy the nested condition (2.10) in Theorem 2 when lag $l \geq 1$. Consider the case of $l = 1$. The t-th CRT (i.e. CRT t) is a permutation test that compares the time $t+1$ outcome of units that cross over at t with the units that cross over after $t+1$: (i) The first CRT compares units that cross over at time 1 and units that cross over at time $3, 4, \ldots$; (ii) The second CRT compares units that cross over at time 2 and units that cross over at time $4, 5 \ldots$; and so on (see Figure 2.2a for an illustration). It is easy to see that the corresponding conditioning sets are not nested. For example, the conditioning set of CRT 1 includes cross over (i.e. starting the treatment) at time 3 but not time 2, while the reverse is true for the conditioning set of CRT 2. The independence of the CRTs thus cannot be established.

The discussion above also suggests that no single CRT can be constructed to test the intersection of the hypotheses $H^{(t)}, t = 1, \ldots, T$ at multiple times. For instance, suppose we would like to test $H^{(1)} \cap H^{(2)}$ with lag $l = 1$ by using potential outcomes at $t = 2$ and 3. Notice that the potential outcome at time 3, $Y_{i3}(\boldsymbol{Z}_i^*)$ is not imputable when $\boldsymbol{Z}_{i,1} = 1$ and $\boldsymbol{Z}_{i,2}^* = 1$ under because the lag is $l = 1$ (see CRT2 in Figure 2.2a), so units treated at time 1 must be excluded from the test (by conditioning on $\boldsymbol{Z}_{i,1} = 0$). The dilemma is that those are precisely the treated units whose outcomes are useful for testing the lag 1 effect at time 2.

This inspires us to modify the tests for the $l = 1$ case as follows: (i) The first CRT compares units that cross over at time 1 and units that cross over at time $3, 5, \ldots$; (ii) The second CRT compares units that cross over at time 2 and units that cross over at time $4, 6, \ldots$; (iii) The third CRT compares units that cross over at time 3 and units that cross over at time $5, 7, \ldots$; and so on (see Figure 2.2b for an illustration). By considering smaller



(a) Non-nested CRTs.



(b) Nested CRTs.

Figure 2.2 Illustration of the CRTs for the lagged treatment effect in a stepped-wedge randomized trial with $T = 8$ time points (lag $l = 1$).

conditioning sets, the odd tests become a sequence of nested CRTs and so do the even tests. Thus, the nesting condition (2.10) in Theorem 2 holds. Moreover, because all the CRTs further condition on

$$g(\boldsymbol{Z}) = \{i \in [N] : Z_{i,t} = 1 \text{ for some odd } t\},$$

it is not difficult to show that

$$(\boldsymbol{Z}_1, \boldsymbol{Z}_3, \dots) \perp\!\!\!\perp (\boldsymbol{Z}_2, \boldsymbol{Z}_4, \dots) \mid g(\boldsymbol{Z}). \tag{2.20}$$

For every t, the t-th CRT only depends on $\boldsymbol{Z}$ through $\boldsymbol{Z}_t$ and uses the randomization distribution of $\boldsymbol{Z}_t$ given $\boldsymbol{Z}_{[t-1]}$ and $g(\boldsymbol{Z})$. Then it is straightforward to verify that every pair of odd tests are conditionally independent and thus satisfy the condition (2.11) in Theorem 2. Similarly, (2.11) also holds for the even tests. Finally, any pair of odd and even tests satisfy (2.11) due to (2.20). Therefore, the p-values of the modified CRTs satisfy (2.13) and can be combined using the standard global testing methods such as Fisher's method (Fisher, 1925).

Algorithm 1 Multiple conditional randomization tests (MCRTs) for testing lag-l treatment effect in stepped-wedge randomized controlled trials

Input: Number of units N, Number of time steps T, Time lag l, Outcomes $\boldsymbol{Y} = (Y_{it} : i \in [N], t \in [T])$, Treatment assignments $\boldsymbol{Z} = (Z_{it} : i \in [N], t \in [T])$, Test statistics T.
Initialization: $J \leftarrow \min(l+1, T-l-1)$ and $\mathcal{I}_t \in \{i \in [N] : Z_{it} = 1\}, \forall t \in [T]$
for $j \in [J]$ **do** ▷ Divide the time steps $[T]$ into J subsets
    $t \leftarrow j$, $\mathcal{C}_j \leftarrow \{t\}$
    while $t + l + 1 \leq T$ **do**
        $t \leftarrow t + l + 1$, $\mathcal{C}_j \leftarrow \mathcal{C}_j \cup \{t\}$
for $j \in [J]$ **do**
    for $k \in \mathcal{C}_j$ **do** ▷ Define the k-th permutation test
        $\mathcal{T}^{(k)} \leftarrow \{t \in \mathcal{C}_j : t \geq k\}$
        $\mathcal{I}_1^{(k)} \leftarrow \mathcal{I}_k$, $\boldsymbol{Y}_{\text{treated}}^{(k)} \leftarrow \{Y_{i,k+l} : i \in \mathcal{I}_1^{(k)}\}$ ▷ Create a treated group
        $\mathcal{I}_0^{(k)} \leftarrow (\bigcup_{t \in \mathcal{T}^{(k)}} \mathcal{I}_t) \setminus \mathcal{I}_k$, $\boldsymbol{Y}_{\text{control}}^{(k)} \leftarrow \{Y_{i,k+l} : i \in \mathcal{I}_0^{(k)}\}$ ▷ Create a control group
        $P^{(k)}(\boldsymbol{Z}, \boldsymbol{Y}) \leftarrow \text{Permutation Test}(\boldsymbol{Y}_{\text{treated}}^{(k)}, \boldsymbol{Y}_{\text{control}}^{(k)}; T)$
Output: P-values $\{P^{(k)} := P^{(k)}(\boldsymbol{Z}, \boldsymbol{Y})\}$

The main idea in the discussion above is that we can further restrict the conditioning sets when the most obvious conditioning sets are not nested. This argument can be easily extended to a longer time lag l. When $l = 1$, we have divided $[T]$ into two subsets of cross-over times (odd and even). To test the lag l effect for $l > 1$, we can simply increase the gap and divide $[T]$ into disjoint subsets: $\mathcal{C}_1 = \{1, l+2, 2l+3, \dots\}, \mathcal{C}_2 = \{2, l+3, 2l+4, \dots\}, \dots$. A formal algorithm for general l is given in Algorithm 1 and will be referred to as multiple conditional randomization tests (MCRTs) in what follows.

2.4.3 Method of combining p-values

A remaining question is how to combine the permutation tests obtained from Algorithm 1 (MCRTs) efficiently. Heard and Rubin-Delanchy (2018) compared several p-value combination methods in the literature. By recasting them as likelihood ratio tests, they demonstrated that the power of a combiner crucially depends on the distribution of the p-values under the alternative hypotheses. In large samples, the behaviour of permutation tests is well studied in the literature. Lehmann and Romano (2006, Theorem 15.2.3) showed that if the test statistics in a permutation test converges in distribution, the permutation distribution will converge to the same limiting distribution in probability. These results form the basis of our investigation of combining CRTs.

Suppose that K permutation p-values $P^{(1)}, \ldots, P^{(K)}$ are obtained from from Algorithm 1. Suppose the kth test statistic, when scaled by $\sqrt{N}$, is asymptotically normal so that it converges in distribution to $T_\infty^{(k)} \sim \mathcal{N}(\tau_l, V_\infty^{(k)})$ under the null hypothesis ($\tau_l = 0$) and some local alternative hypothesis ($\tau_l = h/\sqrt{N}$) indexed by h. The limiting distributions have the same mean but different variances. Suppose that we define the p-value likelihood ratio as the product of the alternative p-value distributions for all k divided by the product of the null p-value distributions for all k. Since the permutation p-values are standard uniform[10], it is easy to verify that the logarithm of this p-value likelihood ratio is proportion to a weighted sum of Z-scores, $T_\infty = \sum_{k=1}^K w_\infty^{(k)} \Phi^{-1}(P^{(k)})$. This motivates a weighted version of Stouffer's method (Stouffer et al., 1949) for combing the p-values. The weights are non-negative and sum to one, thus $T_\infty \sim \mathcal{N}(0,1)$ if the p-values are independent. Then we can reject $\bigcap_{k\in[K]} H_0^{(k)}$ if $\Phi(T_\infty) \leq \alpha$. This test has been shown to be uniformly most powerful for all h in the normal location problem (Heard and Rubin-Delanchy, 2018).

To formalize the discussion above, next we consider the difference-in-means statistics as an example. Following the notation in Algorithm 1 (see lines 6 and 11), the treated group $\mathcal{I}_1^{(k)}$ for the kth CRT crosses over at time k and the control group $\mathcal{I}_0^{(k)}$ crosses over time $t \in \mathcal{T}^{(k)} \setminus \{k\}$, where $\mathcal{T}^{(k)} \setminus \{k\}$ is a subset of $\mathcal{C}_j$ that collects some time points after $k+l$. Let $N_1^{(k)} = |\mathcal{I}_1^{(k)}|$, $N_0^{(k)} = |\mathcal{I}_0^{(k)}|$ and $N^{(k)} = N_0^{(k)} + N_1^{(k)}$ Let $A_i^{(k)} = 1$ or 0 denote unit $i \in \mathcal{I}_0^{(k)} \cup \mathcal{I}_1^{(k)}$ starting the treatment at time k or after $k+l$, i.e., $\boldsymbol{Z}_{i,k+l} = (\mathbf{0}_{k-1}, 1, \mathbf{0}_l)$ or $\mathbf{0}_{k+l}$. To simplify the exposition, we use the abbreviations $Y_i^{(k)} = Y_{i,k+l}$, $Y_i^{(k)}(1) = Y_{i,k+l}(\mathbf{0}_{k-1}, 1, \mathbf{0}_l)$ and $Y_i^{(k)}(0) = Y_{i,k+l}(\mathbf{0}_{k+l})$ in the results below. We further assume that the N units are i.i.d draws from a super-population model, and every unit's potential outcome $Y_i^{(k)}(a)$ is a random copy of some generic $Y^{(k)}(a)$ for $a \in \{0,1\}$ and $k \in K$.

Proposition 8. *Suppose that the K permutation tests in Algorithm 1 use the difference-in-means statistics. We assume that for every $k \in [K]$ and $a \in \{0,1\}$,*

(i) $Y^{(k)}(a)$ has a finite and nonzero variance;

(ii) $N_1^{(k)}/N$ and $N_0^{(k)}/N$ are fixed as $N \to \infty$;

The weighted Z-score combiner T_∞ is then given with weights $w_\infty^{(k)} = \sqrt{\frac{\Lambda^{(k)}}{\sum_{j=1}^K \Lambda^{(j)}}}$, where

$$\Lambda^{(k)} = \left(\frac{N}{N_0^{(k)}} \operatorname{Var}\left[Y^{(k)}(1)\right] + \frac{N}{N_1^{(k)}} \operatorname{Var}\left[Y^{(k)}(0)\right] \right)^{-1} \tag{2.21}$$

is the inverse of the asymptotic variance $V_\infty^{(k)}$ of the statistics in the k-th test.

[10] The p-value from a permutation test is exactly standard uniform by further randomizing the rejection when the p-value is close to any $\alpha \in [0,1]$; see Lehmann and Romano (2006, Equation 5.51).

We can estimate Var $\left[Y^{(k)}(a)\right]$ in $\Lambda^{(k)}$ consistently using the sample variance of $Y_i^{(k)}, \forall i \in \mathcal{I}_a^{(k)}$. We denote the estimator of $\Lambda^{(k)}$ by $\hat{\Lambda}^{(k)}$. The empirical version of T_∞ is given by

$$\hat{T} = \sum_{k=1}^{K} \hat{w}^{(k)} \Phi^{-1}(P^{(k)}) \quad \text{where} \quad \hat{w}^{(k)} = \sqrt{\frac{\hat{\Lambda}^{(k)}}{\sum_{j=1}^{K} \hat{\Lambda}^{(j)}}}. \tag{2.22}$$

Proposition 9. *Suppose that the p-values $P^{(1)}, \ldots, P^{(K)}$ from Algorithm 1 are valid, the combined p-value $\hat{P}(\boldsymbol{Z}, \boldsymbol{Y}) = \Phi(\hat{T})$ is also valid such that under the null hypotheses,*

$$\mathbb{P}\{\hat{P}(\boldsymbol{Z}, \boldsymbol{Y}) \leq \alpha\} \leq \alpha.$$

In general, maximizing the statistical power in combining multiple CRTs is a independent task on top of choosing the most efficient test statistics for the permutation tests. Different test statistics may have different asymptotical distributions and the optimal p-value combiners (if exist) may be different. Nonetheless, a key insight from the discussion above is that we should weight the CRTs appropriately, often according to their sample sizes.

2.5 Experiments

Next, we examine the empirical performance of the methods proposed in Section 2.4 by comparing their power with Bonferroni's Method (Simulation I), checking their validity when there are unit-by-time interactions (Simulation II), and investigating two real trial datasets. We let MCRTs+F and MCRTs+Z denote the combination of MCRTs with Fisher's method (Fisher, 1925) and the weighted Z-score method introduced above, respectively. In both MCRTs+F and MCRTs+Z, the test statistic is the simple difference-in-means.

2.5.1 Simulation I: Size and power

As illustrated in Figure 2.2, the non-nested CRTs have larger control groups than our nested CRTs created in MCRTs. However, the non-nested CRTs are not independent and there are limited ways to combine them. The most widely used method in this case is Bonferroni's method that rejects the intersection of K hypotheses if the smallest p-value is less than α/K. One remaining question is whether the reduced sample size in MCRTs can indeed be compensated by a better p-value combiner. Simulation I is designed to answer this question empirically by investigating varying the sample size N, trial length T, time lag l, and effect sizes τ_l, respectively.

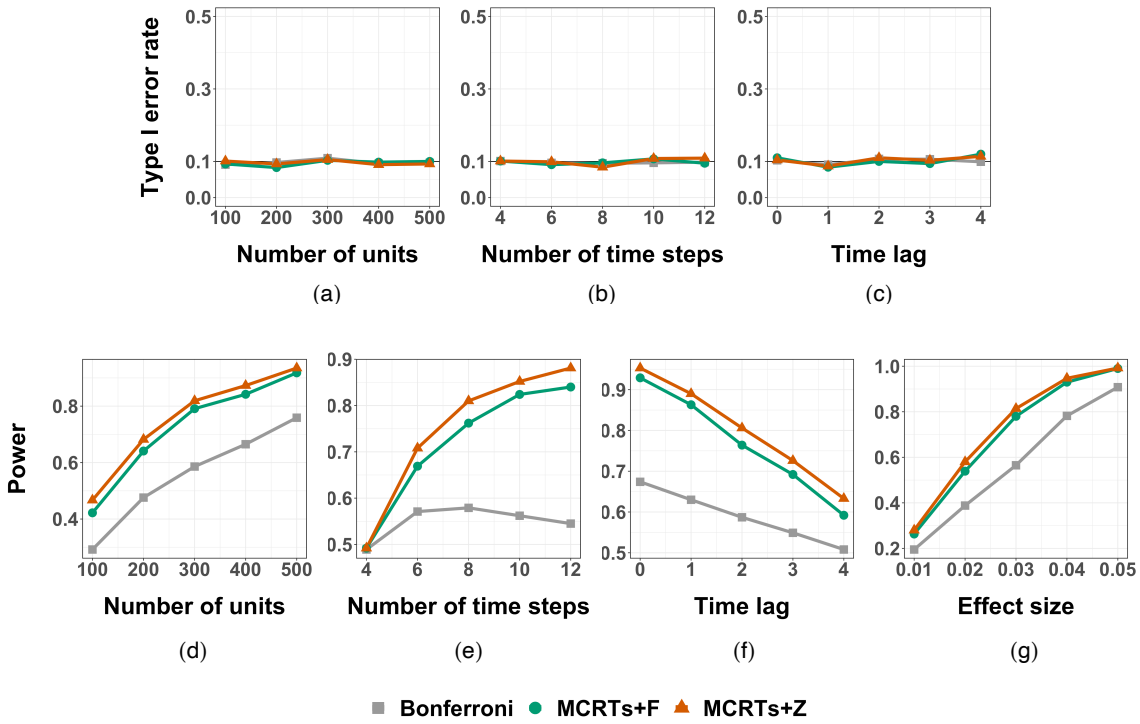


Figure 2.3 Performance of MCRTs+F, MCRTs+Z and Bonferroni's method: type I error rates and powers in testing lagged effects at five different numbers of units, numbers of time steps, time lags and effect sizes. The results were averaged over 1000 independent runs.

In this simulation, we fixed $N_1 = \ldots = N_T = N/T$ and $N_T = N - \sum_{t=1}^{T-1} N_t$. The treatment assignment $\boldsymbol{Z}$ was randomly drawn from $\mathcal{Z}$ given in (2.16). Let A_i to denote the treatment starting time of unit i, i.e., the index of the non-zero entry of Z_i. Outcomes were generated by a linear mixed-effects model,

$$Y_{it} = \mu_i + 0.5(X_i + t) + \sum_{l=0}^{T-1} 1_{\{A_i - t = l\}} \tau_l + \epsilon_{it}, \ i = 1, \ldots, N, \ t = 0, \ldots, T,$$

where $\mu_i \sim \mathcal{N}(0, 0.25)$, $X_i \sim \mathcal{N}(0, 0.25)$ and $\epsilon_{it} \sim \mathcal{N}(0, 0.1)$. This assumes that the baseline outcome Y_{i0} is measured, which is not uncommon in real clinical trials. Basic parameters were varied in the following ranges: (i) number of units $N \in \{100, 200, 300, 400, 500\}$; (ii) number of time steps $T \in \{4, 6, 8, 10, 12\}$; (iii) time lag $l \in \{0, 1, 2, 3, 4\}$; effect size $\tau_l \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Empirical performance of the methods was examined when one of N, T, and l is changed while the other two are fixed at the median of their ranges, respectively. For example, we increased N from 100 to 500 while keeping $T = 8$ and $l = 2$. In these simulations, τ_l was set to 0 and 0.03 to investigate type I error and power of the methods, respectively. One more simulation was created to study the power as the effect size τ_l varies, in

which we keep the first three basic parameters at their median values ($N = 300, T = 8, l = 2$) and increase τ_l from 0.01 to 0.05. For all permutation tests, difference-in-means was used as the test statistic, and the number of permutations was fixed at 1000.

The upper panels of Figure 2.3 show that all the methods control the type I errors at any number of units, time steps and time lags. The lower panel shows that our methods are more powerful than Bonferroni's method in all the simulations. MCRTs+Z is slightly more powerful than MCRTs+F in all experiments. This shows that the weighted z-score is more effective than Fisher's combination method for MCRTs. Panels (d) and (g) show that our methods outperform Bonferroni's method by a wider margin as the sample size or the effect size increases. Panel (e) establishes the same observation for trial length, which can be explained by the fixed sample size (so fewer units start treatment at each step as trial length increases). This is not an ideal scenario for applying Bonferroni's method, as the individual tests have diminishing power. Finally, panel (f) shows that all the methods have smaller power as the lag size increases, and our methods are particularly powerful when the time lag is small. These results are due to facts. First, there are fewer permutation tests available for larger time lags. Moreover, only including the units that are treated after a large time lag, the control groups in the permutation tests are small.

Overall, these simulations support the conclusion that the sample splitting in MCRTs is a worthy sacrifice as a more powerful p-value combiner can be applied. Note that every outcome relevant to the lag l effect is still used in at least one of our permutation tests in MCRTs. So the reduced sample size in MCRTs may not be as damaging as it might first appear.

2.5.2 Simulation II: Overcome misspecification of mixed-effects models

Simulation II is designed to investigate the finite-sample properties of confidence intervals obtained by inverting conditional randomization tests (see Section 2.9 for more details). In particular, we are interested in comparing its efficiency and robustness with mixed-effects models for estimating lagged treatment effect.

In Simulation II, the treatment generating process was kept the same as Simulation I. Simulation parameters are set as $N = 200$, $T = 8$ and lagged effects $(\tau_0, \ldots, \tau_7) = (0.1, 0.3, 0.6, 0.4, 0.2, 0, 0, 0)$. The treatment effect is gradually realized and then decayed to 0 over time. For every $i \in [N]$ and $t = 0, 1, \ldots, T$, we generate the outcome Y_{it} using a

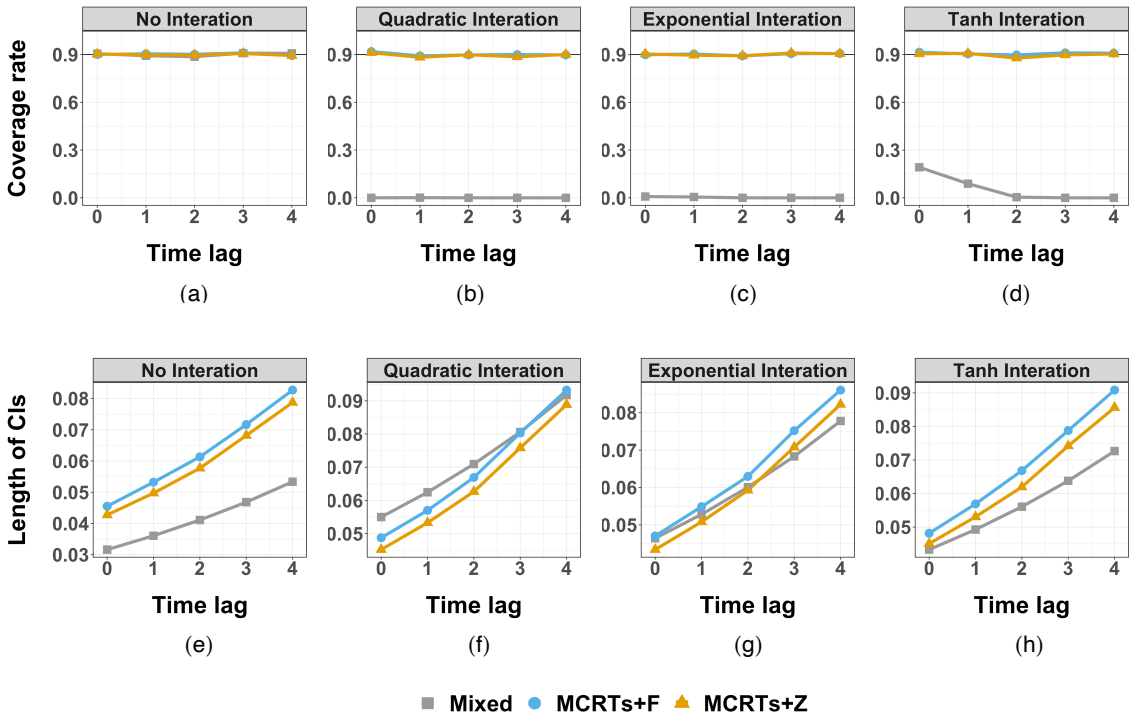


Figure 2.4 Performance of MCRTs+F, MCRTs+Z and the mixed-effects model: coverage rates and lengths of confidence intervals (CIs) under the outcome generating processes with no interaction effect and with three different types of covariate-and-time interactions. The results are averaged over 1000 independent runs.

mixed-effects model,

$$Y_{it} = \mu_i + 0.5(1 - 0.1 \cdot 1_{\{m \neq 0\}})(X_i + t) + 0.1 f_m(X_i + t) + \sum_{l=0}^{7} 1_{\{A_i - t = l\}} \tau_l + \epsilon_{it},$$

where $\mu_i \sim \mathcal{N}(0, 0.25)$, $X_i \sim \mathcal{N}(0, 0.25)$, $\epsilon_{it} \sim \mathcal{N}(0, 0.1)$ and the unit-by-time interaction is given by

$$f_m(X_i + t) = \begin{cases} 0, & \text{if } m = 0, \\ (X_i + t)^2, & \text{if } m = 1, \\ 2\exp[(X_i + t)/2], & \text{if } m = 2, \\ 5\tanh(X_i + t), & \text{if } m = 3, \end{cases}$$

which correspond to no, quadratic, exponential and hyperbolic tangent interactions.

We tasked our methods (MCRTs+F and MCRTs+Z) and a mixed-effects model described below to construct valid 90%-confidence interval (CIs) for the lagged effects $\tau_0, \ldots \tau_4$. Our methods were implemented in exactly the same way as in Simulation I except that the

permutation tests were inverted to obtain interval estimators. The mixed-effects model takes the form

$$Y_{it} = \beta_{0i} + \beta_1 x_i + \beta_2 t + \sum_{l=0}^{7} \xi_l 1_{\{A_i - t = l\}} + \epsilon_{it}, \tag{2.23}$$

where β_{0i} is a random unit effect, β_1, β_2 are fixed effects and $\xi_0, \ldots, \xi_7$ are lagged effects, and were fitted using the R-package `lme4` (Bates et al., 2015). Recently, Kenny et al. (2021) proposed mixed-effects models that can leverage the shapes of time-varying treatment effects. Their models are given by specifying some parametric effect curves with the help of basis functions (e.g. cubic spline). Besides the unit-by-time interaction, the model (2.23) is already correctly specified for modelling the treatment effects and other parts of the outcome generating process above. Excluding some of the lagged effect parameters $\xi_5, \ldots, \xi_7$ from the model may change the effect estimates but not the validity of its CIs.

As the unit-by-time interactions are unknown and fully specifying them would render the model unidentifiable, they are typically not considered in mixed-effect models. If the stepped-wedge trial is randomized at the cluster levels, one can introduce cluster-by-time interaction effects in a mixed-effects model; see Ji et al. (2017, Equation 3.1) as an example. However, the cluster-by-time interactions are not exactly the same as the interaction between time and some covariates of the units. It depends on if the interaction varies within each cluster and the clusters are defined by the covariates which interact with time.

The results of Simulation II are reported in Figure 2.4. The upper panels show that the CIs of the mixed-effects model only achieve the target coverage rate of 90% for all the lagged effects when there is no unit-by-time interaction (panel a). By contrast, the CIs of our methods fulfil the target coverage rate of 90% under all scenarios. The lower panels show that the valid CIs given by our methods have reasonable lengths between 0.05 and 0.10 for covering the true lagged effects $(\tau_0, \ldots, \tau_4) = (0.1, 0.3, 0.6, 0.4, 0.2)$. We also note from panel (e) that when the linear mixed-effects model is specified correctly, it gives valid CIs that are narrower than our nonparametric tests.

2.5.3 Real data applications

We next demonstrate an application of MCRTs+Z in comparison to mixed-effects models (with and without time effect parameters) on real data collected from two stepped-wedge randomized trials. The results are reported as 90%-confidence intervals (CIs) of lagged effects in Figure 2.5.

Trial I was conducted to examine if a chronic care model was more effective than usual care in the Netherlands between 2010 and 2012 (Muntinga et al., 2012). The study consisted

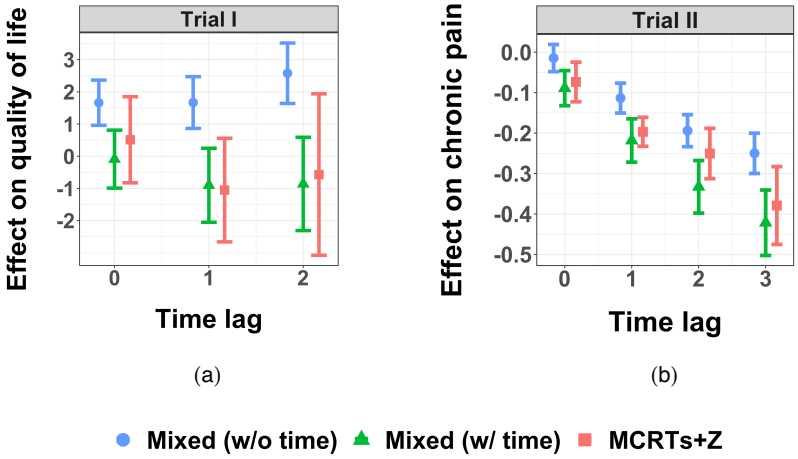


Figure 2.5 Effect estimates from MCRTs+Z and mixed-effects models with and without time effect parameters: 90%-confidence intervals (CIs) of lagged effects on real data collected from four different stepped-wedge randomized trials.

of 1,147 frail older adults in 35 primary care practices. The primary outcome in the dataset was quality of life measured by a mental health component score (MCS) in a 12-item Short Form questionnaire (SF-12). The outcome for each adult was measured at a baseline time point and every 6 months over the study. The trial randomly assigned some practices to start the chronic care model every 6 months. We treated each practice as an experimental unit. We used MCRTs+Z and two mixed-effects models to estimate lag $0, 1$ and 2 effects on the average outcome of each practice, $Y_{it}, i = 1, \ldots, 35$ and $t = 0, 1, \ldots, 4$. In panel (a) of Figure 2.5, the mixed-effects model without using a time effect parameter would conclude that the chronic care model improved the quality of life significantly. However, as noted by Twisk (2021, Section 6.3.1), this model fails to take into account the increase in the quality of life over time, irrespective of the intervention. In contrast, adding a time effect parameter to the mixed-effects model and applying MCRTs+Z produced similar CIs, both showing that the effect of the chronic care model on quality of life is not statistically significant.

Trial II was a stepped-wedge trial performed over the pain clinics of 17 hospitals to estimate the effectiveness of an intervention in reducing chronic pain for patients (Twisk, 2021, Chapter 6.4). The outcome was a pain score from 1 (least pain) and 5 (most pain) and six measurements were taken over time: one at the baseline and five more equally spaced over a period of twenty weeks. After each measurement, the intervention was started in a few randomly selected untreated hospitals. We considered each hospital as an experimental unit and estimated the lag $0, 1, 2$ and 3 effects using hospital-level average outcomes. Panel (b) of Figure 2.5 shows that the CIs of the mixed-effect models with and without a time effect do

not overlap. The CIs of MCRTs+Z lie roughly between those, showing that the effectiveness of the intervention appears to increase over time.

2.6 Conditional randomization tests in the literature

In this section, we apply the general theory in Sections 2.2 and 2.3 to better understand some (conditional) randomization tests that have been proposed in the literature.

2.6.1 Permutation tests for treatment effect

Permutation test is the most well known form of conditional randomization tests. In a permutation test, the p-value is obtained by calculating all possible values of the test statistics under all possible permutations of the observed data points. In our setup, this amounts to use the conditioning sets (suppose $\boldsymbol{Z}$ is a vector of length N)

$$\mathcal{S}_{\boldsymbol{z}} = \{(z_{\sigma(1)}, \ldots, z_{\sigma(N)}) : \sigma \text{ is a permutation of } [N]\}. \tag{2.24}$$

In view of Proposition 1, permutation test is a CRT that conditions on the order statistics of $\boldsymbol{Z}$. In permutation tests, the treatment assignments are typically assumed to be exchangeable,

$$(Z_1, \ldots, Z_N) \overset{d}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(N)}) \text{ for all permutations } \sigma \text{ of } [N],$$

so each permutation of $\boldsymbol{Z}$ has the same probability of being realized under the treatment assignment mechanism $\pi(\cdot)$. See Kalbfleisch (1978) for an alternative formulation of rank-based tests based on marginal and conditional likelihoods. Examples of exchangeable assignments include repeated Bernoulli trials and simple random sampling with or without replacement (let Z_i be the number of times unit i is sampled). Exchangeability makes it straightforward to compute the p-value (2.1), as $\boldsymbol{Z}^*$ is uniformly distributed over $\mathcal{S}_{\boldsymbol{Z}}$ if $\boldsymbol{Z}$ has distinct elements. In this sense, our assumption that the assignment distribution of $\boldsymbol{Z}$ is known (Assumption 8) is more general than exchangeability. See Roach and Valdar (2018) for some recent development on generalized permutation tests in non-exchangeable models.

Notice that in permutation tests, the invariance of $\mathcal{S}_{\boldsymbol{z}}$ in Lemma 1 is satisfied because the permutation group is closed under composition, that is, the composition of two permutations of $[N]$ is still a permutation of $[N]$. This property can be violated when the test conditions on additional events. Southworth et al. (2009) gave a counterexample in which $\boldsymbol{Z}$ is randomized uniformly over $\mathcal{Z} = \{\boldsymbol{z} \in \{0,1\}^N : \boldsymbol{z}^T\mathbf{1} = N/2\}$, so exactly half of the units are treated. They consider the "balanced permutation test" in (Efron et al., 2001, Section 6) that uses

the following conditioning set

$$\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z}^* : \boldsymbol{z}^* \text{ is a permutation of } \boldsymbol{z} \text{ and } \boldsymbol{z}^T\boldsymbol{z}^* = N/4\}. \tag{2.25}$$

Southworth et al. (2009) showed that the standard theory for permutation tests in (Lehmann and Romano, 2006) does not establish the validity of the balanced permutation tests, because $\mathcal{S}_{\boldsymbol{z}}$ is not a group under balanced permutations (nor is $\mathcal{S}_{\boldsymbol{z}} \cup \{\boldsymbol{z}\}$). They also provided numerical examples in which the balanced permutation test has an inflated type I error. With the general theory in Section 2.2 in mind, (2.25) clearly does not satisfy the invariance property in Lemma 1. Moreover, a group structure is not necessary in the construction of a valid randomization test in Section 2.2. The crucial algebraic structure is that $\mathcal{R} = \{\mathcal{S}_{\boldsymbol{z}} : \boldsymbol{z} \in \mathcal{Z}\}$ must be a partition of $\mathcal{Z}$, or equivalently that the conditioning set $\mathcal{S}_{\boldsymbol{z}}$ must be defined by an equivalence relation. This is clearly not satisfied by (2.25).

2.6.2 Permutation tests for independence

Randomization tests are frequently used to test the independence of random variables. In this problem, it is typically assumed that we observe independent and identically distributed random variables $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ and would like to test the null hypothesis that Z_1 and Y_1 are independent. In the classical treatment of this problem (Lehmann and Romano, 2006), the key idea is to establish the following *permutation principle*:

$$(Z_1, \ldots, Z_N, Y_1, \ldots, Y_N) \overset{d}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(N)}, Y_1, \ldots, Y_N) \text{ for all permutations } \sigma \text{ of } [N].$$

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$, and $\boldsymbol{Z}_\sigma = (Z_{\sigma(1)}, \ldots, Z_{\sigma(N)})$. Then given a test statistic $T(\boldsymbol{Z}, \boldsymbol{Y})$, independence is rejected if the following p-value is less than the significance level (recall that there are $N!$ permutations of $[N]$):

$$P(\boldsymbol{Z}, \boldsymbol{Y}) = \frac{1}{N!} \sum_{\sigma} 1_{\{T(\boldsymbol{Z}_\sigma, \boldsymbol{Y}) \leq T(\boldsymbol{Z}, \boldsymbol{Y})\}}. \tag{2.26}$$

We are using the same notation $P(\boldsymbol{Z}, \boldsymbol{Y})$ for the p-value of this permutation test in (2.26), because they are indeed identical. It might seem that the permutation test is solving a different problem from testing a causal null hypothesis—after all, no counterfactuals are involved in testing independence. In fact, Lehmann (1975) referred to the causal inference problem as the *randomization model* and the independence testing problem as the *population model*. Ernst (2004) argued that the reasoning behind these two models is different.

However, a statistical test not only explicitly tests the null hypothesis H_0 but also implicitly tests any assumptions needed to set up the problem. Therefore, the CRT described

in Section 2.2 tests not only a partially sharp null hypothesis about the causal effect but also the assumption that the treatment is randomized (Assumption 8). If we simply "define" the potential outcomes as $\boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{Y}$ for all $\boldsymbol{z} \in \mathcal{Z}$ so $\boldsymbol{W}$ consists of many identical copies of $\boldsymbol{Y}$, the "causal" null hypothesis $H_0 : \boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{Y}(\boldsymbol{z}^*), \forall \boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$ would be automatically satisfied. Recall that the p-value (2.1) of a CRT is given by

$$P(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{W}) \leq T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}, \boldsymbol{W}\},$$

and suppose the test statistic is given by $T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{W}) = T(\boldsymbol{z}^*, \boldsymbol{Y}(\boldsymbol{z}^*))$ as in Section 2.2.5. Because of how the potential outcomes schedule $\boldsymbol{W}$ is defined in this case, the test statistics is equal to $T(\boldsymbol{z}^*, \boldsymbol{Y})$. Furthermore, the independence of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ is equivalent to the independence of $\boldsymbol{Z}$ and $\boldsymbol{W}$. When $\mathcal{S}_{\boldsymbol{Z}}$ is given by all the permutations of $\boldsymbol{Z}$ as in (2.24), $\boldsymbol{Z}^* \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}$ has the same distribution as $\boldsymbol{Z}_\sigma$ where σ is a random permutation. Combining these observations, we see that the permutation test of independence is identical to the conditional randomization test of no causal effect.

In summary, the two formulations of permutation (or more broadly, randomization) tests are unified in our framework. The CRT tests not only a partially sharp null hypothesis about the potential outcomes schedule $\boldsymbol{W}$ but also the independence of $\boldsymbol{Z}$ and $\boldsymbol{W}$ (Assumption 8). In the randomization model, $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ is guaranteed by randomization, so the CRT gives a test of the causal null hypothesis. In the population model, the causal null hypothesis about $\boldsymbol{W}$ is automatically satisfied by definition, so the CRT gives a test of independence. We would like to stress that such a unification is only a mathematical one; the logic and philosophy behind the two formulations are fundamentally different. In the randomization model, inference is justified by the belief that randomness introduced by the experimenter is exogenous. In the population model, inference is justified by mathematical assumptions such as the exchangeability of observations. See Section 2.7 for further discussion.

2.6.3 Randomization tests for conditional independence

Recently, there has been a growing interest in randomization tests for conditional independence (Berrett et al., 2020; Candès et al., 2018; Katsevich and Ramdas, 2020; Liu et al., 2020; Tansey et al., 2021). Similar to the last section, it is assumed that we observe independent and identically distributed random variables $(Z_1, Y_1, X_1), \ldots, (Z_n, Y_n, X_n)$ and would like the test $Z_1 \perp\!\!\!\perp Y_1 \mid X_1$. This can be easily incorporated in our framework by treating $\boldsymbol{X} = (X_1, \ldots, X_n)$ as fixed; see the last paragraph in Section 2.2.1. In this case, the randomization distribution of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ is given by the conditional distribution of $\boldsymbol{Z}$ of $\boldsymbol{X}$. The same randomization test can then be applied; see Candès et al. (2018, Section 4.1) for more detail. Berrett et al.

(2020) extended this test by further conditioning on the order statistics of $\boldsymbol{Z}$, resulting in a permutation test.

As a remark on the terminology, the test in the last paragraph was called "conditional randomization test" by Candès et al. (2018), because the procedure conditions on $\boldsymbol{X}$. However, such conditioning is fundamentally different from the post-experimental conditioning on $\mathcal{S}_{\boldsymbol{Z}}$, which we use to distinguish conditional randomization tests from unconditional ones. When $\boldsymbol{Z}$ is randomized according to $\boldsymbol{X}$, conditioning on $\boldsymbol{X}$ is obligatory in randomization inference because it needs to use the randomness introduced by the experimenter (which is conditional on $\boldsymbol{X}$). On the other hand, conditioning on $\mathcal{S}_{\boldsymbol{Z}}$ (or $g(\boldsymbol{Z})$ in Proposition 1) can be introduced by the analyst to improve the power or make the p-value computable. For this reason, we advocate reserving the terminology "conditional randomization test" for the latter case.

2.6.4 Covariate imbalance and rerandomization

Morgan and Rubin (2012) proposed to use rerandomization to improve covariate balance in experiments and regenerated some interest in randomization inference (Banerjee et al., 2017; Ding et al., 2015; Heckman and Karapakula, 2019). They recounted a conversation between Cochran and Fisher, in which Fisher suggested rerandomizing if some baseline covariates are not well balanced by the randomly chosen assignment. The key insight of Morgan and Rubin (2012) is that the experiment should then be analyzed with the rerandomization taken into account. More specifically, rerandomization is simply a rejection sampling algorithm for randomly choosing $\boldsymbol{Z}$ from

$$\mathcal{Z} = \{\boldsymbol{z} : g(\boldsymbol{z}) \leq \eta\},$$

where $g(\boldsymbol{z})$ measures the covariate imbalance implied by the treatment assignment $\boldsymbol{z}$ and η is the experimenter's level of tolerance. Therefore, we simply need to use the randomization distribution over $\mathcal{Z}$ to carry out the randomization test. This provides an excellent example for our *main thesis* that randomization inference should be based exactly on the randomization introduced by the experimenter and the analyst.

Another way to deal with unlucky draws of treatment assignment is to condition on the covariate imbalance $g(\boldsymbol{Z})$ (Hennessy et al., 2016). This inspires our Proposition 1 in Section 2.2.5. In our terminology, this is a conditional randomization test because the analyst has the liberty to choose which function $g(\boldsymbol{Z})$ to condition on. On the other hand, the randomization test proposed by Morgan and Rubin (2012) is unconditional.

2.6.5 Evidence factors for observational studies

So far our discussion has been focused on randomized experiments, but the same theory also applies to randomization inference for observational studies (Rosenbaum, 2002b). Many observational studies are analyzed under the *ignorability* or *no unmeasured confounders* assumption that the treatment is randomly assigned after conditioning on some covariates $\boldsymbol{X}$. In our notation, this means that $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{X}$. However, the distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ (often called the *propensity score*) is typically unknown and needs to be estimated (Rosenbaum and Rubin, 1983).

Alternatively, one can match the treated units with control units to recreate a paired or blocked randomized experiment. In the simplest setup with a binary treatment and pair matching, each pair has one treated unit and one control unit, but they have exactly the same covariates. So the treatment vector can be written as $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$ (suppose N is even) and is supported on

$$\mathcal{Z} = \left\{ \boldsymbol{z} \in \{0,1\}^N : z_1 + z_2 = 1, \ldots, z_{N-1} + z_N = 1 \right\}.$$

Typically, it is assumed that $\boldsymbol{Z}$ is uniformly distributed over $\mathcal{Z}$, and a randomization test can be conducted by permuting the treatment assignments within the pairs. The uniform distribution over $\mathcal{Z}$ can be justified by, for example, assuming the distribution of Z_i only depends on X_i and are independent across i (Rosenbaum, 2002b, Section 3.2).[11]

When there are concerns about unmeasured confounders, Rosenbaum (1987, 2002b) introduced a general sensitivity analysis framework of the ignorability assumption. In essence, Rosenbaum's sensitivity model allows the distribution $\boldsymbol{Z}$ to be non-uniform over $\mathcal{Z}$, but the "degree" of non-uniformness is bounded. More specifically, it assumes that

$$\frac{1}{1+\Gamma} \leq \mathbb{P}(Z_i = 1) \leq \frac{\Gamma}{1+\Gamma}, \ i = 1, \ldots, N, \tag{2.27}$$

where $\Gamma \geq 1$ is chosen by the analyst.

Let Π be the set of distributions of $\boldsymbol{Z}$ that satisfy (2.27). We denote the p-value in (2.1) as $P(\boldsymbol{Z}, \boldsymbol{Y}; \pi)$ to emphasize that it depends on the unknown randomization distribution π of $\boldsymbol{Z}$. For simplicity, we will assume that the p-value is always computable when discussing

[11] In practice, units within a pair are not always exactly matched, which complicates the randomization inference (Rosenbaum, 2002b, Section 3.6); also see Abadie and Imbens (2006) from the perspective of large sample properties of the matching estimator of average treatment effect. Another complication being omitted here is that the matched sets are typically constructed from a larger set of units. The standard permutation test can be regarded as a CRT that conditions on $\mathcal{Z}$. However, our general theory in Section 2.2 may be inapplicable because $\mathcal{Z}$ depends on treatment variables of a larger set of units and may not induce a partition. To our knowledge, this subtle issue has not been acknowledged in the literature yet.

sensitivity analysis. Because each p-value $P(\boldsymbol{Z}, \boldsymbol{Y}; \pi)$ satisfies (2.3) under the null hypothesis if π is indeed the distribution of $\boldsymbol{Z}$, by applying the union bound,

$$P(\boldsymbol{Z}, \boldsymbol{Y}) = \sup_{\pi \in \Pi} P(\boldsymbol{Z}, \boldsymbol{Y}; \pi) \tag{2.28}$$

satisfies (2.3) (stochastically dominates the uniform distribution over $[0, 1]$) if the actual distribution of $\boldsymbol{Z}$ is indeed in Π. Rosenbaum (1987, 2002b) provided analytical solutions to the optimization problem in (2.28) for a class of sign-score statistics.

In subsequent works, Rosenbaum (2010, 2011, 2017) introduced a concept for observational studies called "evidence factors". This is closely related to our theory of multiple CRTs in Section 2.3. Loosely speaking, a causal theory can usually be represented by several hypotheses. For example, the conclusion "smoking does not cause lung cancer" entails at least two hypotheses: whether or not a person smokes does not alter the risk of lung cancer, and given that the person smokes, the amount of cigarettes he or she smokes does not alter the risk of lung cancer either. Therefore, we can then reject the null causal theory if either hypothesis is rejected. When there are concerns about unmeasured confounders, we may represent the evidence factors by the p-values obtained from suitable sensitivity analyses. This typically requires us to specify a sensitivity model Π for each hypothesis and compute the upper bound in (2.28). Rosenbaum's key observation is that in certain problems, these evidence factors (p-value upper bounds) are "independent" in the sense that they satisfy (2.13) when all the null hypotheses are true. This allows one to test the global null and the partial conjunction hypotheses using standard methods (Karmakar et al., 2019). Rosenbaum (2010, 2011) provided some concrete examples of evidence factors, which essentially correspond to the sequential CRTs described in Section 2.3.3.

Rosenbaum (2017) demonstrated that the permutation groups of a general class of "independent" evidence factors have a knit product structure. Similar to our remark on (Southworth et al., 2009) in Section 2.6.1, we believe that the requirement that $\mathcal{Z}$ is a group or has a knit product structure is not necessary. Instead, the key is to construct sequential CRTs that satisfy the conditions in Proposition 6. Because the conditions in Proposition 6 do not rely on the distribution of $\boldsymbol{Z}$, the key stochastic dominance result (2.13) holds for the p-values $P^{(1)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi), \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi)$ for any distribution π of $\boldsymbol{Z}$. By computing the p-value suprema as in (2.28), we see that the evidence factors satisfy the same stochastic dominance result if the sensitivity model is correct.

2.6.6 Conformal prediction

Conformal prediction is another topic related to randomization inference that is receiving growing interest in statistics and machine learning (Lei et al., 2018a, 2013; Shafer and Vovk, 2008; Vovk et al., 2005). Consider a typical regression problem where the data points $(X_1, Y_1), \ldots, (X_N, Y_N)$ are drawn exchangeably from the same distribution and Y_N is unobserved. We would like to construct a prediction interval $\hat{\mathcal{C}}(X_N)$ for the next observation (X_N, Y_N) such that

$$\mathbb{P}(Y_N \in \hat{\mathcal{C}}(X_N)) \leq 1 - \alpha.$$

The key idea of conformal inference is that the exchangeability of $(X_1, Y_1), \ldots, (X_N, Y_N)$ allows us to test the null hypothesis $H_0 : Y_N = y$ using permutation tests. More concretely, we may fit *any* regression model to $(X_1, Y_1), \ldots, (X_{N-1}, Y_{N-1}), (X_N, y)$ and let the p-value be one minus the percentile of the absolute residual of (X_N, y) among all N absolute regression residuals. Intuitively, a large residual of (X_N, y) indicates that the prediction (X_N, y) "conforms" poorly with the other observations, so $H_0 : Y_N = y$ should be rejected.

We argue that conformal prediction is a special instance of the randomization inference described in Section 2.2. The key idea is to view random sampling as a kind of randomization. Notice that conformal prediction implicitly conditions on the unordered data points $(X_1, Y_1), \ldots, (X_N, Y_N)$. To make this more explicit, let the treatment variable Z be the order of the observations, which is a random permutation of $[N]$ or equivalently a random bijection from $[N]$ to $[N]$. The "potential outcomes" are given by

$$\boldsymbol{Y}(z) = \Big((X_{z(1)}, Y_{z(1)}), \ldots, (X_{z(N)}, Y_{z(N)}) \Big), \tag{2.29}$$

where $Y_{z(N)}$ is unobserved (that is, the observed $\boldsymbol{Y}$ is the first $2N - 1$ elements of $\boldsymbol{Y}(\boldsymbol{Z})$). The notation $\boldsymbol{Y}$ is overloaded here to be consistent with Section 2.2. The "potential outcomes schedule" in our setup is $\boldsymbol{W} = \{\boldsymbol{Y}(z) : \ z \text{ is a permutation of } [N]\}$. The null hypothesis in conformal prediction can be represented as a regression model $f(x)$ such that the missing $Y_{z(N)}$ is imputed by $f(X_{z(N)})$. This is a sharp null hypothesis and our theory in Section 2.2 can be directly used to construct a randomization test. Notice that the key randomization assumption $Z \perp\!\!\!\perp \boldsymbol{W}$ in our theory (Assumption 8) is justified here by the assumption that $(X_1, Y_1), \ldots, (X_N, Y_N)$ are exchangeable. In other words, we may view random sampling as a form of randomizing the order of the observations.

The above argument can be extended to allow "covariate shift" in the next observation X_N (Hu and Lei, 2020; Tibshirani et al., 2019). The key idea is, again, to consider the randomization involved in sampling. For this we need to consider a (potentially infinite) super-population $(X_i, Y_i)_{i \in \mathcal{I}}$. The "treatment" $Z : [N] \to \mathcal{I}$ selects which units are observed.

For example, $Z(1) = i$ means that (X_i, Y_i) is the first data point. The "potential outcomes" are still given by (2.29). By conditioning on unordered Z, we can obtain a conditional randomization test for the unobserved $Y_{z(N)}$. Covariate shift in X_N can be easily incorporated by conditioning on $(X_{Z(1)}, \ldots, X_{Z(N)})$ and deriving the conditional distribution of Z. For example, suppose the first $N-1$ units are sampled from the population according to a covariate distribution $\pi_1(\boldsymbol{x})$ and the last unit is sampled according to another covariate distribution $\pi_2(\boldsymbol{x})$, that is,

$$\begin{aligned} &\mathbb{P}(Z(k) = i) \propto \pi_1(X_i) \text{ for } k = 1, \ldots, N-1, i \in \mathcal{I}, \\ &\mathbb{P}(Z(N) = i) \propto \pi_2(X_i) \text{ for } i \in \mathcal{I}. \end{aligned} \tag{2.30}$$

For simplicity, we assume the data points are sampled without replacement. Then the unordered $Z = \{Z(1), \ldots, Z(N)\}$ is a set $\mathcal{S} = \{s_1, \ldots, s_N\}$ of N distinct units. Let $\boldsymbol{X} = (X_{s_1}, \ldots, X_{s_N})$. The conditional randomization distribution of Z is given by

$$\mathbb{P}\left(Z(N) = s_i \mid \boldsymbol{X}, \mathcal{S}\right) \propto \pi_2(X_{z(N)}) \prod_{j=1}^{N-1} \pi_1(X_{z(j)}) \propto \frac{\pi_2(X_{s_i})}{\pi_1(X_{s_i})}, \ i = 1, \ldots, N,$$

by using the fact that the product $\prod_{j=1}^{N} \pi_1(X_{z(j)}) = \prod_{j=1}^{N} \pi_1(X_{s_j})$ only depends on $\mathcal{S}$. By normalizing the probabilities, we obtain

$$\mathbb{P}\left(Z(N) = s_i \mid \boldsymbol{X}, \mathcal{S}\right) = \frac{\pi_2(X_{s_i})/\pi_1(X_{s_i})}{\sum_{j=1}^{N} \pi_2(X_{s_j})/\pi_1(X_{s_j})}, \ i = 1, \ldots, N.$$

The last display equation was termed "weighted exchangeability" in (Tibshirani et al., 2019). Again, we would like to stress that exchangeability (unweighted or weighted) is not necessary here. What is necessary is a careful consideration of the randomization $\boldsymbol{Z}$ in the problem. This could be either explicit as in a physical intervention, or implicit as in random sampling. Once the randomization distribution of $\boldsymbol{Z}$ is given and a suitable conditioning set $\mathcal{S}_{\boldsymbol{Z}}$ is found, the rest is just a straightforward application of the general CRT.

2.7 Discussion

As described in the Section 2.1, randomization inference is a mode of statistical inference that is based on randomization and nothing more than randomization. This is made precise by trichotomizing the randomness in the data and conditioning on the potential outcomes schedule (the randomness in nature). Through a number of examples, we demonstrate how conditional randomization tests can be applied to classical and modern problems. This sometimes requires an artificial definition of counterfactuals (e.g., Section 2.6.2) or viewing

random sampling as a form of randomization (e.g., Section 2.6.6). Some might argue that such a change of perspective is unnecessary, but we believe it provides a unified understanding of the literature and makes the basis of the inference clear. After all, it is easy for us statisticians to assume the data are independent and identically distributed and forget that this is only a theoretical model. On this, Fisher (1922) offered the following sobering remark: "The postulate of randomness thus resolves itself into the question, 'Of what population is this a random sample?' which must frequently be asked by every practical statistician."

The framework introduced in Sections 2.2 and 2.3 generalizes the existing ideas mentioned in Section 2.1. In Section 2.2 we have described three perspectives on conditioning in randomization tests: conditioning on a set of treatment assignments, conditioning on a σ-algebra, and conditioning on a random variable. They are useful for different purposes: the first perspective is useful when the null hypothesis is partially sharp and one needs to construct a computable p-value; the second perspective is useful to describe the nested structure of multiple CRTs in Section 2.3; and the third perspective allows us to consider post-experimental randomization. In Sections 2.2.5 and 2.3.3, we have summarized many useful practical techniques. Another common technique not mentioned above is sample splitting. A tangential example in Section 2.4 is the method MCRTs which divides the units into subsets in estimating lagged treatment effects from stepped-wedge randomized trials. Experiments in Section 2.5 demonstrate the merits of MCRTs on simulated and real data. All the techniques mentioned above are useful when the most obvious randomization tests are not computable, not nested, or not independent, thereby greatly expanding the scope of randomization inference in complex experiments.

2.8 Technical Proofs

2.8.1 Proof of Lemma 2

Proof. By Definition 1, $\boldsymbol{W}_{\boldsymbol{Z}} = (\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{Z},\boldsymbol{z}^*)}(\boldsymbol{z}^*) : \boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{Z}})$ denote all the potential outcomes imputable from $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{Z})$ under H. Then $\boldsymbol{W}_{\boldsymbol{Z}}$ is fully determined by $\boldsymbol{Z}$ and $\boldsymbol{Y}$. By assumption, $T_{\boldsymbol{Z}}(\boldsymbol{z}^*, \boldsymbol{W})$ only depends on $\boldsymbol{W}$ through $\boldsymbol{W}_{\boldsymbol{Z}}$ for all $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{Z}}$. Thus, $T_{\boldsymbol{Z}}(\boldsymbol{z}^*, \boldsymbol{W})$ is a function of $\boldsymbol{z}^*$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$. With an abuse of notation we denote it as $T_{\boldsymbol{Z}}(\boldsymbol{z}^*, \boldsymbol{Y})$. By the definition of the p-value in (2.1) and $\boldsymbol{Z}^* \perp\!\!\!\perp \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$,

$$\begin{aligned} P(\boldsymbol{Z}, \boldsymbol{W}) &= \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{W}) \le T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}, \boldsymbol{W}\} \\ &= \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{Y}) \le T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{Y}) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}, \boldsymbol{W}\} \\ &= \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*, \boldsymbol{Y}) \le T_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{Y}) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}\} \end{aligned}$$

is a function of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ (since $\mathbb{P}^*$ only depends on the known density $\pi(\cdot)$). □

2.8.2 Proof of Lemma 7

Lemma 7. *Let T be a random variable and $F(t) = \mathbb{P}(T \leq t)$ be its distribution function. Then $F(T)$ has a distribution that stochastically dominates the uniform distribution on $[0, 1]$.*

Proof. Let $F^{-1}(\alpha) = \sup\{t \in \mathbb{R} \mid F(t) \leq \alpha\}$. We claim that $\mathbb{P}\{F(T) \leq \alpha\} = \mathbb{P}\{T < F^{-1}(\alpha)\}$; this can be verified by considering whether T has a positive mass at $F^{-1}(\alpha)$ (equivalently, by considering whether $F(t)$ jumps at $t = F^{-1}(\alpha)$). By using the fact that $F(t)$ is non-decreasing and right-continuous, we have

$$\mathbb{P}\left\{F(T) \leq \alpha\right\} = \mathbb{P}\left\{T < F^{-1}(\alpha)\right\} = \lim_{t \uparrow F^{-1}(\alpha)} F(t) \leq \alpha.$$

□

2.8.3 Proof of Lemma 3

Proof. Consider $\mathcal{S} \in \mathcal{R}^{(j)}$ and $\mathcal{S}' \in \mathcal{R}^{(k)}$ for some $j, k \in [K]$ and $\mathcal{S} \cap \mathcal{S}' \neq \emptyset$. By the definition of refinement and (2.10), $\mathcal{S} \cap \mathcal{S}' \in \underline{R}^{\{j,k\}} \subseteq \mathcal{R}^{(j)} \cup \mathcal{R}^{(k)}$, so there exists integers m such that $\mathcal{S} \cap \mathcal{S}' = \mathcal{S}_m^{(j)}$ or $\mathcal{S}_m^{(k)}$. Because $\mathcal{R}^{(j)}$ and $\mathcal{R}^{(k)}$ are partitions, this means that $\mathcal{S} = \mathcal{S}_m^{(j)}$ or $\mathcal{S}' = \mathcal{S}_m^{(k)}$. In either case, $\mathcal{S} \cap \mathcal{S}' = \mathcal{S}$ or $\mathcal{S}'$.

Now consider any $\boldsymbol{z} \in \mathcal{Z}$ and $j, k \in [K]$. Because $\mathcal{S}_{\boldsymbol{z}}^{(j)}, \mathcal{S}_{\boldsymbol{z}}^{(k)} \in \mathcal{R}^{[K]}$ and $\mathcal{S}_{\boldsymbol{z}}^{(j)} \cap \mathcal{S}_{\boldsymbol{z}}^{(k)} \neq \emptyset$ (because they contain at least $\boldsymbol{z}$), either $\mathcal{S}_{\boldsymbol{z}}^{(j)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(k)}$or $\mathcal{S}_{\boldsymbol{z}}^{(k)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(j)}$ must be true. This means that we can order the conditioning events $\mathcal{S}_{\boldsymbol{z}}^{(k)}, k \in [K]$ according to the relation $\supseteq$. Without loss of generality, we assume that, at $\boldsymbol{z}$,

$$\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(2)} \supseteq \cdots \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K-1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K)}.$$

Then $\bigcap_{k=1}^K \mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{S}_{\boldsymbol{z}}^{(K)}$ and $\bigcup_{k=1}^K \mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{S}_{\boldsymbol{z}}^{(1)}$. Thus the intersection and the union of $\{\mathcal{S}_{\boldsymbol{z}}^{(k)}\}_{k=1}^K$ are contained in $\mathcal{R}^{[K]}$ which collects all $\mathcal{S}_{\boldsymbol{z}}^{(k)}$ by the definition. As this is true for all $\boldsymbol{z} \in \mathcal{Z}$, we have $\underline{\mathcal{R}}^{[K]} \subseteq \mathcal{R}^{[K]}$ and $\overline{\mathcal{R}}^{[K]} \subseteq \mathcal{R}^{[K]}$. □

2.8.4 Proof of Lemma 4

Proof. (i) Suppose $\mathcal{S}'$, $\mathcal{S}''$ are two distinct nodes in $\mathrm{ch}(\mathcal{S})$. By Lemma 3, they are either disjoint or nested. If they are nested, without loss of generality, suppose $\mathcal{S}'' \supset \mathcal{S}'$. However, this contradicts with the definition of the edge $\mathcal{S} \to \mathcal{S}'$, as by Definition 7 there should be no $\mathcal{S}''$ satisfying $\mathcal{S} \supset \mathcal{S}'' \supset \mathcal{S}'$. This shows that any two nodes in $\mathrm{ch}(\mathcal{S})$ are disjoint.

Next we show that the union of the sets in $\mathrm{ch}(\mathcal{S})$ is $\mathcal{S}$. Suppose there exists $\boldsymbol{z} \in \mathcal{S}$ such that $\boldsymbol{z} \notin \mathcal{S}'$ for all $\mathcal{S}' \in \mathrm{ch}(\mathcal{S})$. In consequence, $\boldsymbol{z} \notin \mathcal{S}'$ for all $\mathcal{S}' \in \mathrm{de}(\mathcal{S})$. Similar to the proof of Lemma 3, we can order $\mathcal{S}_{\boldsymbol{z}}^{(k)}, k \in [K]$ according to set inclusion. Without loss of generality, suppose

$$\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(2)} \supseteq \cdots \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K-1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K)}.$$

This shows that $\mathcal{S} = \mathcal{S}_{\boldsymbol{z}}^{(K)}$, so $\mathrm{ch}(\mathcal{S}) = \mathrm{de}(\mathcal{S}) = \varnothing$. This contradicts the assumption.

(ii) Consider any $\mathcal{S} \in \mathcal{R}^{[K]}$, $\mathcal{S}' \in \mathrm{an}(\mathcal{S})$ and $\mathcal{S}'' \in \mathrm{de}(\mathcal{S})$. By definition, $\mathcal{S}' \supset \mathcal{S} \supset \mathcal{S}''$. Because the sets in any partition $\mathcal{R}^{(k)}$ are disjoint, this shows that no pairs of $\mathcal{S}, \mathcal{S}', \mathcal{S}''$ can belong to the same partition $\mathcal{R}^{(k)}$. Thus, $\mathcal{K}(\mathcal{S}'), \mathcal{K}(\mathcal{S}), \mathcal{K}(\mathcal{S}'')$ are disjoint. Because this is true for any $\mathcal{S}' \in \mathrm{an}(\mathcal{S})$ and $\mathcal{S}'' \in \mathrm{de}(\mathcal{S})$, this shows $\mathcal{K}(\mathrm{an}(\mathcal{S}))$, $\mathcal{K}(\mathcal{S})$, and $\mathcal{K}(\mathrm{de}(\mathcal{S}))$ are disjoint.

Now consider any $\boldsymbol{z} \in \mathcal{S}$. By the proof of (i), $\mathcal{S}$ is in a nested sequence consisting of $\mathcal{S}_{\boldsymbol{z}}^{(1)}, \ldots, \mathcal{S}_{\boldsymbol{z}}^{(K)}$. Hence, $\mathcal{K}(\mathrm{an}(\mathcal{S})) \cup \mathcal{K}(\mathcal{S}) \cup \mathcal{K}(\mathrm{de}(\mathcal{S})) = \mathcal{K}(\mathrm{an}(\mathcal{S}) \cup \{\mathcal{S}\} \cup \mathrm{de}(\mathcal{S})) = [K]$.

(iii) The first result follows immediately from the fact that $\mathrm{an}(\mathcal{S}') = \mathrm{an}(\mathcal{S}) \cup \{\mathcal{S}\}$. The second result is true because both $\mathcal{K}(\{\mathcal{S}'\} \cup \mathrm{de}(\mathcal{S}'))$ and $\mathcal{K}(\mathrm{de}(\mathcal{S}))$ are equal to $[K] \setminus \mathcal{K}(\mathrm{an}(\mathcal{S}) \cup \{\mathcal{S}\})$. □

2.8.5 Proof of Lemma 5

Proof. First, we claim that for any $j, k \in [K]$ and $\boldsymbol{z} \in \mathcal{Z}$, given that $\boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}^{(j)} \cap \mathcal{S}_{\boldsymbol{z}}^{(k)}$, $P^{(k)}(\boldsymbol{Z}, \boldsymbol{W})$ only depends on $\boldsymbol{Z}$ through $T_{\boldsymbol{z}}^{(k)}(\boldsymbol{Z}, \boldsymbol{W})$. This is true because of the definition of the p-value (2.1) and Lemma 1 (so $\mathcal{S}_{\boldsymbol{Z}}^{(k)} = \mathcal{S}_{\boldsymbol{z}}^{(k)}$ and $T_{\boldsymbol{Z}}(\cdot, \cdot) = T_{\boldsymbol{z}}(\cdot, \cdot)$). By using this claim, the conditional independence (2.11) implies that

$$P^{(j)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \mathcal{S}_{\boldsymbol{z}}^{(j)} \cap \mathcal{S}_{\boldsymbol{z}}^{(k)}, \boldsymbol{W}, \quad \forall j, k \in [K], \boldsymbol{z} \in \mathcal{Z}.$$

Now fix $\mathcal{S} \in \mathcal{R}^{[K]}$ and consider any $j \in \mathcal{K}(\mathcal{S})$ and $k \in \mathcal{K}(\mathrm{an}(\mathcal{S}) \cup \{\mathcal{S}\}) \setminus \{j\}$. By Lemma 3, $\mathcal{S}_{\boldsymbol{z}}^{(j)} \cap \mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{S}_{\boldsymbol{z}}^{(j)} = \mathcal{S}$ for any $\boldsymbol{z} \in \mathcal{S}$. Then the claimed independence follows immediately. □

2.8.6 Proof of Lemma 6

Proof. Let $\psi^{(k)} = \psi^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) = 1_{\{P^{(k)}(\boldsymbol{Z},\boldsymbol{W}) \le \alpha^{(k)}\}}$ for $k \in [K]$, so the left-hand side of (2.14) can be written as

$$\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha^{(1)}, \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha^{(K)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} = \mathbb{E}\left\{\prod_{k=1}^{K} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\}.$$

We prove Lemma 6 by induction. First, consider any leaf node in the Hasse diagram, that is, any $\mathcal{S} \in \mathcal{R}^{[K]}$ such that $\mathrm{ch}(\mathcal{S}) = \varnothing$. By Lemma 4, $\mathcal{K}(\mathcal{S}) \cup \mathcal{K}(\mathrm{an}(\mathcal{S})) = [K]$. Then by Lemma 5, $\psi^{(j)} \perp\!\!\!\perp \psi^{(k)}$ for any $j \in \mathcal{K}(\mathcal{S})$ and $k \neq j$. By using the validity of each CRT (Theorem 1), we obtain

$$\begin{aligned}\mathbb{E}\left\{\prod_{k=1}^{K} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} &= \mathbb{E}\left\{\prod_{k \in \mathcal{K}(\mathrm{an}(\mathcal{S}))} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\mathcal{S})} \mathbb{E}\left\{\psi^{(j)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \\ &\le \mathbb{E}\left\{\prod_{k \in \mathcal{K}(\mathrm{an}(\mathcal{S}))} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\mathcal{S})} \alpha^{(j)}.\end{aligned}$$

This is exactly (2.14) for a leaf node. Now consider a non-leaf node $\mathcal{S} \in \mathcal{R}^{[K]}$ (so $\mathrm{ch}(\mathcal{S}) \neq \emptyset$) and suppose (2.14) holds for any descendant of $\mathcal{S}$. We have

$$\begin{aligned}&\mathbb{E}\left\{\prod_{k=1}^{K} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \\ &= \mathbb{E}\left\{\sum_{\mathcal{S}' \in \mathrm{ch}(\mathcal{S})} 1_{\{\boldsymbol{Z} \in \mathcal{S}'\}} \mathbb{E}\left\{\prod_{k=1}^{K} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}', \boldsymbol{W}\right\} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} && \text{(By Lemma 4(i))} \\ &\le \mathbb{E}\left\{\sum_{\mathcal{S}' \in \mathrm{ch}(\mathcal{S})} 1_{\{\boldsymbol{Z} \in \mathcal{S}'\}} \mathbb{E}\left\{\prod_{k \in \mathcal{K}(\mathrm{an}(\mathcal{S}'))} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}', \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\{\mathcal{S}'\} \cup \mathrm{de}(\mathcal{S}'))} \alpha^{(j)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \\ &&& \text{(By the induction hypothesis)} \\ &= \mathbb{E}\left\{\prod_{k \in \mathcal{K}(\mathrm{an}(\mathcal{S}) \cup \{\mathcal{S}\})} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\mathrm{de}(\mathcal{S}))} \alpha^{(j)} && \text{(By Lemma 4(iii))} \\ &= \mathbb{E}\left\{\prod_{k \in \mathcal{K}(\mathrm{an}(\mathcal{S}))} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\mathcal{S})} \mathbb{E}\left\{\psi^{(j)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\mathrm{de}(\mathcal{S}))} \alpha^{(j)} \\ &&& \text{(By Lemma 4(ii) and Lemma 5)} \\ &\le \mathbb{E}\left\{\prod_{k \in \mathcal{K}(\mathrm{an}(\mathcal{S}))} \psi^{(k)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\{\mathcal{S}\} \cup \mathrm{de}(\mathcal{S}))} \alpha^{(j)}. && \text{(By Theorem 1)}\end{aligned}$$

By induction, this shows that (2.14) holds for all $\mathcal{S} \in \mathcal{R}^{[K]}$. □

2.8.7 Proof of Theorem 2

Proof. Lemma 3 shows that $\overline{\mathcal{R}}^{[K]} \subseteq \mathcal{R}^{[K]}$, thus (2.14) holds for every $\mathcal{S} \in \overline{\mathcal{R}}^{[K]} \subseteq \mathcal{R}^{[K]}$. Moreover, any $\mathcal{S} \in \overline{\mathcal{R}}^{[K]}$ has no superset in $\mathcal{R}^{[K]}$ and thus has no ancestors in the Hasse diagram. This means that the right-hand side of (2.14) is simply $\prod_{k=1}^{K} \alpha^{(k)}$. Because $\overline{\mathcal{G}}^{[K]}$ is the σ-algebra generated by the events $\{\boldsymbol{Z} \in \mathcal{S}\}$ for $\mathcal{S} \in \overline{\mathcal{R}}^{[K]}$, equation (2.12) holds. Finally, equation (2.13) holds trivially by taking the expectation of (2.12) over $\overline{\mathcal{G}}^{[K]}$. □

2.8.8 Proof of Proposition 8

We denote the units used in the k-th test by $\mathcal{I}^{(k)} = \mathcal{I}_0^{(k)} \cup \mathcal{I}_1^{(k)}$. We denote the treatment variables and outcomes used in the k-th test by $\boldsymbol{A}^{(k)} = (A_i : i \in \mathcal{I}^{(k)})$ and $\boldsymbol{Y}^{(k)} = (Y_i^{(k)} : i \in \mathcal{I}^{(k)})$, respectively. We define the difference-in-means statistics in the k-th test as

$$T(\boldsymbol{A}^{(k)}, \boldsymbol{Y}^{(k)}) = \sqrt{N}(\bar{Y}_1^{(k)} - \bar{Y}_0^{(k)}), \tag{2.31}$$

where the average outcomes are given by

$$\bar{Y}_1^{(k)} = (N_1^{(k)})^{-1} \sum_{i \in \mathcal{I}^{(k)}} A_i^{(k)} Y_{ik} \quad \text{and} \quad \bar{Y}_0^{(k)} = (N_0^{(k)})^{-1} \sum_{i \in \mathcal{I}^{(k)}} (1 - A_i^{(k)}) Y_{ik}.$$

Let $\boldsymbol{W}^{(k)} = (Y_i^{(k)}(0), Y_i^{(k)}(1) : i \in \mathcal{I}^{(k)})$. The randomization distribution in the k-th test is

$$\hat{G}^{(k)}(b^{(k)}) = \mathbb{P}^* \left\{ T(\boldsymbol{A}_*^{(k)}, \boldsymbol{Y}^{(k)}(\boldsymbol{A}_*^{(k)})) \leq b^{(k)} \mid \boldsymbol{W}^{(k)} \right\}$$

where $\boldsymbol{A}_*^{(k)}$ is a permutation of $\boldsymbol{A}^{(k)}$, i.e., $\boldsymbol{A}_*^{(k)}$ is drawn from the same uniform assignment distribution as $\boldsymbol{A}^{(k)}$. Under assumption (i), using the bivariate Central Limit Theorem,

$$\left(T(\boldsymbol{A}^{(k)}, \boldsymbol{Y}^{(k)}), T(\boldsymbol{A}_*^{(k)}, \boldsymbol{Y}^{(k)}(\boldsymbol{A}_*^{(k)})) \right) \xrightarrow{d} (T_\infty^{(k)}, T_{\infty,*}^{(k)}),$$

where $T_\infty^{(k)}$ and $T_{\infty,*}^{(k)}$ are independent, each with a common c.d.f $G^{(k)}(\cdot)$. By Lehmann and Romano (2006, Theorem 15.2.3),

$$\hat{G}^{(k)}(b^{(k)}) \xrightarrow{P} G^{(k)}(b^{(k)})$$

for every $b^{(k)}$ which is a continuity point of $G^{(k)}(\cdot)$.

Under assumptions (i), (ii) and the zero-effect null hypothesis $H_0^{(k)}$, following Lehmann and Romano (2006, Theorem 15.2.5), we have

$$T_\infty^{(k)} \sim g_0^{(k)} = \mathcal{N}(0, V_\infty^{(k)}), \tag{2.32}$$

where the variance $V_\infty^{(k)}$ takes the form

$$V_\infty^{(k)} = 1/\Lambda^{(k)} = \frac{N}{N_0^{(k)}} \operatorname{Var}\left[Y^{(k)}(1)\right] + \frac{N}{N_1^{(k)}} \operatorname{Var}\left[Y^{(k)}(0)\right].$$

This expression of the asymptotic variance can be found in Imbens and Rubin (2015, Section 6.4). For completeness, an alternative derivation is provided in Section 2.8.8. Under assumptions (i), (ii) and the constant effect alternative hypothesis $H_1^{(k)}$ (with $\tau = h/\sqrt{N}$), we have

$$T_\infty^{(k)} \sim g_1^{(k)} = \mathcal{N}(h, V_\infty^{(k)}). \tag{2.33}$$

The c.d.f of $T_\infty^{(k)}$ under $H_0^{(k)}$ and $H_1^{(k)}$ are given by

$$G_0^{(k)}(b^{(k)}) = \Phi\left(b^{(k)}/\sqrt{V_\infty^{(k)}}\right) \quad \text{and} \quad G_1^{(k)}(b^{(k)}) = \Phi\left(\left[b^{(k)} - h\right]/\sqrt{V_\infty^{(k)}}\right).$$

Let $\tilde{b}^{(k)} = \left[G_0^{(k)}\right]^{-1}(p^{(k)})$. The p-value density function under $H_1^{(k)}$ can be rewritten as

$$f_1^{(k)}(p^{(k)}) = g_1^{(k)}(\tilde{b}^{(k)}) \left|\frac{d\tilde{b}^{(k)}}{dp^{(k)}}\right| = g_1^{(k)}(\tilde{b}^{(k)}) \left|\left([G_0^{(k)}]'(\tilde{b}^{(k)})\right)^{-1}\right| = g_1^{(k)}(\tilde{b}^{(k)})/g_0^{(k)}(\tilde{b}^{(k)}).$$

Since the p-values from the permutation tests follow a standard uniform distribution under the null, the log-likelihood ratio of the p-values $P^{(1)} \ldots, P^{(K)}$ is given by

$$\sum_{k=1}^{K} \log\left[f_1^{(k)}(p^{(k)})\right] = \sum_{k=1}^{K} \log\left[\frac{g_1^{(k)}(\tilde{b}^{(k)})}{g_0^{(k)}(\tilde{b}^{(k)})}\right] \propto \sum_{k=1}^{K} \sqrt{\Lambda^{(k)}}\Phi^{-1}(p^{(k)}).$$

Using the p-value weights $\Lambda^{(k)}, k \in [K]$, from the log-likelihood ratio, we obtain

$$T_\infty = \sum_{k=1}^{K} w_\infty^{(k)} \Phi^{-1}(P^{(k)}) \quad \text{where} \quad w_\infty^{(k)} = \sqrt{\frac{\Lambda^{(k)}}{\sum_{j=1}^{K} \Lambda^{(j)}}}. \tag{2.34}$$

Additional proof for Equation (2.32)

Lehmann and Romano (2006, Theorem 15.2.3) uses the fact that under assumptions (i) and (ii), $V_\infty^{(k)} = \operatorname{Var}[T(\boldsymbol{A}^{(k)}, \boldsymbol{Y}^{(k)})]$, but leaves the calculation of $\operatorname{Var}[T(\boldsymbol{A}^{(k)}, \boldsymbol{Y}^{(k)})]$ (their Equation 15.15) as an exercise for readers. Here we provide the calculation to complete

our proof for (2.32). To simplify the exposition, we let $m = N_1^{(k)}$, $n = N_0^{(k)}$. Suppose that $\mathcal{I}^{(k)} = [m+n]$, $\mathcal{I}_1^{(k)} = [m]$ and $\mathcal{I}_0^{(k)} = \{m+1, \ldots, m+n\}$, and that

$$\boldsymbol{Y}^{(k)} = (\underbrace{Y_1^{(k)}, \cdots, Y_m^{(k)}}_{\text{treated outcomes}}, \underbrace{Y_{m+1}^{(k)}, \ldots, Y_{m+n}^{(k)}}_{\text{control outcomes}}).$$

Let $\Pi = [\Pi(1), \ldots, \Pi(m+n)]$ be an independent random permutation of $1, \ldots, m+n$. We rewrite the difference-in-means statistics (2.31) as

$$T := T(\boldsymbol{A}^{(k)}, \boldsymbol{Y}^{(k)}) = \frac{\sqrt{N}}{m} \sum_{i=1}^{m+n} E_i Y_i^{(k)}, \text{ where } E_i = \begin{cases} 1 & \text{if } \Pi(i) \le m \\ -m/n & \text{otherwise} \end{cases}, \forall i \in [m+n].$$

Let D be the number of $i \le m$ such that $\Pi(i) \le m$. The value of T is determined by $m - D$, the number of units swapped between the treated group ($i = 1, \ldots m$) and control group ($i = m+1, \ldots, m+n$). Since all the treated (control) units have the same outcome variance, it does not matter which units are swapped in computing the variance of T.

We first consider the case that $m \le n$. Using the expectation of the hypergeometric distribution, we know

$$\mathbb{E}[D] = \sum_{d=0}^{m} \frac{\binom{m}{d}\binom{n}{m-d}}{\binom{m+n}{m}} d = \frac{m^2}{m+n}.$$

Let $\sigma_1^2 = \text{Var}[Y^{(k)}(1)]$ and $\sigma_0^2 = \text{Var}[Y^{(k)}(0)]$. The variance of T is given by

$$\begin{aligned}
\text{Var}(T) &= \frac{N}{m^2} \sum_{d=0}^{m} \frac{\binom{m}{d}\binom{n}{m-d}}{\binom{m+n}{m}} \left[d\sigma_1^2 + (m-d)\sigma_0^2 + (m-d)\frac{m^2}{n^2}\sigma_1^2 + (n-m+d)\frac{m^2}{n^2}\sigma_0^2 \right] \\
&= \frac{N}{m^2} \left[\frac{m^2}{m+n}\sigma_1^2 + \frac{mn}{m+n}\sigma_0^2 + \frac{mn}{m+n}\frac{m^2}{n^2}\sigma_1^2 + \frac{m^2}{m+n}\sigma_0^2 \right] \\
&= \frac{N}{m^2} \left[\frac{m^2}{n}\sigma_1^2 + m\sigma_0^2 \right] \\
&= \frac{N}{n}\sigma_1^2 + \frac{N}{m}\sigma_0^2.
\end{aligned}$$

If $m > n$, we let D denote the number of $i \in \{m+1, \ldots, m+n\}$ such that $\Pi(i) \in \{m+1, \ldots, m+n\}$. Then, $\mathbb{E}[D] = \frac{n^2}{m+n}$, and

$$\begin{aligned}
\text{Var}(T) &= \frac{N}{m^2} \sum_{d=0}^{n} \frac{\binom{m}{d}\binom{n}{m-d}}{\binom{m+n}{n}} \left[d\frac{m^2}{n^2}\sigma_0^2 + (n-d)\frac{m^2}{n^2}\sigma_1^2 + (n-d)\sigma_0^2 + (m-n+d)\sigma_1^2 \right] \\
&= \frac{N}{m^2} \left[\frac{m^2}{m+n}\sigma_0^2 + \frac{mn}{m+n}\frac{m^2}{n^2}\sigma_1^2 + \frac{mn}{m+n}\sigma_0^2 + \frac{m^2}{m+n}\sigma_1^2 \right] \\
&= \frac{N}{m^2} \left[\frac{m^2}{n}\sigma_1^2 + m\sigma_0^2 \right]
\end{aligned}$$

$$= \frac{N}{n}\sigma_1^2 + \frac{N}{m}\sigma_0^2.$$

The variance is the same for $m \leq n$ and $m > n$. This proves our claim and Equation 15.15 in Lehmann and Romano (2006): $\text{Var}(m^{-1/2}\sum_{i=1}^{m+n} E_i Y_i^{(k)}) = \text{Var}(\sqrt{\frac{m}{N}}T) = \frac{m}{n}\sigma_1^2 + \sigma_0^2$.

2.8.9 Proof of Proposition 9

In Algorithm 1, the time steps in $\mathcal{C}_j$ define a sequence of nested permutation tests. For example, $\mathcal{C}_1 = \{1, 3, 5, 7\}$ defines CRTs 1,3,5 and $\mathcal{C}_2 = \{2, 4, 6, 8\}$ defines CRTs 2,4,6 in Figure 2.2. Suppose that we only have one subset $\mathcal{C} = \{c_1, \ldots, c_K\}$ consists of K time points. We next prove the tests defined on $\mathcal{C}$ are valid to combine. It is suffices to show that the test statistics $\hat{T} = \sum_{k=1}^{K} \hat{w}^{(k)}\Phi^{-1}(P^{(k)})$ stochastically dominates the random variable

$$\tilde{T} := \sum_{k=1}^{K} \hat{w}^{(k)}\Phi^{-1}(U^{(k)}) \sim \mathcal{N}\left(0, \sum_{k=1}^{K} [\hat{w}^{(k)}]^2\right) = \mathcal{N}(0, 1).$$

where each $U^{(k)}$ is a standard uniform random variable. The second equality is achieved by the definition of $\hat{w}^{(k)}, k \in [K]$.

By conditioning on the potential outcomes $(\boldsymbol{W}^{(k)}, k \in [K])$, the weights $\hat{w}^{(k)}, k \in [K]$, are fixed. The derivation follows the same steps as Proposition 6 (the conditioning on $\boldsymbol{W}$'s is suppressed). By construction, $c_1 < \ldots < c_K$ and the conditioning sets $\mathcal{S}^{(1)} \supseteq \cdots \supseteq \mathcal{S}^{(K)}$ for all $\boldsymbol{z} \in \mathcal{Z}$. This implies that $\mathcal{G}^{(1)} \subseteq \cdots \subseteq \mathcal{G}^{(K)}$. Conditioning on $\mathcal{G}^{(k')}$, the term $\sum_{k=1}^{k'-1} \hat{w}^{(k)}\Phi^{-1}(P^{(k)})$ is fixed. By the law of iterated expectation,

$$\begin{aligned}
\mathbb{E}\left[1\left\{\hat{T} \leq b\right\}\right] &= \mathbb{E}\left[1\left\{\hat{w}^{(K)}\Phi^{-1}(P^{(K)}) \leq b - \sum_{k=1}^{K-1} \hat{w}^{(k)}\Phi^{-1}(P^{(k)})\right\}\right] \\
&= \mathbb{E}\left(\mathbb{E}\left[1\left\{\hat{w}^{(K)}\Phi^{-1}(P^{(K)}) \leq b - \sum_{k=1}^{K-1} \hat{w}^{(k)}\Phi^{-1}(P^{(k)})\right\} \;\middle|\; \mathcal{G}^{(K)}\right]\right) \\
&\leq \mathbb{E}_{U^{(K)}}\left(\mathbb{E}\left[1\left\{\hat{w}^{(K)}\Phi^{-1}(U^{(K)}) \leq b - \sum_{k=1}^{K-1} \hat{w}^{(k)}\Phi^{-1}(P^{(k)})\right\} \;\middle|\; U^{(K)}\right]\right) \\
&= \mathbb{E}_{U^{(K)}}\Bigg(\mathbb{E}\Bigg[1\Bigg\{\hat{w}^{(K-1)}\Phi^{-1}(P^{(K-1)}) \leq b - \sum_{k=1}^{K-2} \hat{w}^{(k)}\Phi^{-1}(P^{(k)}) \\
&\qquad\qquad - \hat{w}^{(K)}\Phi^{-1}(U^{(K)})\Bigg\} \;\Bigg|\; \mathcal{G}^{(K-1)}, U^{(K)}\Bigg]\Bigg) \\
&\leq \mathbb{E}_{U^{(K-1)}, U^{(K)}}\Bigg(\mathbb{E}\Bigg[1\Bigg\{\hat{w}^{(K-1)}\Phi^{-1}(U^{(K-1)}) \leq b - \sum_{k=1}^{K-2} \hat{w}^{(k)}\Phi^{-1}(P^{(k)})
\end{aligned}$$

$$- \hat{w}^{(K)}\Phi^{-1}(U^{(K)})\Big\} \;\Big|\; U^{(K-1)}, U^{(K)}\Big]\Big)$$

$$\vdots$$

$$\leq \mathbb{E}_{U^{(1)},\ldots,U^{(K)}}\left(\mathbb{E}\left[1\Big\{\sum_{k=1}^{K}\hat{w}^{(k)}\Phi^{-1}(U^{(k)}) \leq b\Big\} \;\Big|\; U^{(1)},\ldots,U^{(K)}\right]\right)$$

$$= \mathbb{E}\left[1\Big\{\tilde{T} \leq b\Big\}\right].$$

The inequalities are attained by the validity of the p-values $P^{(1)},\ldots,P^{(K)}$, i.e., each $P^{(k)}$ stochastically dominates the standard uniform variable $U^{(k)}$. Since $\hat{T}$ stochastically dominates the standard normal random variable $\tilde{T}$,

$$\mathbb{P}\{\hat{P} \leq \alpha\} = \mathbb{P}\{\Phi(\hat{T}) \leq \alpha\} \leq \mathbb{P}\{\Phi(\tilde{T}) \leq \alpha\} = \alpha \tag{2.35}$$

The last equality is achieved by the fact that $\Phi(\tilde{T})$ is a standard uniform random variable.

Suppose that Algorithm 1 creates multiple subsets $\mathcal{C}_j, j \in [J]$. Applying the same proof to the tests defined on each $\mathcal{C}_j$, we will have

$$\mathbb{E}\left[1\left\{\hat{T}_j \leq b\right\}\right] \leq \mathbb{E}\left[1\left\{\tilde{T}_j \leq b\right\}\right]$$

where $\hat{T}_j = \sum_{c\in\mathcal{C}_j}\hat{w}^{(c)}\Phi(P^{(c)})$ and $\tilde{T}_j = \sum_{c\in\mathcal{C}_j}\hat{w}^{(c)}\Phi(U^{(c)}) \sim \mathcal{N}\left(0, \sum_{c\in\mathcal{C}_j}\left[\hat{w}^{(c)}\right]^2\right)$. Because of sample splitting, we can combine the p-values from the tests based on different $\mathcal{C}_j$. The test statistics $\hat{T} = \sum_{j\in\mathcal{J}}\hat{T}_j$ stochastically dominates the random variable

$$\tilde{T} = \sum_{j\in\mathcal{J}}\tilde{T}_j \sim \mathcal{N}\Big(0, \sum_{j\in\mathcal{J}}\sum_{c\in\mathcal{C}_j}\left[\hat{w}^{(c)}\right]^2\Big) = \mathcal{N}(0,1),$$

which implies that (2.35) still holds for the combined p-value $\hat{P} = \Phi(\hat{T})$.

2.9 Confidence intervals from randomization tests

Here we describe how to invert (a combination of) permutation tests and the complexity involves; see also Ernst (2004, Section 3.4) for an introduction. Consider a completely randomized experiment with treated outcomes $(Y_1,\ldots,Y_m)$ and control outcomes $(Y_{m+1},\ldots,Y_{m+n})$. Consider the constant effect null hypothesis

$$H_0 : Y_i(1) = Y_i(0) + \Delta, \forall i \in [m+n],$$

We can implement a permutation test for H_0 by testing the zero-effect hypothesis on the shifted outcomes $\boldsymbol{Y}_\Delta = (Y_1 - \Delta, \ldots, Y_m - \Delta, Y_{m+1}, \ldots, Y_{m+n})$.

To construct a confidence interval for the true constant treatment effect τ, we can consider all values of Δ for which we do not reject H_0. We define the left and right tails of the randomization distribution of $T(\boldsymbol{Z}^*, \boldsymbol{Y}_\Delta)$ by

$$P_1(\Delta) = \mathbb{P}^*\{T(\boldsymbol{Z}^*, \boldsymbol{Y}_\Delta) \leq T(\boldsymbol{Z}, \boldsymbol{Y}_\Delta)\} \text{ and } P_2(\Delta) = \mathbb{P}^*\{T(\boldsymbol{Z}^*, \boldsymbol{Y}_\Delta) \geq T(\boldsymbol{Z}, \boldsymbol{Y}_\Delta)\},$$

where $\boldsymbol{Z} = [\mathbf{1}_m/m, -\mathbf{1}_n/n]$ is observed assignment, $\boldsymbol{Z}^*$ is a permutation of $\boldsymbol{Z}$, T is the test statistics, e.g., difference-in-means, $T(\boldsymbol{Z}, \boldsymbol{Y}_\Delta) = \boldsymbol{Z}^\top Y_\Delta$. The complexity of computing $P_1(\Delta)$ and $P_2(\Delta)$ with b different permutations is $O(mb + nb)$. The two-sided $(1-\alpha)$-confidence interval for Δ is given by

$$[\Delta_L, \Delta_U] := \left[\min_{P_2(\Delta) > \alpha/2} \Delta, \max_{P_1(\Delta) > \alpha/2} \Delta \right].$$

We use a grid search to approximately find the minimum and maximum Δ in $[\Delta_L, \Delta_U]$. Since $P_1(\Delta)$ and $P_2(\Delta)$ are monotone functions of Δ, we can also consider obtaining the optimal Δ's via a root-finding numerical method (see e.g. Garthwaite, 1996).

Inverting a combination of permutation tests can be done similarly. For example, by searching the same Δ's for every permutation test, the lower confidence bound Δ_L is given by the minimum Δ under which the p-value of the combined test (e.g. $\hat{P}$ in Proposition 9) is larger than $\alpha/2$. The complexity of inverting a combined test scales linearly in terms of the number of tests. The total complexity is manageable as long as the numbers of units and permutations are not very large at the same time. If we have covariates information in the dataset, we can consider fitting a linear regression model (with basis functions for nonlinearity). We can compute the matrix inversion in the least-squares solution once, then update the solution by shifting the outcome vector with different Δ's. Inverting the test returns an interval with coverage probability approximately equal to $1 - \alpha$. When the probability is slightly below $1 - \alpha$, we can decrease α by a small value (e.g. 0.0025) gradually and reconstruct the interval based on the previously searched Δ's. We stop if we obtain an interval with coverage probability above $1 - \alpha$.

Chapter 3

Identifiable representations for the estimation of conditional average treatment effects

Conditional average treatment effects (CATEs) allow us to understand the effect heterogeneity across a large population of individuals. However, typical CATE learners assume all confounding variables are measured in order for the CATE to be identifiable. This requirement can be satisfied by collecting many variables, at the expense of increased sample complexity for estimating CATEs. To combat this, we propose an energy-based model (EBM) that learns a low-dimensional representation of the variables by employing a noise contrastive loss function. With our EBM we introduce a preprocessing step that alleviates the dimensionality curse for *any existing* learner developed for estimating CATEs. We prove that our EBM keeps the representations partially identifiable up to some universal constants. These properties enable the representations to converge and keep the CATE estimates consistent. Experiments demonstrate the convergence as well as show that estimating CATEs on our representations performs better than on the variables or the representations obtained through other dimensionality reduction methods.

3.1 Introduction

Average treatment effect (ATE) is arguably the most popular estimand in the causal inference literature. With the ATE, one measures if a treatment is effective on average over a population of individuals. However, even if we estimate an ATE accurately, we can not conclude if a

treatment is beneficial for a particular individual. In order to get treatment effect estimates for one individual, we condition the ATE on the individual of interest, and arrive at the *conditional* average treatment effect (CATE). CATEs know successful applications in areas such as healthcare and education.

While clinical trials represent the gold standard for causal inference, they often have a small number of individuals and narrow inclusion criteria, rendering them unsuitable for use in estimating the causal effects conditional on some particular individual's confounding variables (covariates). On the other hand, observational datasets are becoming increasingly available, but require careful attention to the biases in the datasets. There is growing interest in leveraging observational data to estimate CATEs, e.g., electronic healthcare records used to determine which patients should get what treatments, or school records to optimize educational policy in low- and high-income communities.

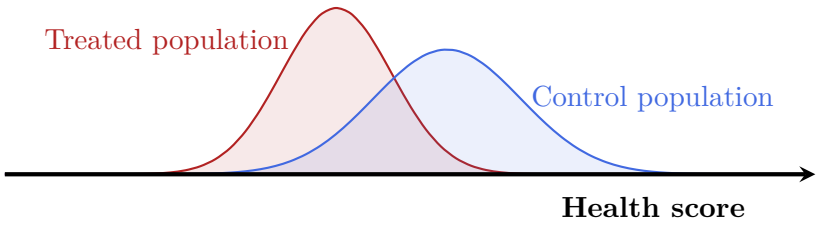


Figure 3.1 Imbalanced treated and control (i.e. untreated) populations. Individuals with lower health scores are more likely to receive the treatment.

A fundamental assumption for valid causal inference on observational data is called the strong ignorability assumption (Rosenbaum and Rubin, 1983, 1984), also known as the unconfoundedness assumption. It assumes independence between the potential outcomes of interest and the treatment variable, conditional on the confounding covariates. Because this assumption is untestable, we often estimate causal effects using all the observed covariates. However, estimating CATEs with moderate or high dimensional covariates is challenging. For example, in Section 3.1 we illustrate the treatment assignment process based on one observed covariate, health score. Here, the assignment process creates a discrepancy between the treated and control populations. That is to say, we rarely observe healthy individuals who receive the treatment, and unhealthy individuals who do not receive the treatment. Then the CATE (i.e., the treatment effect conditional on the health score) becomes difficult to estimate for these individuals. The main reason for this is due to work with finite samples: the probability of observing two comparable individuals in a dataset decreases as the covariates dimension increases. However, many covariates in the high-dimensional space are often generated by some common and low-dimensional (latent) variables.

Contributions. In this chapter, we propose a representation learning method based on a partially randomized energy-based model (EBM) to embed the covariates into a low-dimensional space before estimating CATEs. This preprocessing step can be used alongside any regression model and learner to reduce their dimensionality curse in CATE estimation. We prove that the representation in the partially randomized EBM is *partially identifiable up to some universal constants* for any value of the covariates. To our best knowledge, identifying representations in deep learning models exactly is still infeasible. Existing theory settles on achieving weaker versions of identifiability with the help of some auxiliary information, e.g., time steps and class labels. Auxiliary information does not exist in most observational datasets. We prove that by optimizing the partially randomized EBM with a noise contrastive loss function and a sample splitting strategy, the representations can converge consistently with increasing sample sizes. Experiments on multiple datasets complement our theoretical results. We empirically validate the convergence of the representations with increasing sample sizes. We also show that estimating CATEs based on our representations achieve better performance than directly on the covariates or the representations obtained via a variety of benchmark dimensionality reduction methods.

3.2 Setup

We use the potential outcome framework (Neyman, 1923; Rubin, 1974) to define causal effects. Consider an observational dataset $\mathcal{D} = \{O_i = (X_i, A_i, Y_i) : i \in [N]\}$, where $[N] = \{1, \ldots, N\}$. Each individual i is described by a set of covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, a binary treatment variable $A_i \in \mathcal{A} = \{0, 1\}$ and an observed outcome $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$. We assume that the samples in $\mathcal{D}$ are N i.i.d copies of the random variable

$$O = (X, A, Y) \sim \mathbb{P}(O) = \mathbb{P}(Y \mid A, X)\mathbb{P}(A \mid X)\mathbb{P}(X).$$

We assume every individual i has two potential outcomes, the control outcome $Y_i(0)$ and the treated outcome $Y_i(1)$. The treatment assignment depends on the individuals' covariates, i.e., $A_i \not\perp\!\!\!\perp X_i$. This dependence is quantified via the conditional distribution $e(X_i) = \mathbb{P}(A_i = 1|X_i)$, also termed as the propensity score in the literature. We make the standard assumptions on observational data. We divide $\mathcal{D}$ into a control set and a treated set, $\mathcal{D}_\mathrm{c} = \{(X_i, A_i, Y_i) : A_i = 0, i \in [N]\}$ and $\mathcal{D}_\mathrm{t} = \{(X_i, A_i, Y_i) : A_i = 1, i \in [N]\}$. We denote the sample sizes of $\mathcal{D}_\mathrm{c}$ and $\mathcal{D}_\mathrm{t}$ by $N_\mathrm{c} = |\mathcal{D}_\mathrm{c}|$ and $N_\mathrm{t} = |\mathcal{D}_\mathrm{t}|$. Under Assumption 7, $\mu_a(x) := \mathbb{E}\{Y(a) \mid X = x\} = \mathbb{E}\{Y \mid X = x, A = a\}$ for $a \in \{0, 1\}$. Then we can identify the conditional average treatment effect (CATE) $\tau(x)$ by

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x). \tag{3.1}$$

In nonparametric regression, the dimension and smoothness of the data generating function jointly determine the mean squared error of a regression model (Stone, 1980). The error of the used regression model determines the error of a CATE learner.

Definition 9 (Hölder ball)**.** The Hölder ball $\mathcal{H}_d(s)$ is the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ supported on $\mathcal{X} \subseteq \mathbb{R}^d$ with their partial derivatives satisfying that

$$\left| \frac{\partial^m f}{\partial^{m_1} \cdots \partial^{m_d}}(x) - \frac{\partial^m f}{\partial^{m_1} \cdots \partial^{m_d}}(x') \right| \lesssim \|x - x'\|_2^{s-\lfloor s \rfloor},$$

$\forall x, x' \in \mathcal{X}$ and $m = (m_1, \cdots, m_d)$ s.t. $\sum_{j=1}^d m_j = \lfloor s \rfloor$.

The notation $a \lesssim b$ denotes the relation $a \leq Cb$ for some universal constant C. Essentially, $\mathcal{H}_d(s)$ is the class of smooth functions that are close to their $\lfloor s \rfloor$-order Taylor approximations. We assume μ_0, μ_1, e and τ are s-smooth functions in the Hölder balls $\mathcal{H}_d(s)$ for some non-negative smoothness parameter $s = \alpha_0, \alpha_1, \beta, \gamma$, respectively.

The identification formula (3.1) motivates a common estimation strategy called a "T-learner", where "T" refers to "Two" regression models. A T-learner estimates μ_0 and μ_1 by fitting two separate regression models, $\hat{\mu}_0$ and $\hat{\mu}_1$, on $\mathcal{D}_\text{c}$ and $\mathcal{D}_\text{t}$, respectively. It estimates the CATE as the difference $\hat{\tau}(\cdot) = \hat{\mu}_1(\cdot) - \hat{\mu}_0(\cdot)$. Suppose the mean squared error of $\hat{\mu}_0$ and $\hat{\mu}_1$ are $N_\text{c}^{-\frac{2\alpha_0}{2\alpha_0+d}}$ and $N_\text{t}^{-\frac{2\alpha_1}{2\alpha_1+d}}$, respectively. A T-learner's mean squared error $\mathbb{E}\left\{[\hat{\tau}(X) - \tau(X)]^2\right\}$ is $O(N_\text{c}^{-\frac{2\alpha_0}{2\alpha_0+d}} + N_\text{t}^{-\frac{2\alpha_1}{2\alpha_1+d}})$. There are other advanced learners based on different identification formulas, e.g., X-learner (Künzel et al., 2019), R-learner (Nie and Wager, 2021) and DR-learner (Kennedy, 2020). For example, the identification formula of the DR-learner is based on the uncentered first-order influence function ϕ of the ATE,

$$\phi(O) = \frac{A}{e(X)}\left[Y - \mu_1(X)\right] + \mu_1(X) - \frac{1-A}{1-e(X)}\left[Y - \mu_0(X)\right] - \mu_0(X). \tag{3.2}$$

As discussed in Section 1.3.2, the DR-learner uses $\hat{\phi}(O)$ as the pseudo-outcome for estimating CATEs, where $\hat{\phi}(O)$ is generated by plugging the models $\hat{\mu}_0, \hat{\mu}_1$ and $\hat{e}$ into the expression (3.2). More specifically, splitting $\mathcal{D}$ into three subsets $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$, it estimates μ_0 and μ_1 using $\mathcal{D}_1$, estimates e using $\mathcal{D}_2$, then estimates the CATE τ using $\mathcal{D}_3$ (via regressing $\hat{\phi}(O)$ onto X). Kennedy (2020, Theorem 2) shows that it estimates τ with mean squared error

$$O\left(\tilde{N}_2^{-\frac{2\beta}{2\beta+d}}\left(\tilde{N}_{1,\text{c}}^{-\frac{2\alpha_0}{2\alpha_0+d}} + \tilde{N}_{1,\text{t}}^{-\frac{2\alpha_1}{2\alpha_1+d}}\right) + \tilde{N}_3^{-\frac{2\gamma}{2\gamma+d}}\right),$$

where $\tilde{N}_m = |\mathcal{D}_m|, m = 1, 2, 3$, $\tilde{N}_{1,\text{c}}$ and $\tilde{N}_{1,\text{t}}$ are the numbers of control and treated individuals in $\mathcal{D}_1$. The efficiency loss from sample splitting can be remedied by cross-fitting (Chernozhukov et al., 2018; Nie and Wager, 2021). DR-learner can improve CATE estimation

by leveraging the smoothness of τ. But for an accurate CATE estimator to exist in finite samples, the requirement on the smoothness parameters ($\alpha_0, \alpha_1, \beta$ and γ) is restrictive if the amount of dimensions d is large, regardless of which learner we use.

Observational studies often include covariates that represent the same aspect of an individual. For example, an individual's health status can be represented by some collection of covariates, e.g., blood pressure, temperature and some disease-specific symptoms. These covariates are correlated and contain overlapping information about the individual. This hints at a potentially more sample-efficient estimation strategy: first learning these aspects as a low-dimensional representation of the covariates, then fitting $\hat{\mu}_0, \hat{\mu}_1$ (and $\hat{e}$) on the low dimensional representation to estimate the CATE. The limitation of representation learning is that the representation itself is non-smooth and takes many samples to learn. In supervised learning on a fully labelled dataset, there is no obvious advantage of learning the representations first over learning the label directly. By contrast, observational datasets for CATE estimation are often imbalanced ($N_\mathrm{t} \ll N_\mathrm{c}$) so that the term $N_\mathrm{t}^{-\frac{2\alpha_1}{2\alpha_1+d}}$ in a T-learner's mean squared error is very large. Rather than directly using all the covariates to construct $\hat{\mu}_0, \hat{\mu}_1$ (and $\hat{e}$), we employ representation learning based on *all the samples*, i.e., $N_\mathrm{t} + N_\mathrm{c}$ samples from both the treated and control group. Then by using the low dimensional representation to estimate CATEs, the learners will potentially have smaller errors.

We consider the representation learning model, concatenated together with the outcome (propensity score) model, as an outcome (propensity score) model *based on the observed covariates*. The consistency of the resulting CATE estimator thus depends on the consistency of *both* models. This requires the representation to be identifiable, which is so far still impossible to achieve exactly for overparameterized neural networks. In our next section, we provide an approximate solution to this problem, sufficient for CATE estimation.

3.3 Partially identifiable energy-based models

A parameter is identifiable in a class of statistical models if every model describing the same distribution has the same value of the parameter. If models with different parameter-values give the same distribution, i.e., generate the same observed data in the large data limit, we can no longer find the true model from the data even if the sample size is large (Lewbel, 2019). Identifiability is often achieved by introducing some constraint on the model class, as is also the case here. We will construct partially identifiable representations in a class of partially randomized energy-based models (EBMs). By partially identifiable, we mean if two models give the same distribution, then their representations are only different by some universal constants. A partially randomized EBM is constructed as follows.

Suppose that we want to learn a k-dimensional representation of the covariates ($k < d$)[1]. We let $f_\theta : \mathcal{X} \to \mathbb{R}^k$ be a neural network that generates the k-dimensional data representation. To simplify the exposition in what follows, we assume f_θ is linear, i.e., $f_\theta(x) = \theta x$ for $\theta \in \Theta \subset \mathbb{R}^{k \times d}$. We define k standard EBMs (LeCun et al., 2006) on $\mathcal{X}$ with a *shared* f_θ:

$$p_{\theta,j}(x) = Z_{\theta,j}^{-1} \exp\left[-\beta_j^\top f_\theta(x)\right], \ \forall j \in [k], \tag{3.3}$$

where $Z_{\theta,j} = \int_{\mathcal{X}} \exp\left[-\beta_j^\top f_\theta(x)\right] dx$. We let $\mathcal{P} = \{p_{\theta,j} \mid \beta_j \in \mathbb{R}^k, \theta \in \Theta\}$ and $\mathcal{P}(\beta_j)$ denote the subset of $\mathcal{P}$ with a fixed β_j. We define *a partially randomized EBM* as a mixture of uniformly weighted EBMs $p_{\theta,j} \in \mathcal{P}(\beta_j)$, $j \in [K]$, with a fixed orthogonal $k \times k$ matrix $B = (\beta_1, \ldots, \beta_k)$[2] and a shared f_θ (θ is the only learnable parameter). By construction, the partially randomized EBM satisfies the partial identifiability defined as follows.

Theorem 3. *For any $k \times k$ orthogonal matrix $B = (\beta_1, \ldots, \beta_k)$ and $p_{\theta,j}, p_{\tilde{\theta},j} \in \mathcal{P}(\beta_j)$ such that $p_{\theta,j}(\cdot) = p_{\tilde{\theta},j}(\cdot), \forall j \in [k]$, we have*

$$f_\theta(\cdot) - f_{\tilde{\theta}}(\cdot) = C \quad \textit{for some constant vector } C. \tag{3.4}$$

Next, we will introduce a training strategy for our partially randomized EBM, which will enable the learnt representation model to converge to some limits that are only different by some constants C like f_θ and $f_{\tilde{\theta}}$ in (3.4). This will enable the follow-up CATE estimates to converge consistently because the regression models $\hat{\mu}_0, \hat{\mu}_1$ (and $\hat{e}$) are indifferent to conditioning on a random variable, or the same random variable plus some constant vector. Furthermore, by standardizing the learnt representations, we can fix their mean to 0 and their variance to 1 in any sample size. Given that the representations have mean 0 and variance 1, the representations obtained from different runs of the experiments will have a correlation close to 1 at each dimension in large samples, as will be demonstrated in Section 3.6.2.

3.4 Noise contrastive learning

Fitting energy-based models (EBMs) by maximum likelihood estimation (MLE) is often infeasible because the partition function ($Z_{\theta,j}$) is intractable. Noise Contrastive Estimation (NCE) proposed by Gutmann and Hyvärinen (2010, 2012) is a consistent and computationally efficient alternative. The high-level idea of NCE is to optimize an EBM by contrasting it

[1] The errors of CATE learners depend on the performance of the outcome and propensity score models $\hat{\mu}_0$, $\hat{\mu}_1$ and $\hat{e}$. Because some CATE learners do not use a propensity score model, the dimension k is tuned as a hyper-parameter via cross-validation on the observed outcomes in this chapter.

[2] B is generated by first generating a matrix $B_0 \in \mathbb{R}^{k \times k}$, where each entry is drawn independently from a standard normal distribution, and then taking B as the matrix of eigenvectors of B_0.

with another noise distribution with known and easy-to-sample density. Advanced methods have been proposed to tune the noise distribution, e.g., see Bose et al. (2018); Ceylan and Gutmann (2018); Gao et al. (2020).

Here for every individual $i \in [N]$, we draw b corrupted samples $\tilde{X}_{i1}, \ldots, \tilde{X}_{ib}$ from a noise distribution $p_{\tilde{X}|X}(\tilde{x} \mid X_i)$ defined as follows. Each $\tilde{X}_{ia}$ is generated in two steps: (1) we sample an independent binary variable with some probability for each feature of X_i, used to decide which features of X_i will be corrupted, then (2) corrupt each selected continuous feature by adding white noise drawn from a standard normal distribution, and corrupt each selected categorical feature by uniformly sampling a value from its range. A mathematical description of $p_{\tilde{X}|X}(\tilde{x} \mid X_i)$ is provided in Section 3.9. Overall, the original and corrupted data of individual i is given by

$$\bar{X}_i = (X_i, \tilde{X}_{i1}, \ldots, \tilde{X}_{ib}) \sim p_X(x) \prod_{a=1}^{b} p_{\tilde{X}_a|X}(\tilde{x}_a \mid x).$$

We split the N individuals into k subsets $\mathcal{I}_j, j \in [k]$, to train each of of k models $p_{\theta,j}(x)$ in the partially randomized EBM. Suppose we randomly permute the columns of $\bar{X}_i$ and let $V_i = (V_{ia} : a \in [b+1])$ be the permuted $\bar{X}_i$. Then each column of V_i has equal probability $(b+1)^{-1}$ for being the original sample X_i. We derive the predictive probability of $V_{ia} = X_i$ from the posterior distribution,

$$q_{\theta,j}(a \mid V_i) = \frac{(b+1)^{-1} p_{\theta,j}(V_{ia}) \tilde{p}_{-a}(V_i)}{\sum_{c=1}^{b+1} (b+1)^{-1} p_{\theta,j}(V_{ic}) \tilde{p}_{-c}(V_i)}, \tag{3.5}$$

where $\tilde{p}_{-a}(V_i) = \prod_{a \in [b+1]: a' \neq a} p_{\tilde{X}|X}(V_{ia'} \mid V_{ia})$. It is noteworthy that the intractable partition function $Z_{\theta,j}$ in $p_{\theta,j}$ (in (3.3)) cancels out in the expression of $q_{\theta,j}(a \mid V_i)$. Let $W_i \in \{0,1\}^{b+1}$ indicate which column of V_i is X_i. We can think of $\{(V_i, W_i) : i :\in \mathcal{I}_j\}$ as a set of labeled "images" and optimize the probability $q_{\theta,j}(a \mid V_i)$ to predict W_i. Let $N_j = |\mathcal{I}_j|$. Our objective function is the negative cross-entropy[3],

$$\mathcal{L}_n(\theta) = k^{-1} \sum_{j=1}^{k} \mathcal{L}_{N,j}(\theta),$$

where $\mathcal{L}_{N,j}(\theta)$ is given by

$$\mathcal{L}_{N,j}(\theta) = N_j^{-1} \sum_{i \in \mathcal{I}_j} \sum_{a=1}^{b+1} W_{ia} \log q_{\theta,j}(a \mid V_i) = N_j^{-1} \sum_{i \in \mathcal{I}_j} \log q_{\theta,j}(1 \mid \bar{X}_i). \tag{3.6}$$

[3]This is essentially the ranking objective in (Józefowicz et al., 2016; Ma and Collins, 2018) with a different noise distribution. We reformulate the training strategy as a more intuitive multiclass classification task.

The representation model f_θ is trained on all the samples, even though we split the samples across the models $p_{\theta,j}, j \in [k]$, in our partially randomized EBM. The training strategy here follows the same principle as the other representation learning methods (Vincent, 2011; Vincent et al., 2010): assume the covariates X_i live in some d^*-dimensional manifold ($d^* < d$). If $q_{\theta,j}(a \mid V_i)$ is predictive of W_i, i.e., can distinguish any true sample $X_i \sim p_X(x)$ from its noisy proxy $\tilde{X}_i \sim \tilde{p}_{\tilde{X}|X}(\tilde{x} \mid X_i)$, we have $p_{\theta,j}(x) \approx p_X(x)$ in (3.5). This implies that the low-dimensional representation given by $f_\theta(x)$ is informative of the true covariates X_i; the representation is also predictive of the outcome and treatment because they are generated by the covariates. More formally, Theorem 4 below shows that by our training strategy, $\hat{\theta}_N$ will converge to some θ_0 such that $p_{\theta_0,j}(x) = p_X(x)$ for any $x \in \mathcal{X}$. Then by (3.4) in Theorem 3, the limits of $f_{\hat{\theta}_N}(x)$ are only different by some universal constants.

Theorem 4. *Suppose that the covariate space $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, $f_\theta(x)$ has a compact parameter space $\Theta \subset \mathbb{R}^{k \times d}$, and $f_\theta(x)$ is continuous with respect to θ for any $x \in \mathcal{X}$. Assume that the density function $p_X(x) = p_{\theta_0,j}(x)$ for some $\theta_0 \in \Theta$. For any number of noise samples b and $\hat{\theta}_N \in \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$, $\lim_{N \to \infty} p_{\hat{\theta}_N,j}(x) = p_X(x)$ with probability 1.*

The theorem is proven by showing that $\mathcal{L}_{\infty,j}(\theta)$ is maximized by $q_{\theta,j}(a \mid V_i)$ with $p_{\theta,j}(x) = p_X(x)$, and verify the standard conditions for a consistent M-estimator $\hat{\theta}_N$ under a weaker identifiability assumption; see Section 3.8.2 for more details.

3.5 Related works

Identifiability theory. Khemakhem et al. (2020b) propose two definitions of identifiability for EBMs; weak and strong identifiability (in their Definitions 1 and 2). Their EBM is more complex than ours with β_j as a learnable parameter, while their objective is to identify both β_j and $f_\theta(x)$. This is unnecessary for CATE estimation. Arguably, the partial identifiability we define is stronger than both of their definitions. In their strong identifiability, under some assumptions, each dimension of f_θ is identifiable up to be multiplied by and plus some constants, and each dimension of f_θ can be permuted in any order. They also require a specific network architecture for f_θ. In our work, we use a *simpler* partially randomized EBM to achieve a *stricter* version of identifiability, without restricting the architecture of f_θ.

The works on nonlinear ICA and its generalization (Hyvarinen and Morioka, 2016; Hyvarinen et al., 2019; Khemakhem et al., 2020a; Mita et al., 2021) propose the idea of using contrastive learning for identifiable feature extraction when some auxiliary information (e.g., time steps) about the features is available. We use sample splitting and a noise contrastive loss function for training the partially randomized EBM, assuming *no auxiliary information*

is provided in the observational data. Monti et al. (2020) and Wu and Fukumizu (2020) propose non-linear ICA based methods for causal inference on structural causal models (Pearl, 2009). The setup and problems studied in their works are different from our method which is developed within the potential outcomes framework.

Representation learning. Representation learning is recently applied to balance or match the covariate distribution between the treated and control group in observational data, by minimizing the distributional distance between the group (Shalit et al., 2017), preserving local similarity (Yao et al., 2018), minimizing counterfactual variance (Zhang et al., 2020) and adversarial training (Kallus, 2020). We note that supervised dimensionality reduction in a deep learning model is not reliable because the model can easily overfit the limited outcome data without finding an informative representation of the covariates. Our proposed method works as a preprocessing step to reduce the dimensionality curse for any regression model, including these deep learning models which balance the distribution in their hidden layers.

In statistics, sufficient dimensionality reduction (SDR) (Adragni and Cook, 2009; Cook, 2009; Lee et al., 2013; Li, 1991) has been used in the models for estimating ATE and CATE (Cheng et al., 2020b; Ghosh et al., 2021; Huang and Yang, 2022; Luo et al., 2019; Ma et al., 2019). If the subspace spanned by the columns of a $d \times k$ matrix θ with $k \leq d$ satisfies that $Y \perp\!\!\!\perp X \mid \theta^\top X$, we call this subspace a SDR subspace. The idea of SDR is to project the covariates X onto this subspace before feeding it into a parametric or nonparametric regression model to estimate Y. To achieve the desired conditional independence, θ is jointly learnt with the regression model. This is not straightforward for some of the ML models. e.g., decision tree. Kallus et al. (2018) propose a matrix factorization based method for preprocessing noisy and missing covariates. In contrast with these methods, our method performs nonlinear dimensionality reduction of the covariates, which is more general for the data living in some low-dimensional manifold, including linear subspace. Nabi and Shpitser (2020) and Berrevoets et al. (2020) propose methods to deal with high-dimensional treatment variables, which is not the problem we consider.

Covariates selection. When there are irrelevant covariates in a dataset, data analysis should start with a covariates selection method, e.g., (De Luna et al., 2011; Greenewald et al., 2021; Shortreed and Ertefaie, 2017). However, covariates selection methods often have no guarantee to find the correct adjustment set for causal inference in finite samples. Furthermore, (selected) covariates are correlated, especially when we allow more covariates to be selected in order to satisfy the unconfoundedness assumption. Our representation learning method can be applied to further reduce the dimensionality of the correlated covariates and improve treatment effect estimation. In general, covariates selection and our method are applied in different stages and complement each other in the data analysis process.

3.6 Experiments

We make two claims in this chapter: (1) using our method as a preprocessing step increases the performance of CATE learners; (2) the representation in our model is partially identifiable so that the learnt representations and downstream CATE estimates are consistent. We next test these two claims. Throughout our experiments, we use four different CATE learners: X-Learner, DR-learner, T-Learner, and R-learner (Microsoft Research, 2019). We provide more details of our experiments (on learners and hyperparameters) in Section 3.10.

3.6.1 CATE estimation

Our main contribution is a way to increase performance for *any* learner. We evaluate learners' performance using *precision of estimating heterogeneous effects* (PEHE) introduced by Hill (2011) and now standard in CATE estimation. PEHE is essentially the expected risk $\mathbb{E}\left\{[\hat{\tau}(X) - \tau(X)]^2\right\}$ we define in Section 3.2. Because any individual's treated and control outcomes are never observed jointly, CATEs are unobserved in any real-world data. The literature thus relies on (semi-)synthetic data to evaluate CATE learners.

In our synthetic setup, the generating process of the observed variables $O = (X, A, Y)$ starts by sampling a latent variable $U \sim \mathcal{N}(0, I_{5\times5})$. Then we generate a set of covariates $X = \mathcal{N}(g(U), I_{d\times d})$, two potential outcomes, $\mu_0(U)$ and $\mu_1(U)$ and a treatment assignment $A \sim \text{Ber}[e(U)]$. The observed outcome is given by $Y = \mathcal{N}(A\mu_0(U) + (1-A)\mu_1(U), 1)$. The CATE is given by $\tau(U) = \mu_1(U) - \mu_0(U)$. The function g is a deep ReLU network; μ_0 and μ_1 are one-layer neural networks, with an exp-function on their output layers; e is a one-layer network with a sigmoid-function on its output layer. By generating i.i.d samples from this process, we create a training set (with size N specified in Table 3.1) and a large testing set with 20k samples. Given a training set, we first use it to optimize our partially randomized EBM. Then we preprocess it and apply CATE learners on these lower-dimensional representations. As a comparison, we also apply the same CATE learners on the original covariates.

Lower PEHE across CATE learners. Table 3.1 shows that our method greatly benefits a broad spectrum of CATE learners on the synthetic dataset and semi-synthetic dataset Twins (Almond et al., 2005) with real covariates, especially in small sample sizes. While the gain of using our method diminishes somewhat in larger sample sizes, it is still significant. More importantly, we observe that with our EBM, the performance gaps between different learners shrink significantly. Specifically, R-learner with EBM has the best performance on average over the table while it performs poorly in small samples without

Table 3.1 Results on synthetic data and semi-synthetic data (Twins). Each row reports the average PEHE (lower is better) over ten runs for each CATE learner (standard deviation in script size): both *with* representations (indicated as "✓"), and *without* representation (indicated as "✗"). For each run, we learn a new representation. In the above two blocks, we vary sample sizes and dimensions using our synthetic setup, and in the bottom block, we vary the sample size for the Twins-dataset. The best result is indicated in bold. In green, we emphasize the best results per row, each time *with* EBM.

Methods		**X-Learner**		**DR-Learner**		**T-Learner**		**R-Learner**	
	EBM	✗	✓	✗	✓	✗	✓	✗	✓
d	N	*Synth. data with increasing sample size and increasing dimensions*							
50	100	2.309 ±.00	**1.994** ±.02	4.594 ±.56	**2.017** ±.04	2.441 ±.00	**1.993** ±.01	3.194 ±.26	**1.982** ±.04
100	250	2.779 ±.00	**2.018** ±.01	4.056 ±.32	**2.154** ±.39	2.838 ±.00	**2.019** ±.01	3.702 ±.23	**2.018** ±.01
150	500	2.618 ±.00	**2.000** ±.01	3.030 ±.12	**2.001** ±.01	2.641 ±.00	**2.000** ±.01	2.877 ±.08	**2.000** ±.01
200	1k	2.185 ±.00	**1.940** ±.01	2.283 ±.02	**1.941** ±.01	2.189 ±.00	**1.939** ±.01	2.271 ±.01	**1.940** ±.01
250	1.5k	2.267 ±.00	**1.949** ±.02	2.427 ±.01	**1.976** ±.00	2.271 ±.00	**1.948** ±.01	2.436 ±.02	**1.949** ±.02
	N	*Synth. data with increasing sample size and dimensions fixed at* $d = 100$							
	100	2.134 ±.00	**1.927** ±.01	24.61 ±9.9	**2.096** ±.09	2.279 ±.00	**1.929** ±.01	3.192 ±.13	**1.925** ±.01
	250	2.779 ±.00	**2.018** ±.01	4.056 ±.32	**2.154** ±.39	2.838 ±.00	**2.019** ±.01	3.702 ±.23	**2.018** ±.01
	500	2.155 ±.00	**2.056** ±.02	2.334 ±.07	**2.273** ±.67	2.166 ±.00	**2.053** ±.02	2.271 ±.05	**2.056** ±.02
	1k	2.059 ±.00	**1.964** ±.02	2.105 ±.01	**2.016** ±.16	2.061 ±.00	**1.964** ±.02	2.086 ±.01	**1.965** ±.02
	1.5k	2.013 ±.00	**1.998** ±.02	2.043 ±.01	**1.998** ±.02	2.014 ±.00	**1.998** ±.02	2.024 ±.01	**1.991** ±.02
	N	*Twins* ($d = 48$) *with increasing sample size*							
	500	0.214 ±.00	**0.144** ±.00	0.236 ±.04	**0.182** ±.05	0.221 ±.00	**0.145** ±.00	0.222 ±.02	**0.145** ±.00
	1k	0.294 ±.00	**0.162** ±.00	0.348 ±.12	**0.173** ±.03	0.301 ±.00	**0.162** ±.01	0.532 ±.11	**0.161** ±.00
	1.5k	0.165 ±.00	**0.154** ±.00	0.189 ±.06	**0.159** ±.01	0.165 ±.00	**0.154** ±.00	0.172 ±.01	**0.154** ±.00
	2k	0.167 ±.00	**0.156** ±.00	0.197 ±.03	**0.159** ±.00	0.167 ±.00	**0.156** ±.00	0.222 ±.05	**0.157** ±.00
	2.5k	0.297 ±.00	**0.153** ±.00	0.390 ±.19	**0.156** ±.00	0.297 ±.00	**0.153** ±.00	0.358 ±.22	**0.153** ±.00

EBM. Overall, our experimental results align with our theoretical discussion in Section 3.2: by reducing the dimensionality d to a smaller number, the learners will have lower errors, i.e., lower PEHEs and smaller performance gaps on the testing sets.

Lower PEHE than benchmark dimensionality reduction methods. Based on our previous experiment, a logical next question to ask is whether other dimensionality reduction methods may also help. We compare our EBM method to various linear and nonlinear dimensionality reduction methods in prepossessing the real covariates of the Twins dataset. Specifically, we compare against Principal Components Analysis (PCA), Feature Agglomeration (FA), Spectral Embedding (SE), Isomap, KernelPCA with an RBF kernel, and an Autoencoder (AE). Table 3.2 shows that our EBM method outperforms all the benchmarks significantly over different sample sizes.

Table 3.2 Results using different dimensionality reduction methods. Using an R-learner, we report the PEHE of our EBM and other benchmark methods over 10 runs (standard deviation in script size): PCA, Feature Agglomeration (FA), Spectral Embedding (SE), Isomap, and KernelPCA (K-PCA) and Autoencoder (AE).

Methods	**PCA**	**FA**	**SE**	**Isomap**	**K-PCA**	**AE**	**EBM**
N	*Twins ($d = 48$) with increasing sample size*						
500	1.092 ±.11	1.758 ±1.1	1.011 ±.00	1.006 ±.00	1.015 ±.00	0.580 ±.03	**0.145** ±.00
1k	1.015 ±.00	0.963 ±.00	1.010 ±.00	1.004 ±.00	1.010 ±.00	0.549 ±.04	**0.161** ±.00
1.5k	1.014 ±.00	0.965 ±.00	1.005 ±.00	1.006 ±.00	1.012 ±.00	0.546 ±.04	**0.154** ±.00
2k	1.013 ±.00	0.957 ±.00	1.009 ±.00	1.007 ±.00	1.013 ±.00	0.579 ±.03	**0.157** ±.00
2.5k	1.007 ±.00	0.951 ±.00	1.002 ±.00	1.006 ±.00	1.006 ±.00	0.542 ±.04	**0.153** ±.00

To further validate our proposed method, we repeat the same experiment using **additional regression models and data**, and report consistent results to those we present in this section, in Section 3.10. Overall, we do not find **sample splitting** increase the variance of our method across all our experiments. As we explained below eq. (3.6), the representation model f_θ is trained with all the samples in our objective function.

3.6.2 Partial identifiability of representations

In this section, we empirically validate that our method produces identifiable representations. Having an identifiable method is important for later inspection of the representations, but also to produce consistent CATE learners. Both of which are important in practice.

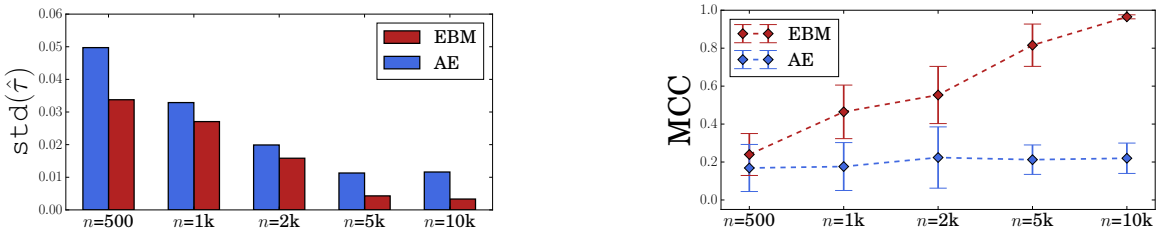


Figure 3.2 Results on identifiability. *Above*— For each model (an autoencoder (AE), and our model (EBM)) we learn ten distinct representations. We then fit an R-Learner on each representation, and calculate the standard deviation of their CATE estimates. Our method has lower standard errors compared to AE. *Below*— We report the mean correlation coefficient (MCC) between the representations on the Twins data (higher is better). Our EBM becomes more consistent with larger samples (error bars indicate standard deviation on MCC), and even tends to 1 in large samples.

Converging CATE estimates. The first panel in Figure 3.2 reports the standard deviation of the CATE-estimates, by an R-learner when fitted on the representations of an autoencoder (AE) and our method (EBM). The representations have the same amount of dimensions ($k = 5$). Figure 3.2 shows that our model decreases the standard deviation with increasing sample size—this is important, as applications require estimates to be consistent.

Converging representations. As discussed at the end of Section 3.4, the learnt representations after standardization should correlate as the sample size increases. We train our EBM ten times using distinct random initializations, while keeping the orthogonal matrix fixed across runs. We subsequently compute the mean correlation coefficient (MCC) between the representations of the test-set from different runs. The MCC is computed by averaging the correlation between *each dimension* in the representations of 20k samples from the test set. Note that the latter is a strict definition as it requires the representation to be consistent for each *individual* dimension.[4] Reported in the second panel, we see that our EBM's MCC grows as the sample size increases, leaving the (unidentifiable) AE behind, indicating that our EBM is identifiable, further confirming our theory.

3.7 Conclusions

We propose a partially randomized EBM to learn a partially identifiable low-dimensional representation of moderate or high-dimensional covariates in CATE estimation. We show theoretically and empirically that by training our EBM with a noise contrastive loss function and a sample splitting strategy, our representations converge to a set of limits differing only by some constants. Experiments on multiple datasets with various dimensions and sample sizes verify our theories and demonstrate a significant performance increase when using our method as a preprocessing step for CATE estimation.

Our work opens a few new directions for future research. First, our method currently operates within the standard setup of observational data in causal inference, while our partial identifiability theory does not rely on the network architecture of f_θ. Extending our work to other high-dimensional settings such as time series or vision could prove useful for many real world applications. Second, as an interpretable approach within CATE estimation, matching is concerned with finding similar individuals across treated and control groups. While effective, matching becomes harder in high-dimensions. Extending our approach to remain interpretable (e.g. by measuring each covariate's influence to each dimension in the representation) can arm matching approaches against the dimensionality curse.

[4] Previous work (Khemakhem et al., 2020b) tests identifiability using the MCC maximized by canonical-correlation analysis (CCA). Here we compute the exact correlation to test our stronger version of identifiability.

3.8 Technical Proofs

3.8.1 Proof of Theorem 3

Proof. Consider two different parameter values θ and $\tilde{\theta}$ such that $p_{\theta,j}(x) = p_{\tilde{\theta},j}(x)$. Using the expression (3.3) and applying logarithm to both sides,

$$\beta_j^\top f_\theta(x) = \beta_j^\top f_{\tilde{\theta}}(x) + \log \frac{Z_{\tilde{\theta},j}}{Z_{\theta,j}}. \tag{3.7}$$

By concatenating the last equation for all $j \in [k]$, we have

$$B^\top f_\theta(x) = B^\top f_{\tilde{\theta}}(x) + G, \tag{3.8}$$

where $G = (\log \frac{Z_{\tilde{\theta},j}}{Z_{\theta,j}} : j \in [k])$ is a k-dimensional vector. By definition, $BB^\top = I_{k\times k}$. Then multiplying the two side of (3.8) by B proves (3.4),

$$f_\theta(x) = f_{\tilde{\theta}}(x) + C \text{ for any } x \in \mathcal{X} \text{ and } C = BG. \tag{3.9}$$

Reversely, multiplying two sides of (3.9) by $\beta_j^\top$, we obtain (3.7),

$$\beta_j^\top f_\theta(x) = \beta_j^\top f_{\tilde{\theta}}(x) + \beta_j^\top BG = \beta_j^\top f_{\tilde{\theta}}(x) + G_j = \beta_j^\top f_{\tilde{\theta}}(x) + \log \frac{Z_{\tilde{\theta},j}}{Z_{\theta,j}}.$$

Then multiplying -1 and applying $\exp(\cdot)$ to both sides,

$$p_{\theta,j}(x) = Z_{\theta,j}^{-1} \exp\left[-\beta_j^\top f_\theta(x)\right] = Z_{\tilde{\theta},j}^{-1} \exp\left[-\beta_j^\top f_{\tilde{\theta}}(x)\right] = p_{\tilde{\theta},j}(x).$$

□

3.8.2 Proof of Theorem 4

Proof. We first show $q_{\theta,j}(a \mid V_i)$ with $p_{\theta,j}(x) = p_X(x)$ is a maximizer of $\mathcal{L}_{\infty,j}(\theta)$, then we show the standard conditions of consistent M-estimators hold for $\mathcal{L}_{N,j}(\theta)$. We assume the sample splitting always keep the individuals in the same folds as $N \to \infty$. This can be achieved by keeping the existing individuals in the same folds and randomly assigning a new individual to a fold as $N \to \infty$. For large N, each fold will have roughly the same number of individuals, and $N_j = |\mathcal{I}_j| \to \infty$ for every $j \in [k]$ as $N \to \infty$.

Step 1. Recall that for every $i \in [N]$, $\bar{X}_i = (X_i, \tilde{X}_{i1}, \ldots, \tilde{X}_{ib}) \sim p_X(x) \prod_{a=1}^b \tilde{p}_{\tilde{X}|X}(\tilde{x} \mid x)$. We randomly permute the columns of $\bar{X}_i$ and let $V_i = (V_{i,1}, \cdots V_{i,b+1})$ be the permuted $\bar{X}_i$. Each column of V_i has equal probability $(b+1)^{-1}$ for being the clean sample X_i. The variable

$W_i \in \{0,1\}^{b+1}$ indicate which column of V_i is X_i. Here, we define a categorical variable $S_i \in [b+1]$ such that $S_i = a$ if $W_{ia} = 1$. We know that

$$p_{S_i}(a) = 1/(b+1), \ \forall a \in [b+1].$$

We define the marginal distribution of $v = (v_c : c \in [b+1])$ as

$$\Lambda_j(v) = \sum_{a=1}^{b+1} p_S(a) p_{V|S}(v|a) = \sum_{a=1}^{b+1} (b+1)^{-1} p_X(v_a) \tilde{p}_{-a}(v),$$

where $\tilde{p}_{-a}(v) = \prod_{a \in [b+1]: a' \neq a} p_{\tilde{X}|X}(v_{a'} \mid v_a)$. We define the posterior probability of $S = a$ as

$$p_{S|V}(a \mid v) = \frac{(b+1)^{-1} p_X(v_a) \tilde{p}_{-a}(v)}{\sum_{c=1}^{b+1} (b+1)^{-1} p_X(v_c) \tilde{p}_{-c}(v)}.$$

This corresponds to the posterior probability $q_{\theta,j}(a \mid v)$ based on the model $p_{\theta,j}(x)$ in (3.5).

As $N \to \infty$, i.e., $N_j \to \infty$, the objective function in (3.6) is given by

$$\mathcal{L}_{\infty,j}(\theta) = \int \Lambda_j(v) \left[\sum_{a=1}^{b+1} p_{S|V}(a \mid v) \log q_{\theta,j}(a \mid v) \right] dv.$$

Because $\Lambda_j(v) > 0$ and Lemma 8, $\mathcal{L}_{\infty,j}(\theta)$ is maximized when

$$q_{\theta,j}(a \mid v) = p_{S|V}(a \mid v), \ \forall a \in [b+1]. \tag{3.10}$$

Suppose $v = (v_{a'} : a' \in [b+1])$ satisfies that $v_{a'} = \xi$ for all $a' \in [b+1] \setminus \{a\}$. Then

$$\tilde{p}_{-c}(v) = \tilde{p}_{\tilde{X}|X}(v_a \mid \xi) \left[p_{\tilde{X}|X}(\xi \mid \xi) \right]^{b-1}, \forall c \in [b+1] \setminus \{a\}$$

and

$$\tilde{p}_{-c}(v) = \tilde{p}_{-c'}(v), \forall c, c' \in [b+1] \setminus \{a\}. \tag{3.11}$$

We continue to rewrite (3.10) as

$$\begin{aligned}
\frac{p_{\theta,j}(v_a)\tilde{p}_{-a}(v)}{\sum_{c=1}^{b+1} p_{\theta,j}(v_c)\tilde{p}_{-c}(v)} &= \frac{p_X(v_a)\tilde{p}_{-a}(v)}{\sum_{c'=1}^{b+1} p_X(v_{c'})\tilde{p}_{-c'}(v)} \\
\frac{p_{\theta,j}(v_a)}{\sum_{c=1}^{b+1} p_{\theta,j}(v_c)\tilde{p}_{-c}(v)} &= \frac{p_X(v_a)}{\sum_{c'=1}^{b+1} p_X(v_{c'})\tilde{p}_{-c'}(v)} \\
\frac{\sum_{c=1}^{b+1} p_{\theta,j}(v_c)\tilde{p}_{-c}(v)}{p_{\theta,j}(v_a)} &= \frac{\sum_{c'=1}^{b+1} p_X(v_{c'})\tilde{p}_{-c'}(v)}{p_X(v_a)} \\
\tilde{p}_{-a}(v_a) + \sum_{c \neq a} \frac{p_{\theta,j}(v_c)\tilde{p}_{-a}(v_c)}{p_{\theta,j}(v_a)} &= \tilde{p}_{-a}(v_a) + \sum_{c' \neq a} \frac{p_X(v_{c'})\tilde{p}_{-a}(v_{c'})}{p_X(v_a)}
\end{aligned}$$

$$\frac{p_{\theta,j}(\xi)}{p_{\theta,j}(v_a)} = \frac{p_X(\xi)}{p_X(v_a)}$$
$$\beta_j^\top \left[f_\theta(\xi) - f_\theta(v_a)\right] = \beta_j^\top \left[f_{\theta_0}(\xi) - f_{\theta_0}(v_a)\right],$$

The fifth line is attained by (3.11). The last line is achieved by the assumption in the theorem. Combing the last equation for all $j \in [k]$ and using the fact that B orthogonal,

$$B^\top \left[f_\theta(\xi) - f_\theta(v_a)\right] = B^\top \left[f_{\theta_0}(\xi) - f_{\theta_0}(v_a)\right]$$
$$f_\theta(\xi) - f_\theta(v_a) = f_{\theta_0}(\xi) - f_{\theta_0}(v_a)$$
$$f_\theta(v_a) = f_{\theta_0}(v_a) + C(\theta, \theta_0).$$

Then,

$$p_{\theta,j}(x) = Z_{\theta,j}^{-1} \exp\left[-\beta_j^\top f_\theta(x)\right] = \frac{\exp\left[-\beta_j^\top f_{\theta_0}(x) - \beta_j^\top C(\theta,\theta_0)\right]}{\int_{\mathcal{X}} \exp\left[-\beta_j^\top f_{\theta_0}(x) - \beta_j^\top C(\theta,\theta_0)\right] dx} = p_X(x).$$

For any $\hat{\theta} \in \arg\max_{\theta\in\Theta} \mathcal{L}_{\infty,j}(\theta)$, we have $p_{\hat{\theta},j}(x) = p_X(x)$ for any $x \in \mathcal{X}$.

Step 2. Because $\mathcal{X}$ and Θ are compact, the value of each covariate and network parameter is bounded. Then the function we optimize in (3.6), $g_j(\bar{x};\theta) = \log q_{\theta,j}(1 \mid \bar{x})$, is bounded for any $\bar{x} = (x, \tilde{x}_{i1}, \ldots, \tilde{x}_{ib})$. Then we denote $\mathcal{L}_{N,j}(\theta)$ in (3.6) by $\mathbb{E}_N\left[g_j(\bar{X};\theta)\right] = N_j^{-1} \sum_{i\in\mathcal{I}_j} g_j(\bar{X}_i;\theta)$. Using the uniform law of large number (ULLN) (Jennrich, 1969, Theorem 2) and (Newey and McFadden, 1994, Lemma 2.4), we have

$$\sup_{\theta\in\Theta} \left|\mathbb{E}_N\left[g_j(\bar{X};\theta)\right] - \mathbb{E}\left[g_j(\bar{X};\theta)\right]\right| \xrightarrow{p} 0.$$

Now we change the exact identifiability assumption in (Wooldridge, 2010, Theorem 12.2) and (Newey and McFadden, 1994, Theorem 2.5) to our partial identifiability assumption. Suppose that there is a subset $\Theta_0 \subset \Theta$ such that for every $\theta_0 \in \Theta_0$, $p_{\theta_0,j}(x)$ gives the same distribution as $p_X(x)$, and any θ_0 is a non-unique maximizer of $\mathbb{E}\left[g_j(\bar{X};\theta)\right]$.

We define an open ball with radius equal to $\eta > 0$ for every $\theta_0 \in \Theta_0$. The region inside and outside these open balls is given by

$$\Theta_\eta = \left\{\theta \in \Theta \;\middle|\; \arg\min_{\theta_0\in\Theta_0} \|\theta - \theta_0\|_2 < \eta\right\} \quad \text{and} \quad \Theta_\eta^c = \left\{\theta \in \Theta \;\middle|\; \arg\min_{\theta_0\in\Theta_0} \|\theta - \theta_0\|_2 \geq \eta\right\}.$$

Using the proof of (Newey and McFadden, 1994, Theorem 2.1), $\forall \epsilon > 0$, $\theta_0 \in \Theta_0$ and $\hat{\theta}_N \in \arg\max \mathbb{E}_N \left[g_j(\bar{X}; \theta) \right]$, we have with probability approaching to 1:

$$\mathbb{E}\left[g_j(\bar{X}; \hat{\theta}_N) \right] > \mathbb{E}\left[g_j(\bar{X}; \theta_0) \right] - \epsilon. \tag{3.12}$$

By the compactness of Θ_η^c and the assumption that f_θ is continuous w.r.t to θ, we have

$$\sup_{\theta \in \Theta_\eta^c} \mathbb{E}\left[g_j(\bar{X}; \theta) \right] = \mathbb{E}\left[g_j(\bar{X}; \theta^*) \right] < \mathbb{E}\left[g_j(\bar{X}; \theta_0) \right] \text{ for some } \theta^* \in \Theta_\eta^c.$$

Thus, by $\epsilon = \mathbb{E}\left[g_j(\bar{X}; \theta_0) \right] - \sup_{\theta \in \Theta_\eta^c} \mathbb{E}\left[g_j(\bar{X}; \theta) \right]$, it follows from (3.12) that with probability approaching to 1,

$$\mathbb{E}\left[g_j(\bar{X}; \hat{\theta}_N) \right] > \sup_{\theta \in \Theta_\eta^c} \mathbb{E}\left[g_j(\bar{X}; \theta) \right] \Rightarrow \hat{\theta}_N \in \Theta_\eta. \tag{3.13}$$

Since (3.13) is true for any $\eta > 0$, we have $\hat{\theta}_N \in \Theta_0$ with probability 1 as $N \to \infty$. We note that the same proof holds if we consider the summation of $\mathbb{E}_N \left[g_j(\bar{X}; \theta) \right]$ over all $j \in [k]$. □

Lemma 8. *Suppose* $w = (w_1, \ldots, w_b) > 0$ *and* $\sum_{a=1}^b w_a = 1$,

$$f(\tilde{w}; w) = \sum_{a=1}^{b} w_a \log \tilde{w}_a \quad \textit{subject to} \quad \tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_b) > 0 \textit{ and } \sum_{a=1}^{b} \tilde{w}_a = 1.$$

Then $f(\tilde{w}; w)$ is maximized at $\tilde{w} = w$.

Proof. Suppose $g(\tilde{w}; w) = \sum_{a=1}^b w_a \log \tilde{w}_a + \lambda \left(\sum_{a=1}^b \tilde{w}_a - 1 \right)$. We have

$$\frac{\partial g(\tilde{w}; w)}{\partial \tilde{w}_c} = 0 \Rightarrow \tilde{w}_c = -\frac{w_c}{\lambda} \quad \text{and} \quad \frac{\partial g(\tilde{w}; w)}{\partial \lambda} = 0 \Rightarrow \sum_{a=1}^{b} \tilde{w}_a = 1.$$

Combining both conditions, we have $\sum_{a=1}^b \tilde{w}_a = -\frac{1}{\lambda} \sum_{a=1}^b w_c = -\frac{1}{\lambda} = 1 \Rightarrow \lambda = -1$. Then, $\tilde{w}_c = -\frac{w_c}{-1} \Rightarrow \tilde{w}_c = w_c$. Then by a second-derivative test on the bordered Hessian of $g(\tilde{w}; w)$, we have $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_b) = (w_1, \ldots, w_b)$ is a maximizer of the function $f(\tilde{w}; w)$. □

3.9 Noise sampler

We draw a noise sample from $p_{\tilde{X}|X}(\tilde{x} \mid X_i)$. For each feature X_{is} of the d-dimensional X_i, $s \in [d]$, we sample an independent binary variable $R_{is} \sim \text{Ber}(q)$ to decide if the s-th feature X_{is} will be corrupted. If $R_{is} = 1$, we will corrupt the s-th feature, otherwise not. Overall,

the first part of $p_{\tilde{X}|X}(\tilde{x} \mid X_i)$ is given by

$$\prod_{s=1}^{d} q^{R_{is}}(1-q)^{1-R_{is}},$$

where q is the only hyperparameter in $p_{\tilde{X}|X}(\tilde{x} \mid X_i)$. We use the same q for all $j \in [k]$. In Section 3.10.2, we describe how q (called perturbation prob.) is selected by validation.

Suppose that $R_{is} = 1$. If the s-th feature is continuous, we will corrupt it by adding a white noise E_{is} drawn from a standard normal distribution, i.e.,

$$E_{is} \sim p_{E_s|R_s}(e_s \mid 1) = \mathcal{N}(0, 1).$$

If the s-th feature is categorical and takes its value in $\mathcal{X}_s$, we will corrupt it by replacing X_{is} with a uniform sample E_{is} drawn from the same range $\mathcal{X}_s$, i.e.,

$$E_{is} \sim p_{E_s|R_s}(e_s \mid 1) = 1/|\mathcal{X}_s|.$$

Suppose that $R_{is} = 0$. We do not corrupt the s-th feature. That is,

- $E_{is} = 0$ and $p_{E_s|R_s}(0 \mid 0) = 1$ if the s-th feature is continuous and $R_{is} = 0$;
- $E_{is} = X_{is}$ and $p_{E_s|R_s,X_s}(X_{is} \mid 0, X_{is}) = 1$ if the s-th feature is categorical and $R_{is} = 0$.

Overall, the probability $p_{\tilde{X}|X}(\tilde{X}_i \mid X_i)$ is computed using R_{is}, E_{is} and X_{is} for all $s \in [d]$,

$$p_{\tilde{X}|X}(\tilde{X}_i \mid X_i) = \prod_{s=1}^{d} q^{R_{is}}(1-q)^{1-R_{is}} p_{E_s|R_s,X_s}(E_{is} \mid R_{is}, X_{is}).$$

where $p_{E_s|R_s,X_s}(E_{is} \mid R_{is}, X_{is})$ only depends on X_{is} if the s-th feature is categorical and $R_{is} = 0$, otherwise $p_{E_s|R_s,X_s}(E_{is} \mid R_{is}, X_{is}) = p_{E_s|R_s}(E_{is} \mid R_{is})$. The corrupted sample $\tilde{X}_i$ is obtained by either adding E_{is} to X_{is} or replacing X_{is} by E_{is} for every $s \in [d]$. We do not need to consider this step when we compute the probability $\tilde{p}_{\tilde{X}|X}(\tilde{X}_i \mid X_i)$.

3.10 Additional experiments & hyperparameters

In Section 3.10.1 we repeat our experiments in Table 3.1 using different regression models to estimate the outcomes and propensity score in the same CATE learners, and using an additional real-world dataset. For hyperparameter settings we refer to Section 3.10.2.

Table 3.3 Results on (semi-)synthetic data with `PowerTransform Regression`. Results are averaged over ten runs with and without the same representations in Table 3.1.

Methods		**X-Learner**		**DR-Learner**		**T-Learner**		**R-Learner**	
	EBM	✗	✓	✗	✓	✗	✓	✗	✓
d	N	*Synth. data with increasing sample size and increasing dimensions*							
50	100	2.267 ±.00	**2.010** ±.03	5.593 ±2.2	**2.015** ±.05	2.455 ±.00	**2.011** ±.03	53.17 ±5.2	**11.16** ±3.9
100	250	2.754 ±.00	**2.019** ±.01	3.963 ±.24	**2.027** ±1.2	2.798 ±.00	**2.020** ±.01	60.10 ±4.2	**10.70** ±3.9
150	500	2.575 ±.00	**2.002** ±.01	2.986 ±.08	**2.001** ±.01	2.595 ±.00	**2.001** ±.01	49.78 ±4.2	**12.31** ±1.8
200	1k	2.197 ±.00	**1.952** ±.00	2.293 ±.03	**1.941** ±.01	2.202 ±.00	**1.951** ±.01	42.76 ±3.9	**2.838** ±.64
250	1.5k	2.288 ±.00	**1.966** ±.04	2.410 ±.04	**1.979** ±.03	2.295 ±.00	**1.968** ±.04	42.03 ±3.6	**2.535** ±.15
	N	*Synth. data with increasing sample size and dimensions fixed at $d = 100$*							
	100	2.150 ±.00	**1.964** ±.02	32.63 ±13	**2.147** ±.15	2.289 ±.00	**1.973** ±.03	58.80 ±5.1	**5.713** ±2.2
	250	2.754 ±.00	**2.019** ±.01	3.963 ±.24	**2.027** ±1.2	2.798 ±.00	**2.020** ±.01	60.10 ±4.2	**10.70** ±3.9
	500	2.150 ±.00	**2.029** ±.02	2.319 ±.06	**2.006** ±.05	2.160 ±.00	**2.028** ±.03	41.85 ±3.4	**3.884** ±.79
	1k	2.053 ±.00	**1.986** ±.02	2.102 ±.01	**1.989** ±.01	2.057 ±.00	**1.987** ±.02	39.49 ±3.2	**2.637** ±.20
	1.5k	2.008 ±.00	**1.999** ±.02	2.354 ±.01	**1.999** ±.52	2.008 ±.00	**2.000** ±.02	37.17 ±2.9	**3.949** ±.77
	N	*Twins ($d = 48$) with increasing sample size*							
	500	0.203 ±.00	**0.187** ±.06	4.383 ±.22	**0.185** ±.02	**0.204** ±.00	0.248 ±.24	300.0 ±16.	**2.176** ±.89
	1k	0.177 ±.00	**0.169** ±.03	0.194 ±.01	**0.163** ±.01	0.177 ±.00	**0.159** ±.02	31.39 ±2.5	**1.029** ±1.1
	1.5k	0.169 ±.00	**0.154** ±.00	**0.172** ±.01	0.183 ±.08	0.169 ±.00	**0.155** ±.00	31.23 ±2.3	**0.459** ±.15
	2k	0.167 ±.00	**0.161** ±.00	0.168 ±.00	**0.163** ±.00	0.168 ±.00	**0.161** ±.00	29.95 ±2.4	**0.629** ±.30
	2.5k	0.169 ±.00	**0.162** ±.00	0.170 ±.00	**0.163** ±.00	0.169 ±.00	**0.162** ±.00	29.79 ±2.4	**0.439** ±.19
	N	*IHDP ($d = 25$) with increasing sample size*							
	100	1.814 ±.01	**1.502** ±.03	3.755 ±.72	**1.637** ±.12	1.845 ±.00	**1.507** ±.05	35.77 ±6.3	**21.36** ±17.
	250	1.713 ±.00	**1.598** ±.04	1.837 ±.09	**1.653** ±.13	1.727 ±.00	**1.593** ±.03	11.71 ±1.5	**9.552** ±4.0
	500	1.603 ±.00	**1.554** ±.02	1.672 ±.05	**1.571** ±.03	1.627 ±.00	**1.556** ±.02	23.45 ±2.5	**15.19** ±3.9

3.10.1 CATE learners with different regression models and datasets

Consider Tables 3.3-3.4-3.5, where we report the PEHE given the same experimental setup as we have in Table 3.1; for additional data (Infant Health Development Program (IHDP) (MacDorman and Atkinson, 1999)), and three additional regression models (`PowerTransform Regression` (Yeo and Johnson, 2000), `Polynomial Regression`, and `Ridge Regression`, respectively). From our results we learn that our EBM is agnostic to the choice of regression model, and is versatile enough to also perform well given other data. These results are promising and should give some assurance regarding our method before application in practice.

As we have in Table 3.1, we ran each CATE learner on ten distinct representations, given different folds of the data, and averaged the results. Note that we have not specifically optimised the EBM's hyperparameters for these different regression models, but rather kept them as they were in Table 3.1 (actual hyperparameter values are reported in Table 3.6). Note that these results are in line with those reported earlier using different regression models.

Table 3.4 Results on (semi-)synthetic data using `Polynomial Regression`. Results are averaged over ten runs with and without the same representations in Table 3.1.

Methods		**X-Learner**		**DR-Learner**		**T-Learner**		**R-Learner**	
	EBM	✗	✓	✗	✓	✗	✓	✗	✓
d	N	*Synth. data with increasing sample size and increasing dimensions*							
50	100	**2.095** ±.00	2.089 ±.09	75.12 ±26.	**2.500** ±.44	2.124 ±.00	**2.095** ±.09	110.8 ±7.1	**26.61** ±7.6
100	250	2.109 ±.00	**2.026** ±.03	11.67 ±1.9	**2.138** ±.30	2.168 ±.00	**2.028** ±.03	50.45 ±4.4	**10.72** ±4.0
150	500	2.048 ±.00	**2.008** ±.02	8.668 ±1.5	**2.006** ±.01	2.142 ±.00	**2.008** ±.02	44.29 ±3.4	**12.35** ±1.8
200	1k	1.964 ±.00	**1.949** ±.02	7.278 ±1.3	**1.949** ±.02	2.088 ±.00	**1.949** ±.02	42.04 ±3.4	**2.973** ±.57
250	1.5k	1.945 ±.00	**1.945** ±.03	6.764 ±1.2	**1.979** ±.03	2.109 ±.00	**1.945** ±.04	41.71 ±3.5	**2.555** ±.18
	N	*Synth. data with increasing sample size and dimensions fixed at $d = 100$*							
	100	2.009 ±.00	**1.987** ±.09	157.1 ±50.	**2.176** ±.13	2.038 ±.00	**1.992** ±.09	54.81 ±4.3	**6.612** ±3.5
	250	2.109 ±.00	**2.019** ±.01	11.67 ±1.9	**2.028** ±.02	2.168 ±.00	**2.019** ±.01	50.46 ±4.4	**9.623** ±3.0
	500	**1.897** ±.00	2.055 ±.05	8.020 ±1.7	**2.003** ±.00	**2.007** ±.00	2.051 ±.05	43.56 ±3.9	**3.758** ±.82
	1k	2.210 ±.00	**1.995** ±.02	5.871 ±.90	**2.289** ±.88	2.338 ±1.4	**1.996** ±.02	40.38 ±3.1	**2.965** ±.69
	1.5k	2.341 ±.00	**2.002** ±.02	4.637 ±.75	**2.156** ±.47	2.474 ±.00	**2.000** ±.02	38.31 ±2.9	**3.945** ±.76
	N	*Twins ($d = 48$) with increasing sample size*							
	500	0.345 ±.00	**0.155** ±.00	4.538 ±1.4	**0.158** ±.00	0.377 ±.00	**0.155** ±.00	116.0 ±14.	**0.847** ±.19
	1k	0.486 ±.00	**0.149** ±.00	1.747 ±.33	**0.157** ±.01	0.529 ±.00	**0.149** ±.00	78.22 ±7.3	**0.482** ±.11
	1.5k	0.455 ±.00	**0.153** ±.00	1.453 ±.40	**0.186** ±.06	0.481 ±.00	**0.153** ±.00	142.3 ±21.	**0.395** ±.13
	2k	0.426 ±.00	**0.159** ±.00	1.109 ±.32	**0.162** ±.00	0.451 ±.00	**0.159** ±.00	38.13 ±4.2	**0.655** ±.33
	2.5k	0.403 ±.00	**0.159** ±.00	0.921 ±.14	**0.163** ±.01	0.418 ±.00	**0.159** ±.00	33.39 ±2.8	**0.459** ±.19
	N	*IHDP ($d = 25$) with increasing sample size*							
	100	1.608 ±.02	**1.565** ±.25	8.076 ±2.4	**1.956** ±.64	1.944 ±.00	**1.542** ±.18	24.53 ±5.1	**11.52** ±8.4
	250	2.335 ±.00	**1.637** ±.02	8.607 ±.82	**1.964** ±.49	2.219 ±.00	**1.627** ±.03	35.37 ±3.8	**18.13** ±7.8
	500	2.177 ±.00	**1.536** ±.00	4.216 ±.35	**1.739** ±.36	2.233 ±.00	**1.535** ±.00	26.06 ±4.5	**10.18** ±5.6

Table 3.5 Results on (semi-)synthetic data with `Ridge Regression`. Results are averaged over ten runs with and without the same representations in Table 3.1.

Methods		**X-Learner**		**DR-Learner**		**T-Learner**		**R-Learner**	
	EBM	✗	✓	✗	✓	✗	✓	✗	✓
d	N	*Synth. data with increasing sample size and increasing dimensions*							
50	100	2.373 ±.00	**2.001** ±.02	10.53 ±4.8	**2.028** ±.06	2.471 ±.00	**1.997** ±.02	53.24 ±5.3	**11.33** ±3.9
100	250	2.802 ±.00	**2.021** ±.01	8.769 ±1.5	**2.041** ±.05	2.871 ±.00	**2.021** ±.01	130.6 ±75.	**22.30** ±8.3
150	500	2.581 ±.00	**2.001** ±.01	3.074 ±.10	**2.001** ±.01	2.601 ±.00	**2.001** ±.01	50.13 ±4.1	**12.31** ±1.8
200	1k	2.187 ±.00	**1.942** ±.01	2.304 ±.03	**1.941** ±.01	2.192 ±.00	**1.941** ±.01	42.77 ±3.8	**2.839** ±.62
250	1.5k	2.270 ±.00	**1.958** ±.02	2.412 ±.04	**1.977** ±.03	2.274 ±.00	**1.957** ±.02	42.03 ±3.5	**2.511** ±.17
	N	*Synth. data with increasing sample size and dimensions fixed at $d = 100$*							
	100	2.124 ±.00	**1.946** ±.03	78.12 ±24.	**2.145** ±.11	2.272 ±.00	**1.948** ±.04	58.24 ±4.9	**5.346** ±1.9
	250	2.802 ±.00	**2.018** ±.00	4.489 ±.26	**2.025** ±.01	2.871 ±.00	**2.019** ±.01	60.27 ±4.1	**9.607** ±3.0
	500	2.152 ±.00	**2.059** ±.04	2.373 ±.08	**2.003** ±.05	2.164 ±.00	**2.056** ±.04	42.13 ±3.5	**3.755** ±.83
	1k	2.062 ±.00	**1.984** ±.02	2.109 ±.01	**2.022** ±.09	2.064 ±.00	**1.983** ±.02	39.49 ±3.2	**2.608** ±.19
	1.5k	2.013 ±.00	**2.003** ±.02	2.052 ±.01	**2.398** ±1.2	2.014 ±.00	**2.001** ±.02	37.19 ±2.9	**3.954** ±.77
	N	*Twins ($d = 48$) with increasing sample size*							
	500	0.182 ±.00	**0.151** ±.00	1.161 ±.20	**0.168** ±.02	0.183 ±.00	**0.151** ±.00	134.6 ±12.	**0.842** ±.22
	1k	0.196 ±.00	**0.159** ±.00	0.261 ±.02	**0.172** ±.01	0.196 ±.00	**0.159** ±.00	58.70 ±4.4	**0.455** ±.14
	1.5k	0.166 ±.00	**0.156** ±.00	0.171 ±.01	**0.159** ±.00	0.166 ±.00	**0.156** ±.00	29.72 ±2.4	**0.415** ±.14
	2k	0.163 ±.00	**0.153** ±.00	0.319 ±.04	**0.157** ±.00	0.163 ±.00	**0.153** ±.00	176.0 ±13.	**0.601** ±.30
	2.5k	0.169 ±.00	**0.162** ±.00	0.321 ±.04	**0.164** ±.00	0.169 ±.00	**0.162** ±.00	207.8 ±17.	**0.479** ±.19
	N	*IHDP ($d = 25$) with increasing sample size*							
	100	1.807 ±.00	**1.673** ±.02	5.933 ±1.2	**2.456** ±.47	1.739 ±.00	**1.675** ±.02	23.56 ±3.2	**14.19** ±9.3
	250	1.659 ±.00	**1.579** ±.03	2.883 ±.12	**2.426** ±.09	1.693 ±.00	**1.577** ±.03	11.01 ±2.0	**7.985** ±3.1
	500	1.625 ±.00	**1.614** ±.01	1.819 ±.16	**1.673** ±.09	1.641 ±.00	**1.610** ±.02	6.614 ±.59	**5.602** ±1.6

3.10.2 Hyperparameters

We report our chosen hyperparameters for each sample-size in Table 3.6. We noticed that the architecture and amount of noisy samples made little difference to performance in PEHE. The perturbation probability, and the value of k *did* make a difference, especially in larger sample sizes. We use EconML (Microsoft Research, 2019) to evaluate the CATE learners. We keep the hyperparameters for each learner as their default. In Table 3.1 we replace each regressor by a `KernelRidge` regressor, and each classifier by a support vector machine (`SVC`); both implemented by Pedregosa et al. (2011).

Table 3.6 Chosen hyperparameters for Table 3.1. We performed hyperparamter sweeps for each setup using a Bayesian optimisation scheme (Biewald, 2020). Our searched ranges are reported in Table 3.7. Twins settings were also used in Figure 3.2, but with a fixed $k = 5$ for both AE and EBM. Each integer (separated by a dash) in "Architecture" indicates layer width; "20-20" thus means a neural network with two hidden layers, each of width 20.

Setup		b	k	Architecture	Perturbation prob.
d	N		Synth. data, increasing dim		
50	100	10	3	20-20-20	0.20
100	250	10	4	20-20-20	0.50
150	500	5	3	20-20	0.20
200	1k	3	15	20-20-20-20	0.50
250	1.5k	3	20	20-20-20	0.50
	N		Synth. data, fixed dim (d=100)		
	100	5	15	20-20-20-20-20-20	0.20
	250	10	4	20-20-20	0.50
	500	3	10	20-20-20-20	0.50
	1k	3	20	20-20	0.35
	1.5k	3	10	20-20	0.30
	N		Twins, increasing N		
	500	5	15	20-20-20-20-20-20	0.45
	1k	5	16	20-20-20-20-20-20	0.55
	1.5k	5	16	20-20-20-20-20-20	0.55
	2k	4	14	20-20-20-20-20-20	0.55
	2.5k	4	12	20-20-20-20-20-20	0.50
	N		IHDP, increasing N		
	100	1	5	36-36-36-36-36-36	0.45
	250	1	5	36-36-36-36-36-36	0.45
	500	1	5	36-36-36-36-36-36	0.45

Table 3.7 Ranges for hyperparameter sweeps in Table 3.6. For each setup: (I) Synth. data, increasing dim, (II) Synth. data, fixed dim (d=100) (III) Twins, increasing dim, and (IV) IHDP, increasing dim; we used a Bayesian optimization (BO) scheme to find our selected hyperparameters. In or BO setup, we maximized the loss (3.6) on a (20%) validation-set.

Setup	b	k	# layers	Perturbation prob.
(I)	$\mathcal{U}(1;2;...;10)$	$\mathcal{U}(3;4;...;25)$	$\mathcal{U}(2;3;4;5;6)$	$\mathcal{U}(0.2;0.8)$
(II)	$\mathcal{U}(1;2;...;10)$	$\mathcal{U}(3;4;...;25)$	$\mathcal{U}(2;3;4;5;6)$	$\mathcal{U}(0.2;0.8)$
(III)	$\mathcal{U}(1;2;...;10)$	$\mathcal{U}(3;4;...;25)$	$\mathcal{U}(2;3;4;5;6)$	$\mathcal{U}(0.2;0.8)$
(IV)	$\mathcal{U}(1;2;...;10)$	$\mathcal{U}(3;4;...;25)$	$\mathcal{U}(2;3;4;5;6)$	$\mathcal{U}(0.2;0.8)$

Chapter 4

Overlapping representations for the estimation of conditional average treatment effects

The choice of making an intervention depends on its potential benefit or harm in comparison to alternatives. Estimating the likely outcome of alternatives from observational data is a challenging problem as *all* outcomes are never observed, and treatment assignment bias precludes the direct comparison of differently intervened groups. Despite their empirical success, we show that algorithms that learn domain-invariant representations of inputs (on which to make predictions) are often inappropriate, and develop generalization bounds that demonstrate the dependence on domain overlap and highlight the need for invertible latent maps. Based on these results, we develop a deep kernel regression algorithm and posterior regularization framework that outperforms a variety of baseline methods in simulations.

4.1 Introduction

Counterfactual estimation poses the question of what would have been the outcome if a different intervention had been applied. In order to make decisions in complex domains, making predictions on the causal effects of different actions and how these may vary across individuals is critical. In this chapter, we focus on the problem of making these predictions based on observational data, which is increasingly available in many domains such as medicine, public policy and advertising. In this setting, past actions, outcomes and context are available, but not the treatment assignment mechanism – we do not know why a given individual was

intervened or not. The treatment assignment mechanism will often be causally affected by context variables that also causally influence the outcome. As an example in Figure 4.1, suppose the context variable (X_i) is individual i's motivation for finding a new job, the treatment (A_i) is a government training program and the outcome (Y_i) is how soon individual i finds a new job. It is often that motivated individuals are more likely to both take advantage of the government training program *and* find a new job soon.

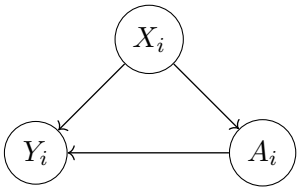


Figure 4.1 Causal graph of three random variables X_i (individual's motivation), A_i (government training program) and Y_i (employment outcome).

Learning from observational data requires adjusting for the covariate shift that exists between groups of individuals that are observed to have received different interventions. The challenge is how to untangle confounding factors and make valid predictions of counterfactual outcomes. Recent methods in machine learning have predominantly focused on learning representations regularized to balance these confounding factors by enforcing domain invariance with distributional distances (Johansson et al., 2016, 2018a; Yao et al., 2018). In this chapter, we argue that domain invariance is often too strict a requirement; overlapping support is sufficient for identifiability of the causal effect and equality in densities is not necessary. We interpret the loss in the predictive power of domain-invariant representations by the loss of information in the input variables that causally influence the treatment assignment, which is also often highly predictive of the treatment effect. Consider the example above for illustration: it is because motivation is predictive of the employment outcome that it confounds the treatment assignment (i.e. the assignment of the government training program). If enforcing domain invariance requires removing the predictive information of the treatment, e.g., motivation, which is also the key information to predict the employment outcome.

We introduce an optimization framework based on regularizing posterior distributions of the treatment effect that includes existing representation learning algorithms for different choices of regularization terms. We take advantage of this framework to introduce a novel type of regularization criterion for the problem of treatment effect estimation: the posterior counterfactual variance for enforcing domain overlap, and invertible representations to preserve the information content of the underlying context. Such an objective enjoys better generalization in small sample regimes, smoother representation surfaces with respect to the outcomes, as can be seen in Figure 4.2, and a Bayesian treatment of parameters which allows consistent uncertainty estimation in predictions. In summary, our contribution is 3-fold: we

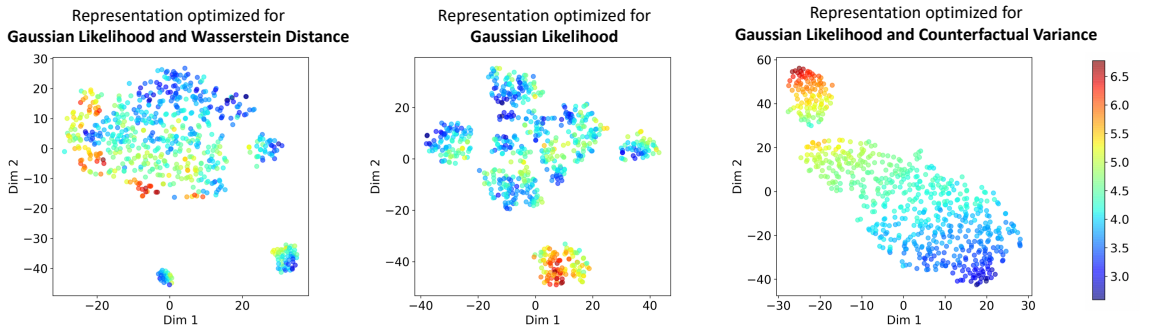


Figure 4.2 T-SNE (Van der Maaten and Hinton, 2008) visualizations of the learnt embeddings for the control potential outcomes of the IHDP dataset. Each panel shows representations regularized by different criteria and the coloured heatmap represents different outcome magnitudes with different colours. The *left panel* shows representations regularized by the Wasserstein distributional distance and results in poor discrimination. The *middle panel* shows representations optimized only for the factual data with the Gaussian likelihood. The *right panel* shows representations regularized by the counterfactual variance, our proposed criterion. Much better separation in outcomes is obtained by regularizing for the predictive variance, in contrast to using integral probability metrics such as the Wasserstein distance.

develop a theory to justify regularizing for the posterior variance to improve generalization error and establish the limitations of distributional distances; we propose to use deep kernels (Wilson et al., 2016) and posterior regularization (Zhu et al., 2014) as a general framework to estimate conditional average treatment effects (CATEs), also called individualized treatment effects (ITEs); we provide an instantiation of this method informed by our generalization bounds, which improves CATE estimation compared with a variety of baseline methods.

4.2 Related work

Due to the ability of deep neural nets to learn rich representations of observed covariates, recent advances in estimating conditional average treatment effects (CATEs) have focused on learning representations invariant to the treatment assignment policy that achieve a small error on the factual data. The hope is that the learnt representation and prediction function can generalize to predict counterfactual outcomes. Several methods follow this approach. Johansson et al. (2016) propose learning a representation of the data that makes the treated and control distributions more similar, fitting a linear ridge-regression model on top of it. Shalit et al. (2017) build on their approach to derive a more flexible family of methods including non-linear hypotheses. However, both methods insist on quantifying divergence between treated and control groups with integral probability metrics.

In this chapter, we share the need for good representations but argue for enforcing support overlap rather than equality in densities. Inspired by nearest-neighbour methods, Yao et al. (2018) learn a representation that preserves local similarity information in feature space and was able to show a decrease in the generalization error in counterfactual estimation. Kallus (2020) develops a representation learning method using a discriminative discrepancy metric which is learnt by solving a game between a weighting and a discriminator network through adversarial training. Adapting Bayesian methods for the problem of conditional average treatment effects has attracted a lot of interest, in particular in the field of medicine where quantifying uncertainty is important. Alaa and van der Schaar (2017) regularize counterfactual predictions through their posterior variance and similarly stress the importance to provide confidence in their estimates using credible intervals, but did not investigate the generalization properties of their method and only allowed for limited expressiveness in their method. Similarly, Jean et al. (2018) use posterior variance regularization to learn from unlabeled data in situations where labelled data is scarce for improved performance.

Our work has also strong connections with work on domain adaptation. In particular, estimating CATEs requires predictions of outcomes over a different distribution from the observed one. Our upper bound of the error in estimating CATEs has similarities with generalization bounds in domain adaptation given by Johansson et al. (2016, 2018a). Johansson et al. (2018a); Zhao et al. (2019) similarly argue against enforcing domain-invariance and related the loss of predictive power of those representations to the loss of information due to the non-invertibility of learnt representations. Covariates matching (Imbens, 2015; Rosenbaum, 1989, 2002a; Stuart, 2010; Stuart et al., 2004) is class of methods in statistics that can directly estimate average causal effects by comparing the matched individuals in an observational studies. Representation learning methods in machine learning and matching methods are similar in terms of measuring the mismatch between the treated and control groups in a study. The former focuses on achieving the best performance of a black-box neural network in estimating treatment effects while the latter is used for better transparency and interpretability in estimating treatment effects. Finally, weighting methods (Crump et al., 2006, 2009; Li et al., 2018) can potentially deal with limited overlap in the estimation of average treatment effects. Extending this approach to estimate CATEs with limited overlap is an interesting direction for future research.

4.3 Setup

Consider a population of N individuals (i.e. units) with each individual i described by a context (a set of covariates) $X_i \in \mathcal{X} \subset \mathbb{R}^d$, a treatment variable $A_i \in \{0, 1\}$ and an outcome variable $Y_i \in \mathcal{Y} \subset \mathbb{R}$. Individual i's outcome to the treatment is a random

variable denoted by $Y_i(1)$, whereas individual i's natural outcome without the treatment is denoted by $Y_i(0)$. Let $\mathcal{I}_a = \{i \in [N] : A_i = a\}$, $a = 0, 1$, denote the control and treated populations, respectively. For $a = 0, 1$, we define $N_a = |\mathcal{I}_a|$, $\boldsymbol{X}_a = (X_i : i \in \mathcal{I}_a)$ and $\boldsymbol{Y}_a = (Y_i : i \in \mathcal{I}_a)$. To estimate the CATE $\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$, we make the standard assumptions in observational studies (Assumption 7). Under this assumption, it holds that $\mu_a(x) := \mathbb{E}[Y(a)|X = x] = \mathbb{E}[Y|X = x, A = a]$. Then we can estimate $\mathbb{E}[Y(a)|X = x]$ by a regression model $\hat{\mu}_a : \mathcal{X} \to \mathcal{Y}$, $a = 0, 1$. More specifically in the setup of Gaussian process regression (Williams and Rasmussen, 2006), for $a = 0, 1$, under a Gaussian prior distribution,

$$\pi_a(x) = \mathcal{N}\left(0, K(x, x)\right), \tag{4.1}$$

with a covariance kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, our Bayesian outcome regression model has a Gaussian posterior distribution

$$\hat{\rho}_a(x) = \mathcal{N}\left(\hat{\mu}_a(x), \hat{\sigma}_a^2(x)\right). \tag{4.2}$$

The posterior distribution $\hat{\rho}_a$ is a distribution conditional on the observed samples $(\boldsymbol{X}_a, \boldsymbol{Y}_a)$. But we omit the conditioning on $(\boldsymbol{X}_a, \boldsymbol{Y}_a)$ in our notation. Intuitively, the kernel function $K(\cdot, \cdot)$ defines the function class $\mathcal{F}$ we will work with; π_a and $\hat{\rho}_a$ are two distributions over $\mathcal{F}$, i.e., a random draw from $\pi_a(\cdot)$ or $\hat{\rho}_a(\cdot)$ is a function $f(\cdot) \in \mathcal{F}$.

Assumption 10. We assume the data space $\mathcal{X} \times \mathcal{Y}$ is compact and any function $f \in \mathcal{F}$ is bounded in l_2 norm, i.e., $\mathcal{F} = \{f : \mathbb{E}\{f^2(X)\} < \infty\}$.

We quantify the accuracy of a CATE estimator $\hat{\tau}$ by its mean squared error in estimating τ, also called the empirical precision in estimating heterogeneous effects (PEPE),

$$\epsilon_{\text{PEHE}} = \int \left(\hat{\tau}(x) - \tau(x)\right)^2 p_X(x) dx.$$

Estimating τ for unobserved individuals involves predictions of both potential outcomes, but we never observe the counterfactual outcomes or the true treatment effect in an observational dataset. This makes the problem of causal inference fundamentally different from supervised learning. In the next section, suppose that we estimate $\tau(x)$ by a T-learner $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$, we analyse the generalization properties of $\hat{\tau}$ with respect to ϵ_{PEHE} under Assumptions 7 and 10. Section 1.3.2 reviews a variety of CATE learners which leverages the smoothness of τ. The method we will present later is a regularization method for the outcome regression models $\hat{\mu}_0$ and $\hat{\mu}_1$. The regularized models can be used to generate pseudo-outcomes for any CATE learner as usual. Our analysis still holds assuming that we use our method in other learners because the errors of outcome regression and pseudo-outcome regression are separated in the error analysis of those advanced learners.

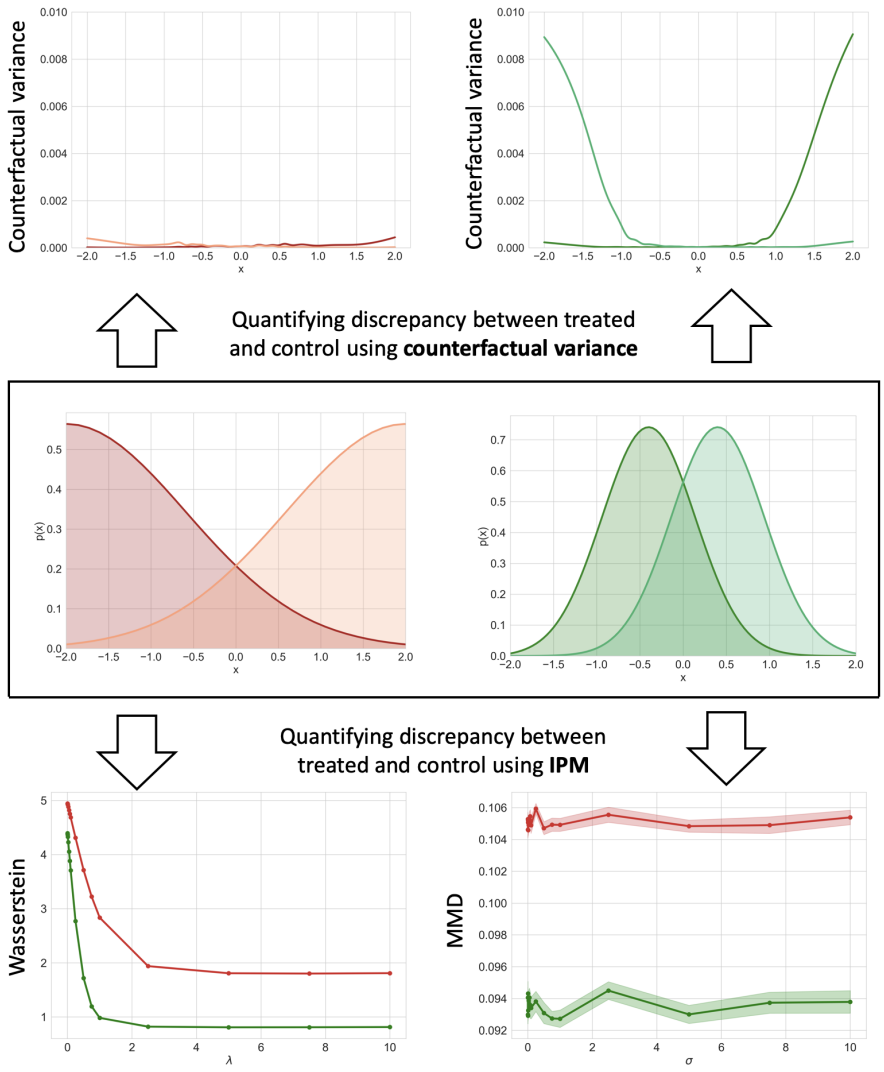


Figure 4.3 Toy example illustrates the shortcomings of distributional distances, like integral probability metrics (IPMs), for regularizing representations in causal inference. Despite the fact that sufficient support is satisfied in the red populations and not in the green populations, IPMs (bottom) give the opposite result, with a larger discrepancy in the red populations than in the green populations. In contrast, the counterfactual variance predicted by $\hat{\sigma}_a^2$ (top) accurately describes the lack of support in the green populations.

4.4 Intuition and theoretical results

Inherent to the approach of learning representations for counterfactual inference is that the representation must trade-off between containing predictive information about factual outcomes while mitigating the information content that drives the treatment assignment policy to ensure good generalization on counterfactuals. In this section, we make several

observations about the deficiencies of enforcing domain invariance for this purpose and propose alternatives based on the posterior counterfactual variance.

Example 4 (Counterfactual variance vs. Distributional distances)**.** In the middle panels of Figure 4.3 we show two simulated datasets. The left-hand/red dataset arises from two truncated normal distributions with a large overlap in the tails; the right-hand/green dataset arises from two ordinary normal distributions with a small overlap in the tails. In both cases, we show the treated and control populations in different shades. In both cases, the outcome is $y = \text{sinc}(4x)$. The red populations satisfy sufficient assumptions for identifiability of causal effects; the green populations do not. However, as shown in the bottom panels, both integral probability metrics (IPMs), the maximum mean discrepancy (MMD) (Gretton et al., 2012) and Wasserstein distances (Villani, 2009) are *smaller* in the green populations than in the red populations. In contrast, the top panel shows that the predictive variance of the counterfactual outcomes, $\hat{\sigma}^2(X_i), \forall i \in \mathcal{I}_{1-a}$, much better describes this lack of overlap. Counterfactual variance is adaptive to the prediction problem of interest, providing a data-dependent measure to quantify distances in the underlying function class, perhaps more precise when the underlying function to be estimated are unknown. IPMs are defined as worst-case distances dependent on a pre-specified function class. We make the observation also that IPMs need to be approximated in practice which may be inaccurate for high-dimensional and small training data samples (Boissard and Le Gouic, 2014).

4.4.1 Generalization bounds

We develop a PAC-Bayes generalization bound (McAllester, 1999; Shawe-Taylor and Williamson, 1997) for $\epsilon_{\text{PEHE}} = \mathbb{E}\left\{[\hat{\mu}_1(X) - \hat{\mu}_0(X) - \tau(X)]^2\right\}$ with $\hat{\mu}_0$ and $\hat{\mu}_1$ defined in (4.2), that shows specifically why minimizing counterfactual variance can improve generalization performance. The proofs the theoretical results in this subsection can be found in Section 4.8.

Theorem 5. *Under Assumptions 7 and 10, for any $\delta \in (0, 1]$ and π_a in (4.1) and $\hat{\rho}_a$ in (4.2), with probability at least $1 - \delta$, we have*

$$\begin{aligned}\epsilon_{\text{PEHE}} \leq \sum_{a=0}^{1} \Bigg[& 2C_a L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) + (C_a + 1) V(\boldsymbol{X}_a; \hat{\rho}_a) + V(\boldsymbol{X}_{1-a}; \hat{\rho}_a) \\ & + \left(\frac{1}{2\sqrt{N_{1-a}}} + \frac{2C_a}{\sqrt{N_a}} \right) \left(2\,\text{KL}(\hat{\rho}_a \| \pi_a) + \ln(2/\delta) + C_2 \right) \Bigg],\end{aligned} \tag{4.3}$$

where the mean squared error $L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) = N_a^{-1} \sum_{i \in \mathcal{I}_a} \left[Y - \hat{\mu}_a(X_i) \right]^2$, the mean factual variance $V(\boldsymbol{X}_a; \hat{\rho}_a) = N_a^{-1} \sum_{i \in \mathcal{I}_a} \hat{\sigma}_a^2(X_i)$, the mean counterfactual variance $V(\boldsymbol{X}_{1-a}; \hat{\rho}_a) = N_{1-a}^{-1} \sum_{i \in \mathcal{I}_{1-a}} \hat{\sigma}_a^2(X_i)$, the constant $C_a = 1 + \sup_{x \in \mathcal{X}} \left[p_{X|A}(x \mid 1-a) / p_{X|A}(x \mid a) \right]$, C_2 is a universal constant and $\text{KL}(\cdot \| \cdot)$ is the Kullback-Leibler divergence.

The empirical terms in the upper bound (4.3) are $L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a)$, $V(\boldsymbol{X}_a; \hat{\rho}_a)$, $V(\boldsymbol{X}_{1-a}; \hat{\rho}_a)$ and $\mathrm{KL}(\hat{\rho}_a \| \pi_a)$. The variance $V(\boldsymbol{X}_a; \hat{\rho}_a)$ is also part of the marginal log-likelihood of our Bayesian model; see (4.6) for details. As the sample size ($N = N_{1-a} + N_a$) increases and we keep minimizing these empirical terms to zero, we would have $\epsilon_{\text{PEHE}} \to 0$, i.e., $\hat{\tau} \to \tau$ in l_2 norm consistently. If the sample size is large, minimizing $\mathrm{KL}(\hat{\rho}_a \| \pi_a)$ is unnecessary.

Suppose that $\hat{\mu}_a(x) = w_a^\top \phi(x)$ where ϕ is a feature map (parameterized by a neural network). Minimizing the counterfactual variance (as a measure of overlap in Figure 4.3) will learn a feature map ϕ that can increase the overlap between the treated and control populations in the space $\mathcal{Z}$. We can expect the representations ϕ of counterfactual data to encode a relatively smoother prediction function $w_a^\top \phi(x)$, as can be seen in Figure 4.2. The estimation of treatment effects is inherently a label-scarce problem as counterfactual data is not observed, representations resulting in smooth prediction curves are especially important to generalize beyond the factual data. This view of the problem of estimating treatment effects emphasizes the need for regularization for good generalization.

Why distributional distances may be inadequate?

While the toy example clearly illustrates the inability of distributional distances to capture domain overlap, we argue that by enforcing equality in full marginals, optimizing for distributional distances may also overly penalize the model's ability to predict factual observations when sufficient data is available, that is, when domain overlap is satisfied.

In the following theorem, we provide a bound on the generalization error of estimating counterfactual outcomes, which illustrates the interplay between distribution mismatch and prediction loss on factual data. We show that distribution mismatch between the treated and control populations becomes decreasingly relevant with increasing sample size.

Theorem 6. *Under Assumptions 7 and 10, for the posterior distribution $\hat{\rho}_a$ in (4.2),*

$$L_{1-a}(\hat{\rho}_a) \leq \frac{1}{2} V_{1-a}(\hat{\rho}_a) + D_{a,\infty} \left[L_a(\hat{\rho}_a) + \frac{1}{2} V_a(\hat{\rho}_a) \right] \tag{4.4}$$

where $L_c(\hat{\rho}_a) = E\left\{[Y - \hat{\mu}_a(X)]^2 \mid A = c\right\}$, $V_c(\hat{\rho}_a) = \mathbb{E}\left\{\hat{\sigma}^2(X) | A = c\right\}$ for $c = a, 1-a$, and $D_{a,\infty} = \sup_{x \in \mathcal{X}} \left[p_{X|A}(x \mid 1-a) / p_{X|A}(x \mid a) \right]$.

The bound Equation (4.4) describes the interaction between the distribution mismatch and the prediction error on factual data $L_a(\hat{\rho}_a) = E\left\{[Y - \hat{\mu}_a(X)]^2 \mid A = a\right\}$. The term $D_{a,\infty}$ is large if, for some x, $p_{X|A}(x \mid a)$ is small while $p_{X|A}(x \mid 1-a)$ is large (that is when there is poor overlap between them), which understandably, makes minimizing the counterfactual

risk $L_{1-a}(\hat{\rho}_a)$ harder because few examples from the other population are observed for a given context. However, note that $D_{a,\infty}$ is multiplied by the expected factual loss $L_a(\hat{\rho}_a)$ (which decreases as N_a increases if $\hat{\mu}_a$ is constant estimator of μ_a). The distribution mismatch thus becomes less important for generalization, if we can minimize the expected factual loss $L_a(\hat{\rho}_a)$ arbitrarily well. This suggests that optimizing for distributional distances between treated and control groups *at the expense* of prediction error on the factual data may be counterproductive. Representations regularized with distributional distances may thus shrink the function class and converge on solutions that, although balanced between treated and control groups, lose their predictive power.

4.4.2 Why encourage preserving information content?

Assumption 7 gives sufficient conditions for identifying treatment effects conditioning on X, but identifiability need not hold with respect to the feature representation $Z = \phi(X)$, even if it does with respect to X. For instance consider $\phi^{-1}(z) := \{x : \phi(x) = z\}$,

$$\begin{aligned} p_{Y(c)|Z,A}(y|z,a) &= \frac{\int_{x\in\phi^{-1}(z)} p_{Y(c)|X,A}(y|x,a)p_{X|A}(x|a)}{\int_{x\in\phi^{-1}(z)} p_{X|A}(x|a)} \\ &= \frac{\int_{x\in\phi^{-1}(z)} p_{Y(c)|X}(y|x)p_{X|A}(x|a)}{\int_{x\in\phi^{-1}(z)} p_{X|A}(x|a)} \\ &\neq p_{Y(c)|Z}(y|z), \end{aligned}$$

In general, with equality only if ϕ is invertible, i.e, $\phi^{-1}(z)$ corresponds to a point $x^* \in \mathcal{X}$ The conditional independence in the unconfoundedness assumption required for estimating treatment effects need not hold for non-invertible transformations. In this sense, we may be introducing unobserved confounders in representation space we hypothesize by the information lost in the map ϕ. Observe also that our objective, the conditional average treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$, is expressed in terms of expectations. Similarly, it holds that in feature space, $\mathbb{E}[Y(1) - Y(0)|Z = \phi(x)] = \int_{x\in\phi^{-1}(z)} \mathbb{E}[Y(1) - Y(0) \mid X = x]dx$ will *not* be equal to our quantity of interest $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ unless ϕ is invertible.

4.5 DKITE

In this section, we describe a method for counterfactual estimation, called DKLITE (Deep Kernel Learning for Individualized Treatment Effects), motivated by our analysis. We keep the method's name the same as in our original publication (Zhang et al., 2020). Individualized treatment effects is another name for CATEs in the literature. Our proposed method works

with a feature map $\phi : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^{d_\phi}$ that defines a kernel function $K(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$. We let ϕ be parameterized by a neural network to encode the information content of input covariates. Neural network-based kernels are called *deep kernels* in the literature (Wilson et al., 2016). The dimension d_ϕ of $\mathcal{Z}$ can be chosen arbitrarily.

For individual i with $A_i = a$, we assume its observed outcome follows the linear model

$$Y_i = w_a^\top \phi(X_i) + \epsilon_{i,a}, \ a = 0, 1, \tag{4.5}$$

where w_a is a d_ϕ-dimensional weight vector and $\epsilon_{i,a} \sim \mathcal{N}(0, \beta_a^{-1})$ is a noise variable, $a = 0, 1$.

4.5.1 Predictive distribution

By using a Gaussian prior distribution $w_a \sim \mathcal{N}(0, \lambda_a^{-1}\mathbf{1})$, the posterior of w_a is given by

$$p(w_a | \boldsymbol{X}_a, \boldsymbol{Y}_a, \phi) = \mathcal{N}\left(\boldsymbol{m}_a, \boldsymbol{K}_a^{-1}\right),$$

where the mean vector $\boldsymbol{m}_a$ and the kernel matrix $\boldsymbol{K}_a$ are given by

$$\boldsymbol{m}_a = \beta_a \boldsymbol{K}_a^{-1} \boldsymbol{\Phi}_a^\top \boldsymbol{Y}_a, \quad \boldsymbol{K}_a = \beta_a \boldsymbol{\Phi}_a^\top \boldsymbol{\Phi}_a + \lambda_a \boldsymbol{I}_{d_\phi \times d_\phi},$$

respectively, and $\boldsymbol{\Phi}_a = (\phi(X_i) : i \in \mathcal{I}_a)$ is the representation of $\boldsymbol{X}_a$. The posterior derivation is based on Bayesian linear regression; see Bishop (2006, Chapter 3.3) for more details.

Then in our assumed model (4.5), the prior distribution $\pi_a(x)$ in (4.1) has a kernel function

$$K(x, x) = \lambda_a^{-1} \phi^\top(x) \phi(x).$$

To predict the outcomes at a given x, we use the posterior distribution $\hat{\rho}_a(x)$ in (4.2) with

$$\hat{\mu}(x) = \boldsymbol{m}_a^\top \phi(x) \text{ and } \hat{\sigma}^2(x) = \phi(x)^\top \boldsymbol{K}_a^{-1} \phi(x) \text{ for } a = 0, 1.$$

Point estimates are given for example by the posterior mean or median, optimal for minimizing squared loss or absolute loss, respectively. Note also that knowledge of the full posterior allows us to quantify our uncertainty around point estimates through credible intervals, especially useful in medicine and public policy, for example.

4.5.2 Learning ϕ

Factual Likelihood. Marginalizing out w_a w.r.t its prior, we can define the negative marginal log-likelihood of (4.5) (Williams and Rasmussen, 2006, Chapter 2.7.1), denoted by $\mathcal{L}_a$. We rewrite $\mathcal{L}_a$ to encode our insights from Section 4.4:

$$\begin{aligned}
\mathcal{L}_a &= -\log\left[p(\boldsymbol{Y}_a \mid \boldsymbol{X}_a, \phi)\right] \\
&= -\frac{d_\phi}{2}\ln\lambda_a - \frac{N_a}{2}\ln\beta_a + \frac{N_a}{2}\ln(2\pi) + \frac{\beta_a}{2}\|\boldsymbol{Y}_a - \boldsymbol{\Phi}_a\boldsymbol{m}_a\|^2 + \frac{\lambda_a}{2}\boldsymbol{m}_a^\top\boldsymbol{m}_a + \frac{1}{2}\ln|\boldsymbol{K}_a| \\
&= \mathrm{KL}(\hat{\rho}_a\|\pi_a) + \frac{d_\phi}{2} - \frac{\lambda_a}{2}\mathrm{Tr}(\boldsymbol{K}_a^{-1}) + \frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{\beta_a}{2}\|\boldsymbol{Y}_a - \boldsymbol{\Phi}_a\boldsymbol{m}_a\|^2 \\
&= \mathrm{KL}(\hat{\rho}_a\|\pi_a) + \frac{1}{2}\mathrm{Tr}(\boldsymbol{K}_a\boldsymbol{K}_a^{-1}) - \frac{\lambda_a}{2}\mathrm{Tr}(\boldsymbol{K}_a^{-1}) + \frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{\beta_a}{2}\|\boldsymbol{Y}_a - \boldsymbol{\Phi}_a\boldsymbol{m}_a\|^2 \\
&= \mathrm{KL}(\hat{\rho}_a\|\pi_a) + \frac{1}{2}\mathrm{Tr}((\boldsymbol{K}_a - \lambda_a\boldsymbol{I})\boldsymbol{K}_a^{-1}) + \frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{\beta_a}{2}\|\boldsymbol{Y}_a - \boldsymbol{\Phi}_a\boldsymbol{m}_a\|^2 \\
&= \mathrm{KL}(\hat{\rho}_a\|\pi_a) + \frac{\beta_a}{2}\mathrm{Tr}((\boldsymbol{\Phi}_a^\top\boldsymbol{\Phi}_a\boldsymbol{K}_a^{-1}) + \frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{\beta_a}{2}\|\boldsymbol{Y}_a - \boldsymbol{\Phi}_a\boldsymbol{m}_a\|^2 \\
&= \mathrm{KL}(\hat{\rho}_a\|\pi_a) + \frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{\beta_a}{2}\mathrm{Tr}((\boldsymbol{\Phi}_a\boldsymbol{K}_a^{-1}\boldsymbol{\Phi}_a^\top) + \frac{\beta_a}{2}\|\boldsymbol{Y}_a - \boldsymbol{\Phi}_a\boldsymbol{m}_a\|^2 \\
&= \mathrm{KL}(\hat{\rho}_a\|\pi_a) + \frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{N_a\beta_a}{2}\Big[\hat{V}_a(\boldsymbol{X}_a;\hat{\rho}_a) + \hat{L}_a(\boldsymbol{X}_a,\boldsymbol{Y}_a;\hat{\rho}_a)\Big],
\end{aligned} \tag{4.6}$$

where the third equality is achieved by

$$\mathrm{KL}(\hat{\rho}_a\|\pi_a) = \frac{\lambda_a}{2}\mathrm{Tr}(\boldsymbol{K}_a^{-1}) + \frac{\lambda_a}{2}\boldsymbol{m}_a^\top\boldsymbol{m}_a - \frac{d_\phi}{2} - \frac{d_\phi}{2}\ln\lambda_a + \frac{1}{2}\ln|\boldsymbol{K}_a|.$$

Summing up the likelihood for $a = 0, 1$, we define

$$\mathcal{L}_{\text{lik}} = \sum_{a=0}^{1}\left(\frac{N_a}{2}\ln(2\pi\beta_a^{-1}) + \frac{N_a\beta_a}{2}\Big[\hat{V}_a(\boldsymbol{X}_a;\hat{\rho}_a) + \hat{L}_a(\boldsymbol{X}_a,\boldsymbol{Y}_a;\hat{\rho}_a)\Big] + \mathrm{KL}(\hat{\rho}_a\|\pi_a)\right)$$

We see from Theorem 5 that the empirical quantities to optimize for in the upper-bound (4.3) are $V(\boldsymbol{X}_{1-a};\hat{\rho}_a)$, $V(\boldsymbol{X}_a;\hat{\rho}_a)$, $L(\boldsymbol{X}_a,\boldsymbol{Y}_a;\hat{\rho}_a)$ and $\mathrm{KL}(\hat{\rho}_a\|\pi_a)$, $a = 0, 1$. The latter three already exist in the final expression of the likelihood $\mathcal{L}$ above.

Counterfactual variance. Therefore, by including an empirical estimate of the counterfactual variance $V(\boldsymbol{X}_{1-a};\hat{\rho}_a)$ as a regularizer in our objective function we are effectively optimizing for all terms given by the PAC-Bayes upper-bound of ϵ_{PEHE} in Theorem 5. We write this regularization term as

$$\mathcal{L}_{\text{var}} = \sum_{a=0}^{1} V(\boldsymbol{X}_{1-a};\hat{\rho}_a) = \sum_{a=0}^{1} N_{1-a}^{-1}\sum_{i\in\mathcal{I}_{1-a}}\hat{\sigma}^2(X_i) \tag{4.7}$$

The neural net ϕ is encouraged to learn a feature representation in which the counterfactual examples are close to the factual examples, thereby reducing the variance in our predictions. The implication is that we are optimizing for representations where counterfactual data tend to cluster around the representations of factual data. This is a way to see how intuitively we are encouraging overlap in support in representation space *without* enforcing equality in densities (i.e. the size of the factual and counterfactual clusters need not coincide).

Invertibility as regularization. While the loss due to non-invertible representations is not directly observable, we may associate it with the information content of x lost in $\phi(x)$ and we found it to be an important source of gain in performance, empirically. We can encourage information content preservation with an additional decoder $\psi : \mathbb{R}^{d_\phi} \to \mathcal{X}$, which is a neural network with the reversed structure of network ϕ trained to reconstruct the input x from $\phi(x)$. The reconstruction loss is given by

$$\mathcal{L}_{\text{rec}} = \sum_{a=0}^{1} N_a^{-1} \sum_{i \in \mathcal{I}_a} \|X_i - \psi(\phi(X_i))\|_2^2 \tag{4.8}$$

We note that there are other advanced techniques (Behrmann et al., 2019; Rezende and Mohamed, 2015) in machine learning that can be used to achieve strict invertibility, so the reconstruction loss $\mathcal{L}_{\text{rec}}$ is not needed if we apply any of these techniques.

Final loss function. Based on the objectives described above, our final loss trades-off between maximizing the likelihood of the observed (factual) data under our model, minimizing the predictive variance of the counterfactual outcomes and minimizing the reconstruction loss of the representations. The loss is given by

$$\mathcal{L}_{\text{fin}} = \mathcal{L}_{\text{lik}} + \alpha_1 \mathcal{L}_{\text{var}} + \alpha_2 \mathcal{L}_{\text{rec}}, \tag{4.9}$$

where $\alpha_1 > 0$ and $\alpha_2 > 0$ are hyperparameters. Standard methods for hyperparameter selection, such as cross-validation, are not generally applicable for choosing hyperparameters because counterfactuals are never observed. As an approximation scheme, we follow the approach of Shalit et al. (2017), and replace the missing counterfactuals with their nearest factual neighbour (in the opposite group) to compute a (approximated) treatment effect for each example and optimize hyperparameters for the PEHE.

Table 4.1 Performance of DKLITE and benchmarks: mean and standard deviation of $\sqrt{\text{PEHE}}$.

Dataset	IHDP		Twins	
Method	In-sample	Out-sample	In-sample	Out-sample
OLS/LR_1	5.8 ± .3	5.8 ± .3	.319 ± .001	.318 ± .007
OLS/LR_2	2.4 ± .1	2.5 ± .1	.320 ± .002	.320 ± .003
BLR	5.8 ± .3	5.8 ± .3	.312 ± .003	.323 ± .018
k-NN	2.1 ± .1	4.1 ± .2	.333 ± .001	.345 ± .007
BART	2.1 ± .1	2.3 ± .1	.347 ± .009	.338 ± .016
R-Forest	4.2 ± .2	6.6 ± .3	.366 ± .002	.321 ± .005
C-Forest	3.8 ± .2	3.8 ± .2	.366 ± .003	.316 ± .011
BNN	2.2 ± .1	2.1 ± .1	.325 ± .003	.321 ± .018
TARNET	.88 ± .02	.95 ± .02	.317 ± .002	.315 ± .003
CAR_{WASS}	.72 ± .02	.76 ± .02	.315 ± .007	.313 ± .008
CMGP	.63 ± .08	.74 ± .11	.320 ± .002	.319 ± .008
DKLITE	**.52 ± .02**	**.65 ± .03**	**.288 ± .001**	**.293 ± .003**

4.6 Experiments

Our experiments will compare DKLITE with benchmark methods, analyze the source of performance gain, and demonstrate the use of the posterior variance for decision-making.

Baseline methods. We compare DKLITE with a total of 11 methods. First we evaluate the least-squares regression using treatment as an additional input feature (OLS/LR_1), we consider separating the least-squares regression for each treatment (OLS/LR_2), we evaluate balancing linear regression (BLR) (Johansson et al., 2016), k-nearest neighbors (k-NN) (Crump et al., 2008), Bayesian additive regression trees (BART) (Chipman et al., 2010), random forests (R-Forest) (Breiman, 2001), causal forests (C-Forest) (Wager and Athey, 2018), balancing neural networks (BNN) (Johansson et al., 2016), treatment-agnostic representation network (TARNET), counterfactual regression with Wasserstein distance (CAR_{WASS}) (Shalit et al., 2017), and multi-task gaussian process (CMGP) (Alaa and van der Schaar, 2017).

Datasets. Causal inference models are often impossible to reliably validate using real-world data due to the absence of counterfactual outcomes. Various established approaches for evaluating causal models have been proposed, which we use for our analysis. We describe these briefly below and refer the reader to the accompanying references and Section 4.10 for further details. We consider *IHDP (747 instances described by 25 covariates)* (Alaa and van der Schaar, 2017; Hill, 2011; Shalit et al., 2017; Yao et al., 2018; Yoon et al., 2018) in which counterfactual outcomes are randomly generated via a predefined probabilistic model;

Twins (11300 instances described by 30 covariates), in which outcomes are observed but the treatment assignment in the dataset is simulated.

Metrics. The metrics used to evaluate each data set differ slightly depending on the available outcome (real or simulated). For IHDP, we use the empirical precision in estimating treatment effects $\hat{\epsilon}_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}(\tau(X_i) - \hat{\tau}(X_i))^2$. For Twins, we use the observed precision in estimating heterogeneous effects, $\tilde{\epsilon}_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2$. We report in-sample performance on the training samples with one missing outcome and out-sample performance on the testing samples with both outcomes missing.

4.6.1 Predictive performance

We report in-sample and out-of-sample performance in Table 4.1. DKLITE targets aspects of the treatment effect estimation problem that have not been considered before. Learning with such an objective outperforms all competing methods on both datasets. The most relevant comparison is perhaps with BNN (Johansson et al., 2016) and CAR_{WASS} (Shalit et al., 2017), neural network models that enforces domain invariance through distributional distances. The performance gain highlights the predictive power of our representations.

4.6.2 Source of gain

Table 4.2 Source of performance gain in DKLITE.

Dataset	$\sqrt{\text{PEHE}}$	$\mathcal{L}_{\text{lik}}$	$\mathcal{L}_{\text{lik}} + \mathcal{L}_{\text{var}}$	$\mathcal{L}_{\text{lik}} + \mathcal{L}_{\text{rec}}$
IHDP	In-sample	1.46 ± .01	.98 ± .04	.96 ± .04
	Out-sample	1.95 ± .14	1.24 ± .09	1.28 ± .13
Twins	In-sample	.308 ± .004	.292 ± .002	.291 ± .001
	Out-sample	.323 ± .008	.294 ± .007	.293 ± .003

In this section, we analyze more deeply the contribution to the performance gain of each component of our loss. We evaluate DKLITE optimized for different components of $\mathcal{L}_{\text{fin}} = \mathcal{L}_{\text{lik}} + \alpha_1\mathcal{L}_{\text{var}} + \alpha_2\mathcal{L}_{\text{rec}}$. As can be seen in Table 4.2, including regularization based on the counterfactual variance ($\mathcal{L}_{\text{lik}} + \alpha_1\mathcal{L}_{\text{var}}$) and reconstruction loss ($\mathcal{L}_{\text{lik}} + \alpha_2\mathcal{L}_{\text{rec}}$), each evaluated separately, already provides a significant gain in performance with respect to optimization on the factual data only ($\mathcal{L}_{\text{lik}}$). Importantly though, combining them ($\mathcal{L}_{\text{fin}}$) improves performance further by an order of magnitude (see DKLITE in Table 4.1), which suggests that $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{rec}}$ capture to some extent orthogonal sources of gain. The gain is especially important on

relatively smaller data sets, such as IHDP with 747 individuals, and to a lesser extent on bigger data sets. These results illustrate the behaviour suggested by (4.4) in Theorem 6: low overlap between groups becomes decreasingly relevant with increasing data set size. In this setting, the error on the factual data ($\mathcal{L}_{\text{lik}}$) drives generalization performance.

4.6.3 Leveraging the predicted uncertainty

Table 4.3 Performance of DKLITE and DKLITE-U on IHDP and Twins.

Dataset	$\sqrt{\text{PEHE}}$	DKLITE	DKLITE-U
IHDP	In-sample	.52 ± .02	.46 ± .02
	Out-sample	.60 ± .03	.53 ± .02
Twins	In-sample	.288 ± .001	.287 ± .001
	Out-sample	.293 ± .003	.292 ± .002

Data-driven solutions for decision support have most often been proposed without methods to quantify and control the uncertainty in a decision. In contrast, for example, in medicine, a physician knows whether she/he is uncertain about a case and will consult more experienced colleagues if needed. We use this idea to show that uncertainty informed treatment effect estimation can improve performance. An instantiation of this approach, termed DKLITE-U, is given by referring to the 10 % most uncertain predictions for further scrutiny. The uncertainty is measured by the posterior variances $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$. Performance in comparison to DKLITE is given in Table 4.3. Note that especially on small data sets, such as IHDP, DKLITE-U achieves a better PEHE by removing the 10 % most uncertain samples, which indicates that the posterior variances can capture the true CATE estimation error.

4.7 Discussion

In many domains, understanding the effect of interventions at an individual level is crucial, but predicting those potential outcomes is challenging. Despite their empirical success, we find that methods enforcing representations to satisfy domain invariance are often too strong a requirement for causal predictions. This stems from the fact that overlapping support is sufficient for identifiability of the causal effect and equality in densities is not necessary. We have proposed generalization bounds that show the dependence on domain overlap through the counterfactual variance which we interpret as a proxy for domain overlap, and highlighted the need for invertible latent maps. These results motivated the novel posterior regularization

method that we incorporated into a deep kernel learning framework, which led to a superior empirical performance in estimating CATEs.

4.8 Technical Proofs

4.8.1 Preliminary: PAC-Bayes theory

We first introduce the concepts called *Gibbs model* and *Bayesian model* in PAC-Bayesian theory (McAllester, 1999; Shawe-Taylor and Williamson, 1997). Given a posterior distribution $\hat{\rho}$ over the functions in a function class $\mathcal{F}$, a Gibbs model $G_{\hat{\rho}}$ makes its prediction for every example by randomly sampling a function $f \in \mathcal{F}$. A Bayesian model $B_{\hat{\rho}}$ makes its prediction by averaging all the functions in $\mathcal{F}$ with respect to $\hat{\rho}$. We define the expected risks

$$R(G_{\hat{\rho}}) = \mathbb{E}_{f\sim\hat{\rho}}\left\{\mathbb{E}\left[(Y - f(X))^2\right]\right\} \text{ and } R(B_{\hat{\rho}}) = \mathbb{E}\left\{\left(Y - \mathbb{E}_{f\sim\hat{\rho}}[f(X)]\right)^2\right\}.$$

The following is the PAC-Bayes theorem in supervised learning (Alquier et al., 2016; Germain et al., 2016; Pentina and Lampert, 2014) that we will apply in our proofs.

Theorem 7. *Given a function class $\mathcal{F}$ and a prior distribution π on $\mathcal{F}$. For any $\delta \in (0, 1]$, $\kappa > 0$ and posterior distribution $\hat{\rho}$ on $\mathcal{F}$, with probability at least $1 - \delta$,*

$$R(G_{\hat{\rho}}) \leq R_N(G_{\hat{\rho}}) + \left\{\mathrm{KL}(\hat{\rho}\|\pi) + \ln(1/\delta) + N \ln \mathbb{E}_{f\sim\pi}\left[\mathbb{E}\left(e^{\frac{\kappa}{N}\bar{R}_f(X,Y)}\right)\right]\right\}/\kappa \tag{4.10}$$

where $R_N(G_{\hat{\rho}}) = \mathbb{E}_{f\sim\hat{\rho}}\left\{\frac{1}{N}\sum_{i=1}^{N}[Y - f(X_i)]^2\right\}$ and

$$\bar{R}_f(X, Y) = [Y - f(X)]^2 - \mathbb{E}\left\{[Y - f(X)]^2\right\}.$$

By Jensen's inequality, $R(B_{\hat{\rho}}) \leq R(G_{\hat{\rho}})$. An upper bound for $R(G_{\hat{\rho}})$ is also an upper bound for $R(B_{\hat{\rho}})$. To make the upper bound in (4.10) fully empirical, we need to upper bound the moment generating function $\mathbb{E}\left(e^{\frac{\kappa}{N}\bar{R}_f(X,Y)}\right)$ for the mean-zero random variable $\bar{R}_f(X, Y)$, e.g., by applying Hoeffding's lemma if every $f \in \mathcal{F}$ is bounded.

To extend the PAC-Bayes theory to our problem, for $a \in \{0, 1\}$ and $c = a, 1 - a$, we define the expected factual risks ($c = a$) and counterfactual risks ($c = 1 - a$):

$$R_c(B_{\hat{\rho}_a}) = \mathbb{E}\left\{(Y - \hat{\mu}_a(X))^2 \mid A = c\right\}, \tag{4.11}$$

and

$$R_c(G_{\hat{\rho}_a}) = \mathbb{E}_{f\sim\hat{\rho}_a}\left\{\mathbb{E}\left[(Y - f(X))^2 \mid A = c\right]\right\}. \tag{4.12}$$

4.8.2 Proof of Theorem 5

Proof. By Lemma 9, we have an upper bound of ϵ_{PEHE} in terms of $R_c(B_{\hat{\rho}_a})$, $a, c \in \{0, 1\}$:

$$\epsilon_{\text{PEHE}} \leq 2 \sum_{a=0}^{1} \left[R_a(B_{\hat{\rho}_a}) + R_{1-a}(B_{\hat{\rho}_a}) \right] \leq 2 \sum_{a=0}^{1} \left[R_a(G_{\hat{\rho}_a}) + R_{1-a}(G_{\hat{\rho}_a}) \right]. \tag{4.13}$$

Combing the PAC-Bayes bounds for the expected factual risk $R_a(G_{\hat{\rho}_a})$ in Lemma 10 and the expected counterfactual risk $R_{1-a}(G_{\hat{\rho}_a})$ in Lemma 11, we have

$$\begin{aligned}
& R_a(G_{\hat{\rho}_a}) + R_{1-a}(G_{\hat{\rho}_a}) \\
\leq\ & (D_{a,\infty} + 1) L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) + (D_{a,\infty}/2 + 1) V(\boldsymbol{X}_a; \hat{\rho}_a) + \frac{1}{2} \hat{V}_a(\boldsymbol{X}_{1-a}; \hat{\rho}_a) \\
& + \left(\frac{1/4}{\sqrt{N_{1-a}}} + \frac{D_{a,\infty}}{\sqrt{N_a}} \right) (2\,\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(2/\delta) + \xi_2) \\
& + \frac{1}{\sqrt{N_a}} \left[\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(2/\delta) + \xi_1 \right] \\
\leq\ & (D_{a,\infty} + 1) L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) + (D_{a,\infty}/2 + 1) V(\boldsymbol{X}_a; \hat{\rho}_a) + \frac{1}{2} V(\boldsymbol{X}_{1-a}; \hat{\rho}_a) \\
& + \left(\frac{1/4}{\sqrt{N_{1-a}}} + \frac{D_{a,\infty} + 1}{\sqrt{N_a}} \right) (2\,\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(2/\delta) + 2C)
\end{aligned}$$

where $D_{a,\infty} = \sup_{x \in \mathcal{X}} \left[p_{X|A}(x \mid 1-a) / p_{X|A}(x \mid a) \right]$, the second inequality is achieved by $C = \max(\xi_1, \xi_2)$ and the fact that the Kullback-Leibler divergence is non-ngeative. Substituting the upper bound of $R_a(G_{\hat{\rho}_a}) + R_{1-a}(G_{\hat{\rho}_a})$ into (4.13) with $C_a = D_{a,\infty} + 1$, we otain (4.3). □

4.8.3 Proof of Theorem 6

Proof. By Jensen's inequality and Lemma 12,

$$R_{1-a}(B_{\hat{\rho}_a}) \leq R_{1-a}(G_{\hat{\rho}_a}) \leq \frac{1}{4} D_{1-a}(G_{\hat{\rho}_a}) + D_{a,\infty} L_a(G_{\hat{\rho}_a}). \tag{4.14}$$

where $R_{1-a}(B_{\hat{\rho}_a})$ and $R_{1-a}(G_{\hat{\rho}_a})$ are defined in (4.11) and (4.12), respectively. For the posterior distribution $\hat{\rho}_a$ in (4.2),

$$\begin{aligned}
D_{1-a}(G_{\hat{\rho}_a}) &= \mathbb{E}_{(f_1, f_2) \sim \hat{\rho}_a^2} \left\{ \mathbb{E} \left[\left(f_1(X) - f_2(X) \right)^2 \,\middle|\, A = 1 - a \right] \right\} \\
&= 2\mathbb{E}\{\hat{\sigma}_a^2(X) + \mu_a^2(X) \mid A = 1 - a\} - 2\mathbb{E}\{\hat{\mu}_a^2(X) \mid A = 1 - a\} \\
&= 2\mathbb{E}\{\hat{\sigma}_a^2(X) \mid A = 1 - a\} \\
&= 2V_{1-a}(\hat{\rho}_a)
\end{aligned}$$

and

$$\begin{aligned} L_a(G_{\hat{\rho}_a}) &= \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2}\left\{\mathbb{E}\left[\left(\frac{f_1(X)+f_2(X)}{2}-Y\right)^2 \;\middle|\; A=a\right]\right\} \\ &= \mathbb{E}\left[\frac{2\hat{\sigma}_a^2(X)+2\hat{\mu}_a^2(X)+2\hat{\mu}_a^2(X)}{4}-2\hat{\mu}_a(X)Y+Y^2 \;\middle|\; A=a\right] \\ &= \mathbb{E}\left[\left(Y-\hat{\mu}_a(X)\right)^2 \mid A=a\right]+\frac{1}{2}\mathbb{E}\left[\hat{\sigma}_a^2(X)\mid A=a\right] \\ &= L_a(\hat{\rho}_a)+\frac{1}{2}V_a(\hat{\rho}_a) \end{aligned}$$

Substituting the expressions of $D_{1-a}(G_{\hat{\rho}_a})$ and $L_a(G_{\hat{\rho}_a})$ into (4.14), we obtain (4.4). □

4.9 Supporting lemmas

4.9.1 Proof of Lemma 9

The expected Precision in Estimation of Heterogeneous Effects, PEHE, can be bounded using the expected factual risk $R_a(B_{\hat{\rho}_a})$ and the expected counterfactual risk $R_{1-a}(B_{\hat{\rho}_a})$.

Lemma 9. *For the Bayesian models $B_{\hat{\rho}_a}$ defined by $\hat{\rho}_a$ in (4.2), $a\in\{0,1\}$, it holds that*

$$\epsilon_{\text{PEHE}} \leq 2\sum_{a=0}^{1}\left[R_a(B_{\hat{\rho}_a})+R_{1-a}(B_{\hat{\rho}_a})\right] \tag{4.15}$$

Proof.

$$\begin{aligned} \epsilon_{\text{PEHE}} &= \int\left(\hat{\tau}(x)-\tau(x)\right)^2 p_X(x)dx \\ &= \int\left[\left(\hat{\mu}_1(x)-\hat{\mu}_0(x)\right)-\left(\mu_1(x)-\mu_0(x)\right)\right]^2 p_X(x)dx \\ &= \int\left[\left(\hat{\mu}_1(x)-\mu_1(x)\right)+\left(\hat{\mu}_0(x)-\mu_0(x)\right)\right]^2 p_X(x)dx \\ &\leq 2\sum_{a=0}^{1}\int\left(\hat{\mu}_a(x)-\mu_a(x)\right)^2 p_X(x)dx \\ &= 2\sum_{a=0}^{1}\sum_{a'=0}^{1}\int\left(\hat{\mu}_a(x)-\mu_a(x)\right)^2 p_{X,A}(x,a')dx \\ &= 2\sum_{a=0}^{1}\int\left(\hat{\mu}_a(x)-\mu_a(x)\right)^2 p_{X,A}(x,a)dx+2\sum_{a=0}^{1}\int\left(\hat{\mu}_a(x)-\mu_a(x)\right)^2 p_{X,A}(x,1-a)dx \\ &\leq 2\sum_{t=0}^{1}\left[R_a(B_{\hat{\rho}_a})+R_{1-a}(B_{\hat{\rho}_a})-\beta_a^{-1}-\beta_{1-a}^{-1}\right] \end{aligned}$$

$$\leq 2\sum_{t=0}^{1}\left[R_a(B_{\hat{\rho}_a}) + R_{1-a}(B_{\hat{\rho}_a})\right]$$

The second last equality is achieved as follows.

$$\begin{aligned}
R_c(B_{\hat{\rho}_a}) &= \int \left(\hat{\mu}_a(x) - y\right)^2 p_{X,Y(a)|A}(x, y \mid c)dxdy \\
&= \int \left(\hat{\mu}_a(x) - \mu_a(x) - \epsilon_a\right)^2 \mathcal{N}(\epsilon_a; 0, \beta_a^{-1}) p_{X|A}(x \mid c)dxd\epsilon_a \\
&= \int \Big(\left(\hat{\mu}_a(x) - \mu_a(x)\right)^2 - 2\epsilon_a(\hat{\mu}_a(x) - \mu_a(x)) + \epsilon_a^2\Big)\mathcal{N}(\epsilon_a; 0, \beta_a^{-1}) p_{X|A}(x \mid c)dxd\epsilon_a \\
&= \int \left(\mu_a(x) - \mu_a(x)\right)^2 p_{X|A}(x \mid c)dx + \beta_a^{-1}.
\end{aligned}$$

which implies that $\int \left(\mu_a(x) - \mu_a(x)\right)^2 p_{X|A}(x \mid c)dx = R_c(B_{\hat{\rho}_a}) - \beta_a^{-1}$. □

4.9.2 Proof of Lemma 10

Lemma 10. *Under Assumption 10, for any $\delta \in (0, 1]$ and π_a in (4.1) and $\hat{\rho}_a$ in (4.2), with probability at least $1 - \delta$,*

$$R_a(G_{\hat{\rho}_a}) \leq L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) + V(\boldsymbol{X}_a; \hat{\rho}_a) + \left[\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(1/\delta) + \xi_1\right]/\sqrt{N_a}, \tag{4.16}$$

where ξ_1 is a universal constant.

Proof. Factual risk minimization is essentially a supervised learning problem where we have access to the outcome information. Under Assumption 10, the random variable $\bar{R}_{a,f}(X, Y) := [Y - f(X)]^2 - \mathbb{E}\left\{[Y - f(X)]^2 \mid A = a\right\}$ is bounded. Suppose that $\bar{R}_{a,f}(X, Y) \in [b_1, b_2]$ for some constants $b_1, b_2 > 0$. By Hoeffding's lemma, the moment generating function of the mean-zero $\bar{R}_{a,f}(X, Y)$,

$$\mathbb{E}\left[e^{\frac{\kappa}{N}\bar{R}_{a,f}(X,Y)}\right] \leq \exp\left[\frac{\kappa^2(b_2 - b_1)^2}{8N_a^2}\right].$$

Then applying Theorem 7, we have

$$\begin{aligned}
R_a(G_{\hat{\rho}_a}) &\leq R_{a,N}(G_{\hat{\rho}_a}) + \left[\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(1/\delta) + \frac{\kappa^2(b_2 - b_1)^2}{8N_a}\right]/\kappa \\
&= R_{a,N}(G_{\hat{\rho}_a}) + \left[\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(1/\delta) + \xi_1\right]/\sqrt{N_a},
\end{aligned}$$

where $\kappa = \sqrt{N_a}$ and $\xi_1 = (b_2 - b_1)^2/8$ and

$$\begin{aligned} R_{a,N}(G_{\hat{\rho}_a}) &= \mathbb{E}_{f\sim\hat{\rho}}\left\{\frac{1}{N_a}\sum_{i=1}^{N} 1\{A_i = a\}\left[Y - f(X_i)\right]^2\right\} \\ &= \frac{1}{N_a}\sum_{i=1}^{N} 1\{A_i = a\}\Big[\left(Y_i - \hat{\mu}_a(X_i)\right)^2 + \hat{\sigma}_a^2(X_i)\Big] \\ &= L(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) + V(\boldsymbol{X}_a; \hat{\rho}_a). \end{aligned}$$

□

4.9.3 Proof of Lemma 11

Unlike the factual risk $R_a(G_{\hat{\rho}_a})$ in Lemma 10, the PAC-Bayes bound for supervised learning is not directly applicable to the counterfactual risk $R_{1-a}(G_{\hat{\rho}_a})$. This issue is resolved by importance weighting (Cortes et al., 2010; Germain et al., 2013).

Lemma 11. *Under Assumptions 7 and 10, for any $\delta \in (0, 1]$, π_a in (4.1) and $\hat{\rho}_a$ in (4.2), with probability at least $1 - \delta$, we have*

$$\begin{aligned} R_{1-a}(G_{\hat{\rho}_a}) \leq \frac{1}{2}V_a(\boldsymbol{X}_{1-a}; \hat{\rho}_a) + D_{a,\infty}\left[L_a(\boldsymbol{X}_a, \boldsymbol{Y}_a; \hat{\rho}_a) + \frac{1}{2}V_a(\boldsymbol{X}_a; \hat{\rho}_a)\right] \\ + \left(\frac{1/4}{\sqrt{N_{1-a}}} + \frac{D_{a,\infty}}{\sqrt{N_a}}\right)\left(2\,\mathrm{KL}(\hat{\rho}_a\|\pi_a) + \ln(1/\delta) + \xi_2\right), \end{aligned} \tag{4.17}$$

where $D_{a,\infty} = \sup_{x\in\mathcal{X}}\left[p_{X|A}(x \mid 1-a)/p_{X|A}(x \mid a)\right]$ and ξ_2 is a universal constant.

Proof. In the upper bound of $R_{1-a}(G_{\hat{\rho}_a})$ from Lemma 12, $D_{1-a}(G_{\hat{\rho}_a})$ does not depend on any outcome and $L_a(G_{\hat{\rho}_a})$ is the expected factual risk. Then we can bound both terms using Theorem 7. First, we define the empirical version of $D_{1-a}(G_{\hat{\rho}_a})$ and $L_a(G_{\hat{\rho}_a})$,

$$\begin{aligned} D_{1-a,N}(G_{\hat{\rho}_a}) &= \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2}\left\{N_{1-a}^{-1}\sum_{i=1}^{N} 1\{A_i = 1-a\}\left[f_1(X_i) - f_2(X_i)\right]^2\right\} \\ &= \frac{1}{N_{1-a}}\sum_{i=1}^{N} 1\{A_i = 1-a\}\left[2\hat{\sigma}_a^2(X_i) + 2\hat{\mu}_a^2(X_i) - 2\hat{\mu}_a^2(X_i)\right] \\ &= 2V_a(\boldsymbol{X}_{1-a}; \hat{\rho}_a) \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} L_{a,N}(G_{\hat{\rho}_a}) &= \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2}\left\{\frac{1}{N_a}\sum_{i=1}^{N}1\{A_i=a\}\left[\frac{f_1(X_i)+f_2(X_i)}{2}-Y_i\right]^2\right\} \\ &= \frac{1}{N_a}\sum_{i=1}^{N}1\{A_i=a\}\left[\hat{\mu}_a^2(X_i)+\frac{1}{2}\hat{\sigma}_a^2(X_i)-2\hat{\mu}_a(X_i)Y_i+Y_i^2\right] \\ &= \frac{1}{N_a}\sum_{i=1}^{N}1\{A_i=a\}\left[(Y_i-\hat{\mu}_a(X_i))^2+\frac{1}{2}\hat{\sigma}_a^2(X_i)\right] \\ &= L_a(\boldsymbol{X}_a,\boldsymbol{Y}_a;\hat{\rho}_a)+\frac{1}{2}V_a(\boldsymbol{X}_a;\hat{\rho}_a) \end{aligned} \tag{4.19}$$

Under Assumption 10 the random variable

$$\bar{D}_{1-a,f_1,f_2}(X,Y) := [f_1(X)-f_2(X)]^2 - \mathbb{E}\left\{[f_1(X)-f_2(X)]^2 \mid A=a\right\}$$

is bounded, and that $\bar{D}_{1-a,f_1,f_2}(X,Y)\in[b_3,b_4]$ for some constants $b_3,b_4>0$. By Hoeffding's lemma, the moment generating function of the mean-zero $\bar{D}_{1-a,f_1,f_2}(X,Y)$,

$$\mathbb{E}\left[e^{\frac{\kappa}{N_{1-a}}\bar{D}_{1-a,f_1,f_2}(X,Y)}\right] \le \exp\left[\frac{\kappa^2(b_4-b_3)^2}{8N_{1-a}^2}\right].$$

Then applying Theorem 7 with $\kappa=\sqrt{N_{1-a}}$, we have

$$D_{1-a}(G_{\hat{\rho}_a}) \le D_{1-a,N}(G_{\hat{\rho}_a}) + \left[2\,\mathrm{KL}(\hat{\rho}_a\|\pi_a)+\ln(1/\delta)+(b_4-b_3)^2/8\right]/\sqrt{N_{1-a}}, \tag{4.20}$$

where the double KL divergence is from the fact that

$$\mathrm{KL}(\hat{\rho}^2\|\pi^2) = \int\hat{\rho}(f_1)\log\frac{\hat{\rho}(f_1)}{\pi(f_1)}df_1 + \int\hat{\rho}(f_2)\log\frac{\hat{\rho}(f_2)}{\pi(f_2)}df_2 = 2\,\mathrm{KL}(\hat{\rho}\|\pi).$$

Similarly, suppose that

$$\bar{L}_{a,f_1,f_2}(X,Y) := \left(\frac{f_1(X)+f_2(X)}{2}-Y\right)^2 - \mathbb{E}\left\{\left(\frac{f_1(X)+f_2(X)}{2}-Y\right)^2 \;\middle|\; A=a\right\} \in [b_5,b_6],$$

for some constants $b_5,b_6>0$. By Hoeffding's lemma,

$$\mathbb{E}\left[e^{\frac{\kappa}{N_a}\bar{L}_{a,f_1,f_2}(X,Y)}\right] \le \exp\left[\frac{\kappa^2(b_6-b_5)^2}{8N_{1-a}^2}\right].$$

Then applying Theorem 7 with $\kappa=\sqrt{N_a}$, we have

$$L_a(G_{\hat{\rho}_a}) \le L_{a,N}(G_{\hat{\rho}_a}) + \left[2\,\mathrm{KL}(\hat{\rho}_a\|\pi_a)+\ln(1/\delta)+(b_6-b_5)^2/8\right]/\sqrt{N_a} \tag{4.21}$$

Substituting (4.20) and (4.21) into the upper bound in Lemma 12, we obtain

$$R_{1-a}(G_{\hat{\rho}_a}) \leq \frac{1}{4} D_{1-a,N}(G_{\hat{\rho}_a}) + D_{a,\infty} L_{a,N}(G_{\hat{\rho}_a}) + \left(\frac{1/4}{\sqrt{N_{1-a}}} + \frac{D_{a,\infty}}{\sqrt{N_a}} \right) (2\,\mathrm{KL}(\hat{\rho}_a \| \pi_a) + \ln(1/\delta) + \xi_2),$$

with $\xi_2 = \max\{(b_4 - b_5)^2, (b_6 - b_5)^2\}/8$. Substituting the expression of $D_{1-a,N}(G_{\hat{\rho}_a})$ in (4.18) and the expression of $L_{a,N}(G_{\hat{\rho}_a})$ (4.19) into the last equation, we obtain (4.17). □

4.9.4 Proof of Lemma 12

Lemma 12. *Under Assumptions 7 and 10, it holds that*

$$R_{1-a}(G_{\hat{\rho}_a}) \leq \frac{1}{4} D_{1-a}(G_{\hat{\rho}_a}) + D_{a,\infty} L_a(G_{\hat{\rho}_a}) \tag{4.22}$$

where $D_{a,\infty} = \sup_{x \in \mathcal{X}} \left[p_{X|A}(x \mid 1-a) / p_{X|A}(x \mid a) \right]$,

$$D_{1-a}(G_{\hat{\rho}_a}) = \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2} \left\{ \mathbb{E}\left[(f_1(X) - f_2(X))^2 \;\middle|\; A = 1-a \right] \right\}, \text{ and}$$

$$L_a(G_{\hat{\rho}_a}) = \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2} \left\{ \mathbb{E}\left[\left(\frac{f_1(X) + f_2(X)}{2} - Y \right)^2 \;\middle|\; A = a \right] \right\}.$$

Proof. We rewrite $R_{1-a}(G_{\hat{\rho}_a})$ as

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{\rho}_a} \left\{ \mathbb{E}\left[(f(x) - Y)^2 \mid A = 1-a \right] \right\} \\
&= \frac{1}{2} \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2} \left\{ \mathbb{E}\left[(f_1(X) - Y)^2 + (f_2(X) - Y)^2 \mid A = 1-a \right] \right\} \\
&= \frac{1}{2} \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2} \left\{ \mathbb{E}\left[f_1^2(X) + f_2^2(X) - 2Y f_1(X) - 2Y f_2(X) + 2Y^2 \mid A = 1-a \right] \right\} \\
&= \frac{1}{2} \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2} \left\{ \mathbb{E}\left[\frac{1}{2}\Big(f_1^2(X) + f_2^2(X) - 2 f_1(X) f_2(X) \Big) + \frac{1}{2}\Big(f_1^2(X) + f_2^2(X) \right.\right. \\
&\quad \left.\left. + 2 f_1(X) f_2(X) - 4Y(f_1(X) + f_2(X)) + 4Y^2 \Big) \;\middle|\; A = 1-a \right] \right\} \\
&= \frac{1}{2} \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2} \left\{ \mathbb{E}\left[\frac{1}{2} (f_1(X) - f_2(X))^2 + 2 \left(\frac{f_1(X) + f_2(X)}{2} - Y \right)^2 \;\middle|\; A = 1-a \right] \right\} \\
&= \frac{1}{4} D_{1-a}(G_{\hat{\rho}_a}) + L_{1-a}(G_{\hat{\rho}_a}).
\end{aligned}$$

where

$$
\begin{aligned}
L_{1-a}(G_{\hat{\rho}_a}) &= \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2}\left\{\mathbb{E}\left[\left(\frac{f_1(X)+f_2(X)}{2}-Y\right)^2 \;\middle|\; A=1-a\right]\right\} \\
&= \mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2}\left\{\int \frac{p_{X|A}(x\mid 1-a)}{p_{X|A}(x\mid a)} p_{X|A}(x\mid a) l_{f_1,f_2}(x,y)dx\right\} \\
&\leq \left\{\int \left(\frac{p_{X|A}(x\mid 1-a)}{p_{X|A}(x\mid a)}\right)^{\alpha} dx\right\}^{\frac{1}{\alpha}} \cdot \left\{\mathbb{E}_{(f_1,f_2)\sim\hat{\rho}_a^2}\left[\mathbb{E}\left(l_{f_1,f_2}^{\frac{\alpha}{\alpha-1}}(X,Y)\;\middle|\; A=a\right)\right]\right\}^{\frac{\alpha-1}{\alpha}}
\end{aligned}
$$

where $l_{f_1,f_2}(x,y) = ([f_1(x)+f_2(x)]/2-y)^2$. The inequality is attained by applying Hölder's inequality and the positivity assumption. Taking $\alpha\to\infty$, we obtain (4.22). □

4.10 Further experimental details

Following the hyperparameter optimization method in (Shalit et al., 2017), we choose the hyperparameters using a random search over the hyperparameter space in Table 4.4. The missing counterfactual outcomes in the PEHE loss are approximated by the observed outcome of the nearest neighbour in the opposite group.

Table 4.4 Hyperparameters and ranges of DKLITE

Hyperparameters	Range
Variance regularization parameter α_1	$\{0.001, 0.01, 0.1, 1, 10, 25, 50, 75, 100\}$
Reconstruction regularization parameter α_2	$\{0.001, 0.01, 0.1, 1, 10, 25, 50, 75, 100\}$
Number of hidden layers	$\{1, 2, 3\}$
Number of hidden units	$\{50, 100, 150, 200\}$
Dimension of the feature map	$\{25, 50, 75, 100\}$
Regression Form	$\{\text{Primal}, \text{Dual}\}$

IHDP. *Counterfactual outcomes are randomly generated via a predefined probabilistic model* (Alaa and van der Schaar, 2017; Hill, 2011; Shalit et al., 2017; Yao et al., 2018). The objective is to estimate the effects of specialist home visits to individuals on their future cognitive test scores. Patient covariates X were collected from the actual randomized experiment but the overall cohort was made artificially imbalanced by removing a subset of the treated population. The dataset comprises 747 units (139 treated, 608 control) and 25 covariates measuring aspects of children and their mothers. Outcomes $Y(0)$ and $Y(1)$ are obtained by implementing the setting "A" in the NPCI package (Dorie, 2016).

Twins. *Outcomes are observed but the treatment assignment is simulated.* The objective is to predict the mortality of each of one of two twins in their first year. We consider the treated twin to be the one with higher weight at birth and, since we have records for both twins, we treat their outcomes as two potential outcomes, i.e. $Y(1)$ and $Y(0)$. Now, in order to simulate an observational study, we need to select one of the two twins for inclusion in our data, that is defining $\mathbb{P}(A \mid X)$. We do so by sampling from $A \mid X \sim \text{Bern}(\text{Sigmoid}(W^\top X + \epsilon))$ where $W^\top \sim \text{Uniform}((-0.1, 0.1)^{30\times 1})$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. The final data contains 11400 individuals with 30 measured covariates relating to their parents, pregnancy and birth.

For the IHDP dataset, we average over 1000 realizations of the outcomes with 63/27/10% train/validation/test split. For the Twins dataset, each dataset is divided 56/24/20% into training/validation/testing sets, and we report the results averaged over 100 realizations.

Chapter 5

Robust recursive partitioning: from conditional average treatment effects to interpretable subgroups

Subgroup analysis of treatment effects plays an important role in applications from medicine to public policy to recommender systems. It allows physicians to identify groups of patients for whom a given drug or treatment is likely to be effective and groups of patients for which it is not. In this chapter, we propose a new method called R2P, which can turn black-box conditional average treatment effect (CATE) estimates into interpretable subgroups. In R2P, we quantify the errors in the CATE estimates by a distribution-free technique called conformal prediction and make use of the quantified uncertainties to recursively partition the covariates space for identifying subgroups. Experiments using synthetic and semi-synthetic datasets demonstrate that R2P can construct partitions in which individual treatment effects are more homogeneous within subgroups and more heterogeneous across subgroups, compared with the partitions produced by various baseline methods. Moreover, leveraging the predictive power of black-box machine learning models, R2P produces narrower valid confidence intervals for CATEs than the model-specific baseline methods.

5.1 Introduction

The understanding of treatment effects plays an important role in shaping interventions and treatments in areas from clinical trials (Rothwell, 2005; Zhou et al., 2017) to recommender systems (Lada et al., 2019) to public policy (Grimmer et al., 2017). In many settings, the

relevant population is diverse, and different parts of the population display different reactions to treatment. In such settings, *heterogeneous treatment effect (HTE) analysis*, also called *subgroup analysis*, is used to find subgroups consisting of subjects who have similar covariates and display similar treatment outcomes (Foster et al., 2011; Imai and Ratkovic, 2013). The identification of subgroups is informative of itself; it also improves the interpretation of treatment effects across the entire population and makes it possible to develop more effective interventions and treatments and to improve the design of further experiments. In a clinical trial, for example, HTE analysis can identify subgroups of the population for which the studied treatment is effective, even when it is found to be ineffective for the population in general (Hobbs et al., 2011).

To identify subjects who have similar covariates and display similar treatment outcomes, it is necessary to create reliable estimates of the treatment effects conditional on individual covariates, e.g., *conditional average treatment effects* (CATEs). The existing works on subgroup analysis (Athey and Imbens, 2016; Johansson et al., 2018b; Su et al., 2009; Tran and Zheleva, 2019) proceed by estimating CATEs with a specific machine learning model (e.g. decision tree) and recursively partitioning the subject population for subgroup identification. In these methods, the criteria for partitioning maximize the heterogeneity of treatment effects *across* subgroups, using a sample mean estimator. In particular, the population (or any previously identified subgroups) would be partitioned into smaller subgroups provided that the sample means of these subgroups are sufficiently different. These methods focus on inter-group heterogeneity with less attention to intra-group homogeneity.

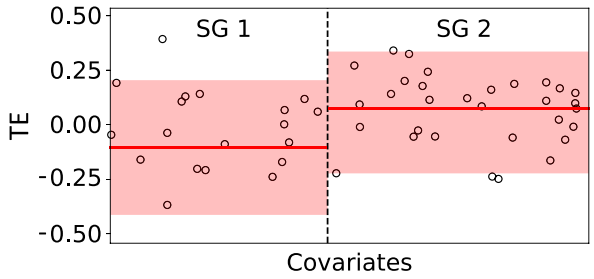


Figure 5.1 Two subgroups identified by the method (Tran and Zheleva, 2019). The solid red line shows the CATE estimates and 95% confidence intervals filled in red.

An important problem with this approach is that, because it overly relies on inter-group heterogeneity based on sample means, it may lead to *false discovery*. To illustrate, consider the toy example depicted in Figure 5.1. In this example, individual treatment effects (the dots in the figure) are generated by i.i.d random draws from a Gaussian distribution with mean 0 and standard deviation 0.1. In truth, the treatment under consideration is in fact *totally ineffective and innocuous*; on average, it has *no effect at all* and the treatment effects

are entirely uncorrelated with the single covariate (shown on the horizontal axis). However, if the observed data – the realization of the random draws – happens to be the one shown in Figure 5.1, standard methods will typically partition the population as shown in the figure, thereby "discovering" a segment of the population for whom the treatment is effective and a complementary segment where the treatment is dangerous. Obviously, decisions based on such a false discovery are not useful. Note that this false discovery occurs because, although the outcome variations between the two groups are indeed substantially different, the outcome variations within each group are just as different – but the latter variation is entirely ignored in the creation of subgroups. We propose a robust recursive partitioning (R2P) method that can avoid false discoveries. R2P has several distinctive characteristics summarized below.

R2P discovers interpretable subgroups in a way that is not tied to any *particular* CATE estimator. This is in sharp contrast with previous methods (Athey and Imbens, 2016; Johansson et al., 2018b; Seibold et al., 2016; Su et al., 2009; Tran and Zheleva, 2019), each of which relies on a *specific* CATE estimator. R2P can leverage *any* black-box model for subgroup analysis, e.g. an CATE estimator based on random forest (Wager and Athey, 2018), or on multi-task Gaussian processes (Alaa and van der Schaar, 2017) or on deep neural networks (Shalit et al., 2017; Yoon et al., 2018; Zhang et al., 2020). Many of these CATE estimators are based on non-interpretable black-box models. R2P divides individuals (units) into subgroups with respect to a tree structure which is much easier to interpret than black boxes. Leveraging an uncertainty quantification method called split conformal prediction (Lei et al., 2018b), R2P employs a novel criterion we call *confident homogeneity* to create partitions that take into account both heterogeneity across groups and homogeneity within groups. Extensive experiments using synthetic and semi-synthetic datasets demonstrate that R2P outperforms baseline methods, by more robustly identifying subgroups while providing much narrower valid confidence intervals for CATEs.

5.2 Robust recursive partitioning

To highlight the core design principles, we begin by introducing robust recursive partitioning (R2P) in the regression setting; we extend to the more complicated CATE estimation setting in the next section. We consider a standard regression problem with a d-dimensional covariate space $\mathcal{X} \subseteq \mathbb{R}^d$ and an outcome space $\mathcal{Y} \subseteq \mathbb{R}$. We are given a dataset of N samples, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$. We assume the samples are independently drawn from an unknown distribution $\mathbb{P}(X, Y)$ defined on $\mathcal{X} \times \mathcal{Y}$. We are interested in estimating the conditional mean $\mu(x) = \mathbb{E}[Y|X = x]$ for $x \in \mathcal{X}$. We denote the estimator by $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$; $\hat{\mu}$ predicts an outcome $\hat{Y}_{N+1} = \hat{\mu}(X_{N+1})$ for a new testing sample X_{N+1}. To quantify the error in this prediction, we apply the split conformal prediction (SCP) algorithm (Lei et al., 2018b) (also

called inductive conformal prediction in (Papadopoulos et al., 2002; Vovk et al., 2005)) to construct a prediction interval $\hat{C}$ for Y_{N+1} with marginal coverage guarantee in finite samples.

In SCP, we take as given a *miscoverage rate* $\alpha \in (0,1)$. We split the samples in $\mathcal{D}$ into a training set $\mathcal{I}_1$ and a validation set $\mathcal{I}_2$. We let the two sets have the same sample size. We obtain the outcome estimator $\hat{\mu}^{\mathcal{I}_1}$ using $\mathcal{I}_1$ and compute the absolute residual of $\hat{\mu}^{\mathcal{I}_1}$ on each validation sample in $\mathcal{I}_2$. On a new sample X_{N+1}, the prediction interval of SCP is given by

$$\begin{aligned}\hat{C}(X_{N+1}) &= \left[\hat{\mu}^{\text{lo}}(X_{N+1}), \hat{\mu}^{\text{up}}(X_{N+1})\right] \\ &= \left[\hat{\mu}^{\mathcal{I}_1}(X_{N+1}) - \hat{Q}^{\mathcal{I}_2}_{1-\alpha}, \hat{\mu}^{\mathcal{I}_1}(X_{N+1}) + \hat{Q}^{\mathcal{I}_2}_{1-\alpha}\right],\end{aligned} \tag{5.1}$$

where $\hat{Q}^{\mathcal{I}_2}_{1-\alpha}$ is defined as the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th quantile of the residual set $\{|Y_i - \hat{\mu}^{\mathcal{I}_1}(X_i)| : i \in \mathcal{I}_2\}$, i.e., the $\lfloor(|\mathcal{I}_2|+1)\alpha\rfloor$-th largest absolute residual. Assuming that the $N+1$ samples are drawn exchangeably from $\mathbb{P}(X,Y)$, the prediction interval (5.1) satisfies the following marginal coverage guarantee,

$$\mathbb{P}[Y_{N+1} \in \hat{C}(X_{N+1})] \geq 1-\alpha. \tag{5.2}$$

To understand why (5.2) holds, we consider the following example.

Example 5. Let $Z_1, \ldots, Z_M, Z_{M+1}$ be random variables exchangeably drawn from $\mathbb{P}(Z)$ and $Z_{(1)}, \ldots, Z_{(M+1)}$ denote the order statistics of $Z_1, \ldots, Z_{M+1}$. We define the $(1-\alpha)$-th quantile of $Z_{(1)}, \ldots, Z_{(M+1)}$ as

$$\hat{Q}_{1-\alpha} = \begin{cases} Z_{(\lceil (M+1)(1-\alpha)\rceil}, & \text{if } \lceil (M+1)(1-\alpha)\rceil \leq M \\ \infty, & \text{otherwise} \end{cases}$$

The key observation is that the rank of Z_{M+1} among $Z_1, \ldots, Z_{M+1}$ is uniformly distributed over the set $\{1, ..., M+1\}$ under the exchangeability assumption that the joint distribution of $Z_1, \ldots, Z_{M+1}$ is invariant of the sampling order of $Z_1, \ldots, Z_{M+1}$. Thus, for any given miscoverage level $\alpha \in (0,1)$, we have

$$\mathbb{P}[Z_{M+1} \leq \hat{Q}_{1-\alpha}] \leq 1-\alpha, \tag{5.3}$$

by summing the uniform distribution up to $\hat{Q}_{1-\alpha}$. Suppose that $M = N/2$, and $Z_i = |Y_i - \hat{\mu}^{\mathcal{I}_1}(X_i)|$ for $i \in \mathcal{I}_2$ and $Z_{M+1} = |Y_{N+1} - \hat{\mu}^{\mathcal{I}_1}(X_{N+1})|$. The coverage guarantee (5.2) is essentially a result from (5.3) and the construction of $\hat{C}$ in (5.1).

Marginal coverage vs. conditional coverage. To illustrate the guarantee implied by (5.2), assume that we are given 1000 testing samples and set α to 0.05. SCP prescribes a prediction interval for each sample in a way that we can expect the prediction to be within

the associated prediction interval on at least 950 samples. However, this coverage guarantee is marginal over the covariate space $\mathcal{X}$. Suppose that we divide $\mathcal{X}$ into two subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ in a way that $\mathcal{X}_1$ has 800 samples and $\mathcal{X}_2$ has 200 samples, it might be the case that 790 samples in $\mathcal{X}_1$ are covered but only 160 samples in $\mathcal{X}_2$ are covered. In this case, 80% of the samples in $\mathcal{X}_2$ would be covered, which is lower than the target coverage rate 95%. Lei and Wasserman (2014); Vovk (2012) show that conformal prediction can not achieve the conditional coverage guarantee at a given x in a distribution-free setup. Nevertheless, we will see later that the intervals with marginal coverage guarantee are still useful for subgroup analysis by informing us about a region that contains 95% of the outcomes.

Prediction intervals vs. confidence intervals. Before moving on to subgroup analysis, we want to first clarify the difference between prediction intervals and confidence intervals. The prediction interval given by SCP is designed to cover the random variable Y_{n+1} with high probability. And Y_{n+1} is not a parameter in some parametric models, so $\hat{C}(X_{N+1})$ is not a confidence interval we often refer to. However, in practice, a prediction interval for Y_{N+1} can still be applied to cover $\mu(X_{N+1}) = \mathbb{E}[Y_{N+1} \mid X_{N+1}]$. This is because the mean of $Y_{N+1} \mid X_{N+1}$ often has less variability than the random variable $Y_{N+1} \mid X_{N+1}$. Of course, there are special cases. For example, the mean of a distribution can undergo a large shift if the distribution is contaminated with outliers. Then a prediction interval will fail to cover the mean with any guarantee. Some additional assumption (e.g. continuity or smoothness) on the outcome distribution is needed to establish the equivalence between the two types of intervals. However, in the literature, conformal prediction methods are mainly studied in a distribution-free manner. In what follows, we will use prediction intervals from SCP as confidence intervals for conditional means since the CATE function is defined as the difference between two conditional means. And we only verify the marginal coverage guarantee of our confidence intervals empirically.

5.2.1 Robust heterogeneity analysis

Let $\Pi = \{l_j\}$ be a partition of the covariate space $\mathcal{X}$. Let $\mathcal{D}_l = \{(X_i, Y_i) \in \mathcal{D} | X_i \in l\}$ collect the samples in the subgroup l. Let $\mathcal{I}_2^l$ be the subset of samples in $\mathcal{I}_2$ that belongs to the subgroup l. We obtain the interval $\hat{C}_l(X)$ for the subgroup l by setting the upper and lower endpoints to be

$$\hat{\mu}_l^{\text{up}}(X) = \hat{\mu}^{\mathcal{I}_1}(X) + \hat{Q}_{1-\alpha}^{\mathcal{I}_2^l} \text{ and } \hat{\mu}_l^{\text{lo}}(X) = \hat{\mu}^{\mathcal{I}_1}(X) - \hat{Q}_{1-\alpha}^{\mathcal{I}_2^l}. \tag{5.4}$$

Compared with the interval (5.1), $\hat{C}_l(X)$ better captures the local error of $\hat{\mu}^{\mathcal{I}_1}$ in the subgroup l. If we further partition the subgroup l, we can further subsample $\mathcal{I}_2^l$ to construct the intervals. There is almost no additional computational cost because all the validation

errors for $\mathcal{I}_2$ is computed in the first place. All we need to do is simply compute the quantile $\hat{Q}_{1-\alpha}$'s locally. (To avoid notational complications, omit reference to the subsets $\mathcal{I}_1$ and $\mathcal{I}_2^l$ hereafter. Throughout, we follow the convention that the confidence bound have been computed based on the split.) To estimate the center of the subgroup l, we use the average outcome $\hat{\mu}_{l,\text{mean}} = \mathbb{E}[\hat{\mu}(X) \mid X \in l]$. We define the *expected absolute deviation* of the subgroup l as $S_l = \mathbb{E}[v_l(X) \mid X \in l]$, where

$$v_l(x) = \left(\hat{\mu}_{l,\text{mean}} - \hat{\mu}_l^{\text{up}}(x)\right) \mathbb{I}\left[\hat{\mu}_{l,\text{mean}} > \hat{\mu}_l^{\text{up}}(x)\right] + \left(\hat{\mu}_l^{\text{lo}}(x) - \hat{\mu}_{l,\text{mean}}\right) \mathbb{I}\left[\hat{\mu}_{l,\text{mean}} < \hat{\mu}_l^{\text{lo}}(x)\right].$$

By definition, $\hat{\mu}_l^{\text{up}}(x)$ is larger than $\hat{\mu}_l^{\text{lo}}(x)$ as long as the quantile $\hat{Q}_{1-\alpha} > 0$. When the first indicator function is 1, i.e. the average outcome (the group center) $\hat{\mu}_{l,\text{mean}}$ is larger than the upper bound $\hat{\mu}_l^{\text{up}}(x)$ at x, we are confident that the outcome value at x is significantly smaller than the group mean. Similarly, when the second indicator function is 1, we are certain that the outcome value at x is significantly larger than the group mean. When $\hat{C}_l(x) = [\hat{\mu}_l^{\text{lo}}(x), \hat{\mu}_l^{\text{up}}(x)]$ contains $\hat{\mu}_{l,\text{mean}}$, both indicator functions are 0 and $v_l(x) = 0$.

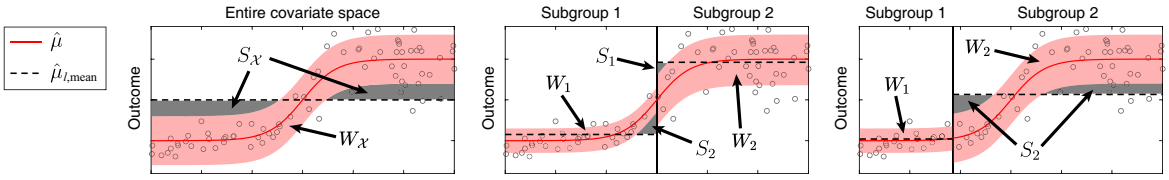


Figure 5.2 Illustration of the space partition and confident homogeneity in R2P. The regions shaded in red and grey represent W_l and S_l, respectively. Start by partitioning the covariate space $\mathcal{X}$ (the left panel). The partition with smaller impurity (the middle panel) makes the heterogeneity across subgroups and the homogeneity within subgroups stronger than others with larger impurity (e.g., the right panel).

The deviation S_l evaluates the outcome homogeneity in the subgroup l by measuring the proportion of $x \in l$ at which a confidence interval $\hat{C}_l(x)$ does not cover the group mean $\hat{\mu}_{l,\text{mean}}$. If this value is small, the outcome for the subgroup l is not significantly different from the group mean. Our criterion for partitioning is more conservative than the mean difference $|\hat{\mu}_l(x) - \hat{\mu}_{l,\text{mean}}|$, and has the potential to provide greater protection against false discoveries of subgroups, e.g., the one in Figure 5.1. From the left panel of Figure 5.2, we can see how conformal prediction plays a role in our procedure. The confidence intervals given by SCP cover the red region with marginal guarantee, which means that 95% of the outcomes are in this red region. When we initialize the partition, we treat the entire covariate space $\mathcal{X}$ as a group, but the group mean (the dashed line in the left panel) falls outside the red region.

This indicates that the outcome homogeneity in $\mathcal{X}$ is very low and a partition (e.g. the one in the middle panel) is needed to increase the outcome homogeneity within each subgroup.

However, minimizing S_l is not enough to maximize subgroup homogeneity. If the intervals $\hat{C}_l(x)$ for all $x \in l$ are very wide and contain the average outcome $\hat{\mu}_{l,\text{mean}}$, the outcome homogeneity can be very low even though $S_l = \mathbb{E}[v_l(X) \mid X \in l]$ is 0. To resolve this issue, as we partition the covariate space, we jointly minimize S_l and the expected confidence interval width $W_l = \mathbb{E}[|\hat{C}_l(X)| \mid X \in l]$. Overall, we formalize our partitioning problem as

$$\underset{\Pi}{\text{minimize}} \sum_{l \in \Pi} \lambda W_l + (1-\lambda) S_l, \tag{5.5}$$

where $\lambda \in [0,1]$ is a hyperparameter that balances the impact of W_l and S_l. We call the weighted sum, $\lambda W_l + (1-\lambda)S_l$, the impurity of the *confident homogeneity* for the subgroup l. Figure 5.2 illustrates how minimizing the impurity of the confident homogeneity improves both homogeneity within each subgroup and heterogeneity across subgroups. We can see both W_l (the area of the red region) and S_l (the area of the grey region) are minimized by the partition in the middle panel (compared with the partition in the right panel). We can also see from the middle panel that by reconstructing the intervals using the local samples $\mathcal{I}_2^l$, $l = 1, 2$, the intervals reflect the local errors better than the intervals in the left panel.

5.2.2 Confident homogeneity

We now describe our robust recursive partitioning (R2P) method for solving the mimization problem (5.5). There may be more than one partition that achieves the minimum; because a larger number of subgroups is harder to interpret, we will always choose the minimizer with the smallest number of subgroups.

We begin with the trivial partition $\Pi = \{\mathcal{X}\}$. We let Π_c denote the set of subgroups whose objectives in (5.5) can be potentially improved. In the initialization step, we set $\Pi_c = \Pi$ and apply SCP on $\mathcal{D}$ to obtain the confidence interval (5.1). With the intervals for the samples in the subgroup l, we estimate W_l and S_l by

$$\hat{W}_l = \frac{1}{N_2^l} \sum_{i \in \mathcal{I}_2^l} |\hat{C}_l(X_i)| \quad \text{and} \quad \hat{S}_l = \frac{1}{N_2^l} \sum_{i \in \mathcal{I}_2^l} v_l(X_i), \tag{5.6}$$

where $N_2^l = |\mathcal{I}_2^l|$ and $|\hat{C}_l(X_i)|$ is the width of $\hat{C}_l(X_i)$. After initialization, we recursively partition $\mathcal{X}$ by splitting the subgroups in Π_c with respect to the criterion in (5.5). To split

Algorithm 2 Robust Recursive Partitioning

Input: Samples $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$, miscoverage rate $\alpha \in (0,1)$, $\Pi = \{\mathcal{X}\}$
Initialization: $\Pi_c = \Pi$, split $\mathcal{D}$ into $\mathcal{I}_1$ and $\mathcal{I}_2$, obtain a regression model $\hat{\mu}$ using $\mathcal{I}_1$, construct the confidence interval (5.1), $\hat{W}_{\mathcal{X}}$ and $\hat{S}_{\mathcal{X}}$ using $\mathcal{I}_2$
for $l \in \Pi_c$ **do**
    Obtain $\hat{W}^*_{l^\pm}$, $\hat{S}^*_{l^\pm}$, i^*, and ϕ^*
    if $\lambda \hat{W}^*_{l^\pm} + (1-\lambda)\hat{S}^*_{l^\pm} \leq (1-\gamma)\left[\lambda \hat{W}_l + (1-\lambda)\hat{S}_l\right]$ **then**
        Partition l into $l^+(i^*, \phi^*)$ and $l^-(i^*, \phi^*)$
        $\Pi \leftarrow \Pi \cup \{l^+(i^*, \phi^*), l^-(i^*, \phi^*)\} \setminus \{l\}$
        $\Pi_c \leftarrow \Pi_c \cup \{l^+(i^*, \phi^*), l^-(i^*, \phi^*)\}$
    $\Pi_c \leftarrow \Pi_c \setminus \{l\}$
Output: Π, $\hat{\mu}$, and $\hat{C}_l$ for all $l \in \Pi$

each subgroup $l \in \Pi_c$, we first consider the two disjoint subsets from the subgroup l given by

$$l_k^+(\phi) = \{x \in l | x_k \geq \phi\} \text{ and } l_k^-(\phi) = \{x \in l | x_k < \phi\},$$

where $\phi \in (x_k^{l,\min}, x_k^{l,\max})$ is the threshold for splitting, x_k is the k-th covariate and $x_k^{l,\min}$ and $x_k^{l,\max}$ are the minimum and maximum values of the k-th covariate within the subgroup l, respectively. We split $\mathcal{I}_2^l$ into two subsets with respect to $l_k^+(\phi)$ and $l_k^-(\phi)$:

$$\mathcal{I}_2^{l_k^+(\phi)} = \left\{(X_i, Y_i) \in \mathcal{I}_2^l : X_i \in l_k^+(\phi)\right\} \text{ and } \mathcal{I}_2^{l_k^-(\phi)} = \left\{(X_i, Y_i) \in \mathcal{I}_2^l : X_i \in l_k^-(\phi)\right\}.$$

Using the absolute residuals for the samples in $\mathcal{I}_2^{l_k^+(\phi)}$ and $\mathcal{I}_2^{l_k^-(\phi)}$, we can construct the confidence intervals $\hat{C}_{l_k^+(\phi)}(x)$ and $\hat{C}_{l_k^-(\phi)}(x)$ and the associated quantities in the objective function, $\hat{W}_{l_k^+(\phi)}, \hat{W}_{l_k^-(\phi)}, \hat{S}_{l_k^+(\phi)}$ and $\hat{S}_{l_k^-(\phi)}$. Then we find the optimal covariate k_l^* and threshold ϕ_l^* for splitting subgroup l as

$$(k_l^*, \phi_l^*) = \underset{(k,\phi)}{\arg\min}\, \lambda\left(\hat{W}_{l_k^+(\phi)} + \hat{W}_{l_k^-(\phi)}\right) + (1-\lambda)\left(\hat{S}_{l_k^+(\phi)} + \hat{S}_{l_k^-(\phi)}\right).$$

For (k_l^*, ϕ_l^*), we compute $\hat{W}^*_{l^\pm} = \hat{W}_{l^+_{k^*}(\phi^*)} + \hat{W}_{l^-_{k^*}(\phi^*)}$ and $\hat{S}^*_{l^\pm} = \hat{S}_{l^+_{k^*}(\phi^*)} + \hat{S}_{l^-_{k^*}(\phi^*)}$. To improve the objective in (5.5), we split the subgroup l into $l_{k^*}^+(\phi^*)$ and $l_{k^*}^-(\phi^*)$ only if the reduction in the impurity of the confident homogeneity is sufficiently large:

$$\frac{\lambda \hat{W}^*_{l^\pm} + (1-\lambda)\hat{S}^*_{l^\pm}}{\lambda \hat{W}_l + (1-\lambda)\hat{S}_l} \leq 1 - \gamma \tag{5.7}$$

Here, $\gamma \in [0, 1)$ is a hyperparameter for regularization. We refer to (5.7) as the *confident criterion* which prevents overfitting that leads to a large number of small subgroups.

After the splitting decision, we remove l from Π_c; if we have split l, we remove l from Π and add the two split sets to both Π and Π_c. We continue recursively until Π_c is empty, i.e., no further splitting is productive. When the procedure stops, we will have obtain an estimator $\hat{\mu}$ and a partition Π and a confidence interval $\hat{C}_l(x) = \left[\hat{\mu}(x) - \hat{Q}^l_{1-\alpha}, \hat{\mu}(x) + \hat{Q}^l_{1-\alpha}\right]$ for every $l \in \Pi$. The entire procedure of our R2P method is summarized in Algorithm 2.

5.3 Robust recursive partitioning for CATEs

We now apply the R2P method to convert conditional average treatment effect (CATE) estimates into subgroups. We consider a setup with N units (i.e. samples). For every unit $i \in \{1, 2, ..., N\}$, there exists a pair of treated and control potential outcomes, $Y_i(1)$ and $Y_i(0)$. Every unit has a set of covariates X_i, a treatment variable $A_i \in \{0, 1\}$ and an observed outcome Y_i. The observational data is $\mathcal{D} = \{(X_i, A_i, Y_i)\}_{i=1}^N$. Under Assumption 7, the CATE at x is given by

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mu_1(x) - \mu_0(x),$$

where $\mu_a(x) = \mathbb{E}[Y \mid X = x, A = a]$ for $a = 0, 1$. We estimate each μ_a by a regression model $\hat{\mu}_a$, then estimate $\tau(x)$ by $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$.

To construct an accurate $\hat{\tau}(x)$, we can let $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ be one of the powerful machine learning models in the literature (Alaa and van der Schaar, 2017; Athey and Imbens, 2016; Shalit et al., 2017; Wager and Athey, 2018; Yoon et al., 2018; Zhang et al., 2020). We set the target coverage rate of $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ as $\sqrt{1-\alpha}$. We can construct a confidence interval for $\mu_0(x)$ and $\mu_1(x)$ using SCP, respectively. Denote the $\sqrt{1-\alpha}$ confidence intervals by

$$\hat{C}_1(x) = \left[\hat{\mu}_1(x) - \hat{Q}^1_{\sqrt{1-\alpha}}, \hat{\mu}_1(x) + \hat{Q}^1_{\sqrt{1-\alpha}}\right] \text{ and } \hat{C}_0(x) = \left[\hat{\mu}_0(x) - \hat{Q}^0_{\sqrt{1-\alpha}}, \hat{\mu}_0(x) + \hat{Q}^0_{\sqrt{1-\alpha}}\right].$$

We set the confidence interval for $\tau(x)$ to be

$$\hat{C}_\tau(x) = \left[\hat{\mu}_1(x) - \hat{\mu}_0(x) - \hat{Q}^1_{\sqrt{1-\alpha}} - \hat{Q}^0_{\sqrt{1-\alpha}}, \hat{\mu}_1(x) - \hat{\mu}_0(x) + \hat{Q}^1_{\sqrt{1-\alpha}} + \hat{Q}^0_{\sqrt{1-\alpha}}\right].$$

In the confidence interval $\hat{C}_\tau(x)$, the upper endpoint is given as the difference between the upper endpoint of $\hat{C}_1$ and the lower endpoint of $\hat{C}_0$, and the lower endpoint is given as the difference between the lower endpoint of $\hat{C}_1$ and the upper endpoint of $\hat{C}_0$. If the coverage rates for $\hat{C}_1$ and $\hat{C}_0$ are $\sqrt{1-\alpha}$, the coverage rate for $\hat{C}_\tau$ will be $1-\alpha$. A parallel work by Lei and Candès (2021) develops a less conservative conformal prediction method for individual treatment effects based on the weighted conformal prediction method in (Tibshirani et al.,

2019). Our contribution in this chapter is different by converting CATE estimates into interpretable subgroups, rather than prediction intervals.

With $\hat{\tau}(x)$ and its confidence interval $\hat{C}_{\tau,l}(x)$ for each subgroup l, we can calculate the quantities $\hat{W}_l$ and $\hat{S}_l$ in (5.6), and solve the robust partitioning problem in (5.5) by applying the R2P method in Algorithm 2, with two minor changes: 1) each sample in the dataset is a triple (X_i, A_i, Y_i), and the model $\hat{\mu}(x)$ is replaced by $\hat{\tau}(x)$. We call this modified method R2P-CATE. If the confidence intervals overlap across the subgroups, we can not conclude that the subgroups are well-identified. If the intervals have little or no overlap across subgroups, we can conclude that the subgroups are well-identified. By identifying subgroups this way, R2P is more robust against inconclusive and false discoveries in applications, e.g., drug development, which may save a large amount of resources spent on confirmatory trials.

5.4 Related works

Subgroup analysis methods with recursive partitioning have been widely studied based on regression trees (Athey and Imbens, 2016; Johansson et al., 2018b; Su et al., 2009; Tran and Zheleva, 2019). In these methods, once the subgroups (i.e., leaves in the tree structure) are constructed, the treatment effects are estimated by the corresponding sample mean estimator on each leaf. To represent the non-linearity such as interactions between treatment and covariates, Seibold et al. (2016) integrate a parametric model into the regression trees for subgroup analysis. However, such an approach only works for limited types of models, which is not particularly satisfying given the fact that causal inference is more accurate by estimating the outcomes with machine learning models (Alaa and van der Schaar, 2017; Shalit et al., 2017; Zhang et al., 2020). The global model interpretation method proposed by Yang et al. (2018) can analyze the subgroup structure of arbitrary models but it depends on local model interpreters and does not consider an application to treatment effects estimation.

A variety of criteria have been proposed in regression trees for recursive partitioning in the literature. The adaptive criterion (Johansson et al., 2018b) identifies subgroups with heterogeneous treatment effects by maximizing the heterogeneity across subgroups. The honest criterion (Athey and Imbens, 2016) splits the samples into two subsets, then use the first subset to build a tree structure and the second subset for estimating the treatment effects. This criterion can prevent overfitting and eliminate the bias in the adaptive criterion. Modelling the interactions between the treatment and covariate variables is proposed in the partitioning criterion (Su et al., 2009). Yang et al. (2018) propose to use the contribution matrix of the samples from local model interpreters for partitioning. Some of these criteria construct confidence intervals using the estimated variances, but these intervals may fail to

achieve the coverage guarantee in finite samples. Johansson et al. (2018b) propose a conformal prediction method to construct confidence intervals is proposed for regression trees. The adaptive criterion they use for partitioning does not take into account the confidence intervals. The confident criterion in R2P is different from these criteria. It constructs subgroups based on both heterogeneity and homogeneity measured by the confidence intervals.

5.5 Experiments

In this section, we evaluate R2P-CATE (abbreviated as R2P) by comparing its performance with some baseline subgroup analysis methods. Specifically, we compare R2P-CATE with four baselines: standard regression trees for causal effects (CT-A) (Breiman et al., 1984), conformal regression trees for causal effects (CCT) (Johansson et al., 2018b), causal trees with honest criterion (CT-H) (Athey and Imbens, 2016), and causal trees with generalization costs (CT-L) (Tran and Zheleva, 2019). Details of the baseline methods are provided in Section 5.7.2. For the CATE estimator of R2P, we use the causal multi-task Gaussian process (CMGP) (Alaa and van der Schaar, 2017). Because individual treatment effects are never observed in any real data, we use two synthetic and two semi-synthetic datasets. The first synthetic dataset (Synthetic dataset A) is taken from the article (Athey and Imbens, 2016). Dataset A has little homogeneity within subgroups. We offer the second synthetic dataset (Synthetic dataset B) that has more covariates and greater homogeneity within subgroups than dataset A. Dataset B is inspired by the initial clinical trial developed for remdesivir (Wang et al., 2020) which is a treatment for COVID-19. The trial reports a result that remdesivir leads to a faster time of clinical improvement for patients with a shorter time from symptom onset to starting the trial. The two semi-synthetic datasets are based on real-world covariates; the first uses the Infant Health and Development Program (IHDP) dataset (Hill, 2011) and the second uses the Collaborative Perinatal Project (CPP) dataset (Dorie et al., 2019). More details of the datasets can be found in Section 5.7.1. We next present experimental results averaged over 50 independent runs on these datasets.

The optimal ground truth of subgroups depends on multiple objectives, including homogeneity, heterogeneity, and the number of subgroups. In the literature, the usual metric used is variance, because greater heterogeneity across subgroups and homogeneity within each subgroup on a large testing set generally indicate a good subgroup analysis method. We denote the set of test samples by $\mathcal{D}^*$ and the test samples that belong to the subgroup l as $\mathcal{D}_l^*$. We define the mean and variance of treatment effects on the test samples in the subgroup l as $\text{Mean}(\mathcal{D}_l^*)$ and $\text{Var}(\mathcal{D}_l^*)$, respectively. We define the *heterogeneity across subgroups* V^{across} as the sample variance of the group means $\{\text{Mean}(\mathcal{D}_l^*)\}_{l\in\Pi}$. We measure the *homogeneity within subgroups* by the *average intra-subgroup variance* $V^{\text{in}}(\mathcal{D}^*) = |\Pi|^{-1}\sum_{l\in\Pi}\text{Var}(\mathcal{D}_l^*)$. We

Table 5.1 Performance of R2P and baseline methods: the measures V^{across} and V^{in}, the widths and coverage rates of confidence intervals. The best results are highlighted in bold.

Metrics	V^{across}	V^{in}	# SGs	CI width	Coverage (%)
Synthetic dataset A					
R2P	**0.22±.01**	**0.03±.001**	4.9±.16	**0.08±.003**	98.98±.24
CCT	0.18±.02	0.05±.01	4.4±.24	7.42±.48	100.0±.00
CT-A	0.19±.02	0.04±.01	4.7±.21	3.96±.16	99.99±.02
CT-H	0.12±.03	0.11±.02	3.1±.39	4.39±.22	99.98±.02
CT-L	0.12±.02	0.10±.02	2.9±.35	5.22±.02	99.97±.06
Synthetic dataset B					
R2P	**2.39±.04**	**0.12±.01**	5.0±.16	**0.88±.06**	98.86±.23
CCT	1.97±.14	0.58±.15	5.0±.13	5.95±.59	99.86±.13
CT-A	2.24±.06	0.30±.05	5.1±.15	2.77±.20	97.73±.76
CT-H	2.07±.13	0.53±.13	4.5±.15	3.38±.32	98.20±.73
CT-L	0.80±.26	1.77±.27	3.1±.28	6.92±.53	99.44±.47
IHDP dataset					
R2P	**0.46±.04**	**0.38±.03**	4.1±.12	**1.27±.22**	97.93±.39
CCT	0.30±.04	0.53±.05	4.3±.13	5.70±.23	99.59±.12
CT-A	0.31±.04	0.57±.05	4.1±.08	3.71±.08	97.41±.42
CT-H	0.28±.05	0.56±.05	3.8±.14	3.76±.14	97.76±.40
CT-L	0.27±.06	0.64±.05	2.8±.23	4.75±.15	98.97±.30
CPP dataset					
R2P	**0.06±.02**	**0.10±.01**	5.7±.30	**1.11±.13**	98.52±.34
CCT	0.03±.02	0.12±.01	6.4±.20	3.60±.12	99.54±.23
CT-A	0.03±.01	0.12±.01	6.6±.18	2.45±.06	96.60±.50
CT-H	0.01±.00	0.14±.01	5.2±.23	2.67±.06	98.01±.40
CT-L	0.01±.01	0.14±.01	2.9±.29	3.23±.07	99.49±.23

also report the average number of subgroups discovered by the methods. We set α to 0.05, so we demand a 95% marginal coverage of CATEs on the testing samples.

Table 5.1 reports the performance of R2P and the baseline methods on the four datasets described above. Keep in mind that *larger* V^{across} means greater heterogeneity *across* subgroups, while *smaller* V^{in} means greater homogeneity *within* subgroups. As the table shows, R2P displays by far the best performance on all four datasets: the greatest heterogeneity across subgroups, the greatest homogeneity within subgroups, and the narrowest confidence intervals. It delivers all these results while identifying comparable numbers of subgroups. We conclude that R2P identifies subgroups more effectively than the baseline methods. All the methods achieve the 95% target coverage rate. The confidence intervals from R2P are much

narrower than the other methods. This reflects one of the strengths of R2P, which is the ability to leverage *any* black-box machine learning models to estimate CATEs.

However, the confidence intervals from R2P are not well-calibrated, i.e. covering more than 95% of the samples. This is due to two reasons mentioned above: (1) the intervals from SCP are targeted to cover the random variable (individual treatment effect) $Y_i(1) - Y_i(0)$ rather than the conditional mean (CATE); (2) we combine the $(\sqrt{1-\alpha})$-confidence intervals for μ_0 and μ_1 to construct the $(1-\alpha)$-confidence interval for τ. In general, applying distribution-free uncertainty quantification techniques (e.g. conformal prediction) to some unobserved parameters is challenging because we take into account the most complex and simple data generating distribution at the same time; see Barber (2020); Lee and Barber (2021) for theoretical results on binary regression.

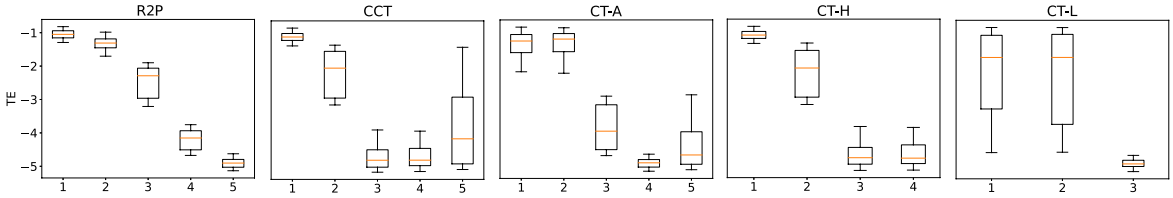


Figure 5.3 Treatment effects for the identified subgroups on Synthetic dataset B. Each box represents the range between the 25th and 75th percentiles of the treatment effects on the test samples; each whisker represents the range between the 5th and 95th percentiles.

The effectiveness of R2P can also be seen in Figure 5.3, which provides, for R2P and each of the four baseline methods, boxplots of the distribution of treatment effects for each identified subgroup on Synthetic dataset B. R2P identifies subgroups reliably: the distributions of treatment effects are non-overlapping or well-discriminated across subgroups. By contrast, the other methods have false discoveries of overlapping subgroups.

Table 5.2 Normalized V^{in} of R2P

Dataset	Synthetic dataset A	Synthetic dataset B	IHDP	CPP
$\tilde{V}^{\text{in}}$	0.110±.005	0.046±.003	0.459±.033	0.691±.076

To indicate the gain from subgroup analysis obtained by R2P, and hence to indicate the effectiveness of recursive partitioning, we compare V^{in}, the homogeneity within subgroups obtained by R2P in Table 5.1, against the homogeneity within the entire population, V^{pop}. We divide V^{in} by V^{pop} to obtain the normalized V^{in}, denoted by $\tilde{V}^{\text{in}}$ in Table 5.2. R2P reduces the average intra-subgroup variance by 89% and more than 95% on Synthetic datasets A and

B, respectively. On the IHDP and CPP datasets, R2P reduces the average intra-subgroup variance by more than 50% and 30%, respectively.

5.6 Conclusion

The understanding of treatment effects plays an important role in many areas, especially in medicine and public policy. In medicine, subgroup analysis makes it possible to identify groups of patients suffering from a particular disease for whom a particular drug is effective and safe and other groups for whom the same drug is ineffective and unsafe. Similarly, subgroup analysis may make it possible to identify groups of patients for whom one course of treatment (e.g. a particular mode of radiotherapy or chemotherapy) is preferable to another. In public policy, subgroup analysis can identify groups of people or geographic regions for which particular interventions (e.g., providing mosquito nets to combat malaria) are likely to be successful or unsuccessful. Our subgroup analysis method R2P improves over the baseline methods and therefore has an impact on a variety of applications.

We believe R2P opens up some future research directions for subgroup analysis with conformal prediction. In R2P, the intervals and subgroups are constructed using the same validation data, which breaks the data exchangeability to establish subgroup-level coverage guarantees. One possible solution to this problem is by applying concentration inequalities (Kim et al., 2021; Park et al., 2020; Vovk, 2012) and controlling the partition complexity. R2P can be implemented on any CATE learners reviewed in Section 1.3.2. If a learner estimates $\hat{\tau}$ directly, we can let $\hat{\mu}_1 = \mu_0 + \hat{\tau}$ and construct the confidence interval for $\hat{\tau}$ as in R2P-CATE above. It is interesting to study how to better construct intervals and subgroups by conformal prediction when using advanced CATE learners.

5.7 Further experimental details

5.7.1 Datasets

Synthetic dataset A. We first consider the synthetic treatment effect model proposed in (Athey and Imbens, 2016). The potential outcome $Y_i(a)$ for $a \in \{0, 1\}$ is given by

$$Y_i(a) = \eta(X_i) + \frac{1}{2}(2a - 1)\kappa(X_i) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, 0.01)$, $X_{i,k} \sim \mathcal{N}(0, 1)$, and $\eta(\cdot)$ and $\kappa(\cdot)$ are the design functions. The outcome Y_i is determined by the design functions. The functions $\eta(\cdot)$ and $\kappa(\cdot)$ are the control

outcome and conditional average treatment effect for some given covariates, respectively. We consider the following design functions with two covariates,

$$\eta(X_i) = \frac{1}{2}X_{i1} + X_{i2} \text{ and } \kappa(X_i) = \frac{1}{2}X_{i1}.$$

In the experiments, we generate 300 samples for training and 1000 samples for testing.

Synthetic dataset B. We introduce a synthetic model based on the initial clinical trial results of remdesivir to COVID-19 (Wang et al., 2020). The result shows that remdesivir results in a faster time to clinical improvement for the patients with a shorter time from symptom onset to starting the trial. Since the clinical trial data is not public, we construct a synthetic model following this result. We consider the following 10 baseline covariates: age $\sim \mathcal{N}(66, 4)$, white blood cell count ($\times 10^9$ per L) $\sim \mathcal{N}(66, 4)$, lymphocyte count ($\times 10^9$ per L) $\sim \mathcal{N}(0.8, 0.1)$, platelet count ($\times 10^9$ per L) $\sim \mathcal{N}(183, 20.4)$, serum creatinine (U/L) $\sim \mathcal{N}(68, 6.6)$, aspartate aminotransferase (U/L) $\sim \mathcal{N}(31, 5.1)$, alanine aminotransferase (U/L) $\sim \mathcal{N}(26, 5.1)$, lactate dehydrogenase (U/L) $\sim \mathcal{N}(339, 51)$, creatine kinase (U/L) $\sim \mathcal{N}(76, 21)$, and time from symptom onset to starting the trial (days) $\sim$ Unif$(4, 14)$.

We use a logistic function on the time covariate to produce different effectiveness (i.e., the faster time to clinical improvement with a shorter time from symptom onset to the trial). The control and treatment outcomes are given by

$$Y(0) \sim \mathcal{N}(\beta X_{-0} + 1/(1 + e^{-(X_0 - 9)}) + 5, 0.1), \text{ and}$$

$$Y(1) \sim \mathcal{N}(\beta X_{-0} + 5/(1 + e^{-(X_0 - 9)}), 0.1),$$

where X_{-0} represents the matrix of the standardized (zero-mean and unit standard deviation) covariates except for the time covariate X_0. The coefficients in β are randomly sampled among the values $(0, 0.1, 0.2, 0.3, 0.4)$ with the probability $(0.6, 0.1, 0.1, 0.1, 0.1)$, respectively. This synthetic model is constructed to be consistent with the trial result in (Wang et al., 2020) such that the time to clinical improvement (i.e., the treatment effect) becomes faster with a shorter time from symptom onset to the trial.

IHDP dataset. The Infant Health and Development Program (IHDP) is a randomized experiment intended to enhance the cognitive and health status of low-birth-weight, premature infants through intensive high-quality child care and home visits from a trained provider. Based on the real experimental data about the impact of the IHDP on the subjects' IQ scores at the age of three, the semi-synthetic (simulated) dataset is developed and has been used to evaluate methods for treatment effects estimation (Alaa and van der Schaar, 2017; Hill, 2011; Louizos et al., 2017; Shalit et al., 2017). All outcomes are simulated using the real covariates. The dataset consists of 747 subjects (608 untreated and 139 treated), and 25 input covariates

for each subject. We generated the outcomes using the standard non-linear mean outcomes of "Response surface B" setting in (Hill, 2011). A noise $\epsilon \sim \mathcal{N}(0, 0.1)$ is added to each observed outcome. In the experiments, we use 80% samples for training and 20% samples for testing.

CPP dataset. In the 2016 Atlantic Causal Inference Conference competition (ACIC), a semi-synthetic dataset is created based on the data from the Collaborative Perinatal Project (CPP) (Dorie et al., 2019). It consists of multiple datasets that are generated by distinct data generating processes (causal graphs) and random seeds. Each dataset consists of 4802 observations with 58 covariates of which 3 are categorical, 5 are binary, 27 are count data, and the remaining 23 are continuous. The factual and counterfactual samples are drawn from a generative model and a noise $\epsilon \sim \mathcal{N}(0, 0.1)$ is added to each observed outcome. In the experiments, we use the dataset with index 1 provided in (Dorie et al., 2019) and drop the rows whose $Y(1)$ or $Y(0)$ above the 99% quantile or below the 1% quantile to avoid the outliers. The dataset consists of 35% treated units and 65% control units. We randomly pick 500 samples for training and 300 samples for testing in this dataset.

5.7.2 Benchmark methods

Robust recursive partitioning for CATE (R2P-CATE). In R2P-CATE, our CATE estimator $\hat{\tau}(x)$ is the causal multi-task Gaussian process (CMGP) in (Alaa and van der Schaar, 2017). Using $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ in CMGP, we construct the confidence interval for $\hat{\tau}(x)$ as described in Section 5.3. We set λ in (5.5) to 0.5 and γ in (5.7) to 0.05.

Standard regression trees for causal effects (CT-A). Because the standard regression trees in (Breiman et al., 1984) is not developed for estimating treatment effects, we implement a modified version of the standard regression trees for causal effects estimation in (Athey and Imbens, 2016). In this modified version, the regression trees recursively partitions according to a criterion based on the mean squared error (MSE) of the treatment effects. In the literature, this criterion is called the adaptive criterion. In the experiments, we set the minimum number of training samples in each leaf as 20 since CT-A does not need to split the data samples into two subsets for validation as in other methods.

Conformal regression trees for causal effects (CCT). We modify the conformal regression trees in (Johansson et al., 2018b) for our experiments of treatment effect estimation. We implemented CCT by applying the split conformal prediction method to CT-A.

Causal trees with honest criterion (CT-H). We implement the causal trees method in (Athey and Imbens, 2016). The method modifies the standard regression trees for causal

effects in which an honest criterion is used instead of the adaptive criterion. It divides tree-building and treatment effect estimation into two steps. The samples are split into two subsets: training samples to build the trees and samples to estimate treatment effects. This two-step procedure makes the tree-building and the treatment effect estimation process independent, which prevents overfitting in treatment effect estimation.

Causal trees with generalization costs (CT-L). We implement causal trees with a criterion considering generalization costs in (Tran and Zheleva, 2019). This method is a modified version of the causal trees in (Athey and Imbens, 2016). It splits the data samples into the training and validation samples and builds the trees using the training samples while penalizing based on generalization ability using the validation samples. In the experiments, we use half of the samples for the training and validation, respectively.

In R2P-CATE, CCT, CT-H and CT-L, we set the minimum number of training samples in each leaf as 10. We also implement a pruning step to prevent overfitting.

5.8 Additional experiments

5.8.1 Average overlap of treatment effects across subgroups

The average overlap of treatment effects across subgroups indicates whether the subgroups are false discoveries. Specifically, we define a treatment effect interval of each subgroup l as $[a_l(p), b_l(q)]$, where $a_l(p)$ and $b_l(q)$ are p-th and q-th percentiles of the treatment effects in the subgroup l. We define the average overlap of treatment effects across subgroups as the overlapped width of the treatment effect intervals between all the pairs of the subgroups. We provide the average overlap of R2P and the baselines for all datasets with $p = 20$ and $q = 80$. Table 5.3 shows that the average overlap in R2P is significantly small than the baselines, which implies that R2P performs best for identifying non-overlapping subgroups.

Table 5.3 Average overlap of treatment effects across subgroups.

	Synthetic dataset A	Synthetic dataset B	IHDP	CPP
R2P	**0.45±.06**	**0.14±.03**	**0.32±.04**	**0.23±.03**
CCT	1.35±.04	0.63±.15	0.81±.09	0.55±.04
CT-A	1.13±.21	0.44±.09	0.59±.08	0.47±.05
CT-H	0.60±.20	0.60±.16	0.76±.10	0.45±.05
CT-L	0.87±.18	2.27±.55	0.46±.10	0.24±.04

5.8.2 Maximum depth for partitioning

We set the maximum depth of each method to be 2, which limits the maximum number of identified subgroups by 4. Table 5.4 shows that overall, R2P has both the highest variance across subgroups and the lowest intra-subgroup variance. This implies that every partition R2P makes is more effective for identifying subgroups than the other methods. In Synthetic dataset B, The variance across subgroups in CT-A is slightly larger than R2P. But the intra-subgroup variance of CT-A is much larger than R2P.

Table 5.4 Results with maximum depth for partitioning.

Metrics	V^{across}	V^{in}	# SGs	CI width	Coverage (%)
Synthetic dataset A					
R2P	**0.27±.01**	**0.04±.001**	4.0±.04	**0.09±.002**	99.06±.23
CCT	0.20±.02	0.07±.01	3.4±.22	8.28±.42	100.0±.00
CT-A	0.24±.02	0.06±.01	3.6±.15	4.29±.16	99.99±.02
CT-H	0.14±.03	0.11±.02	2.5±.27	4.71±.18	99.99±.01
CT-L	0.13±.03	0.12±.02	2.4±.25	5.49±.16	99.99±.01
Synthetic dataset B					
R2P	2.19±.04	**0.14±.01**	4.0±.00	**0.98±.07**	99.28±.16
CCT	2.10±.08	0.43±.09	3.9±.09	6.05±.46	99.99±.01
CT-A	**2.23±.06**	0.30±.06	3.9±.08	2.97±.20	98.68±.52
CT-H	2.13±.09	0.45±.09	3.8±.12	3.45±.25	98.66±.68
CT-L	0.82±.23	1.74±.25	2.8±.19	6.71±.53	99.70±.30
IHDP dataset					
R2P	**0.52±.05**	**0.42±.05**	3.5±.14	**1.21±.13**	97.24±.50
CCT	0.28±.05	0.62±.06	3.5±.14	6.11±.21	99.55±.14
CT-A	0.33±.04	0.58±.05	3.7±.13	3.64±.08	97.25±.43
CT-H	0.30±.05	0.60±.05	3.5±.14	3.72±.12	97.17±.43
CT-L	0.30±.06	0.67±.04	2.7±.18	4.72±.17	98.99±.23
CPP dataset					
R2P	**0.05±.02**	**0.12±.01**	3.7±.13	**1.22±.14**	99.04±.23
CCT	**0.05±.03**	0.13±.01	3.4±.14	3.66±.13	99.50±.21
CT-A	0.02±.01	0.13±.01	3.5±.14	2.44±.05	96.17±.51
CT-H	0.01±.00	0.14±.01	3.4±.14	2.59±.06	97.31±.56
CT-L	0.02±.02	0.13±.01	2.7±.21	3.25±.07	99.55±.22

5.8.3 Different CATE estimators

We provide the results of R2Ps with different CATE estimators in Table 5.5. For this, we integrate R2P with the CATE estimators based on dragonnet (DN) (Shi et al., 2019), random forest (RF), and CT-H (Athey and Imbens, 2016). To evaluate the precision of the CATE estimation, we use the empirical precision in the estimation of heterogeneous effect (PEHE) (Hill, 2011), PEHE $= N^{-1} \sum_{i=1}^{N} [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \tau(x)]^2$. The lower PEHE implies a more accurate CATE estimation. In the table, we can see that with a more accurate CATE estimator, R2P constructs better subgroups.

Table 5.5 Results of R2Ps with different CATE estimators.

Metrics	V^{across}	V^{in}	# SGs	CI width	$\sqrt{\text{PEHE}}$
Synthetic dataset A					
R2P	0.22±.01	0.03±.001	4.9±.16	0.08±.003	0.01±.00
R2P-DN	0.21±.02	0.05±.01	4.9±.27	0.71±.05	0.06±.00
R2P-RF	0.18±.02	0.08±.01	4.9±.26	2.91±.12	0.33±.01
R2P-CT-H	0.12±.02	0.07±.03	4.3±.49	8.87±.32	0.51±.02
Synthetic dataset B					
R2P	2.39±.04	0.12±.01	5.0±.16	0.88±.06	0.16±.01
R2P-DN	2.32±.06	0.19±.03	5.1±.17	2.24±.09	0.41±.02
R2P-RF	2.20±.08	0.33±.07	5.1±.15	3.10±.32	0.42±.04
R2P-CT-H	1.05±.25	1.51±.25	4.6±.41	6.42±.51	0.92±.12
IHDP dataset					
R2P	0.46±.04	0.38±.03	4.1±.12	1.27±.22	0.22±.02
R2P-DN	0.41±.03	0.44±.04	4.3±.13	1.92±.09	0.33±.01
R2P-RF	0.32±.05	0.55±.05	4.0±.32	3.07±.13	0.39±.02
R2P-CT-H	0.09±.04	0.75±.05	2.7±.50	6.56±.25	0.83±.03
CPP dataset					
R2P	0.06±.02	0.10±.01	5.7±.30	1.11±.13	0.13±.01
R2P-DN	0.06±.02	0.10±.01	6.0±.26	1.52±.07	0.19±.01
R2P-RF	0.04±.02	0.12±.01	6.0±.27	1.80±.06	0.23±.01
R2P-CT-H	0.00±.00	0.14±.00	1.0±.004	4.20±.10	0.41±.01

5.8.4 Interpretable versus non-interpretable subgroups

One naive way to construct subgroups is by dividing the covariate space with respect to the quantiles of estimated CATEs. However, this approach fails to satisfy the essential

requirement of subgroup analysis: interpretability. The estimates from a black-box CATE estimator are non-interpretable. Similarly, the subgroups defined by the estimated quantiles also fail to explain why the samples are assigned to a particular subgroup (in terms of the covariates). To demonstrate this problem clearly, in Figure 5.4, we divide the covariate space of the IHDP dataset into four subgroups based on the estimated quantiles by CMGP, $[0, 25), [25, 50), [50, 75)$ and $[75, 100]$. The colours indicate which subgroup each sample belongs to. The subgroups are overlapped and hard to interpret in terms of the covariates. In contrast, R2P constructs easy-to-interpret subgroups based on a tree structure.

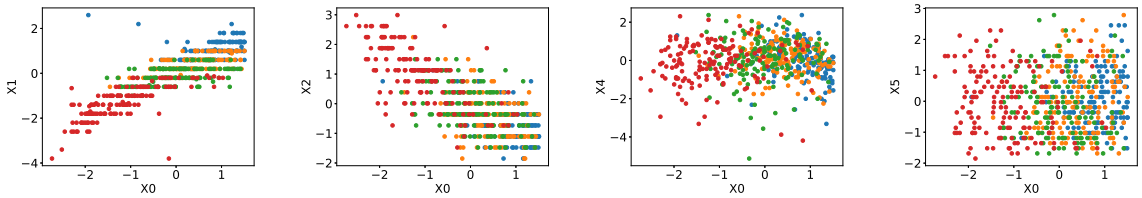
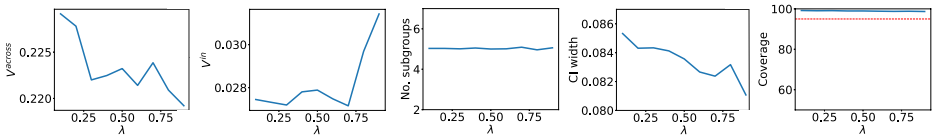


Figure 5.4 Subgroups on four intervals of the estimated quantiles by CMGP.

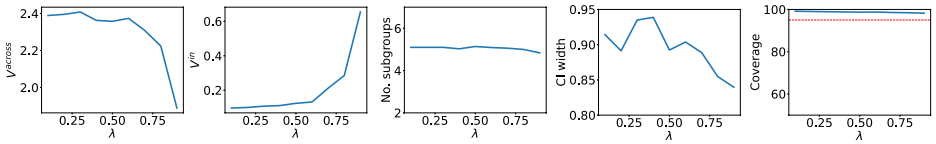
5.8.5 Impacts of hyperparameters

Here we evaluate the impacts of the hyperparameters γ and λ on the performance of R2P. We provide the results with the hyperparameters $\gamma \in \{0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.5\}$ and $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We use the same experiment setup above and repeat each experiment 50 times for each hyperparameter.

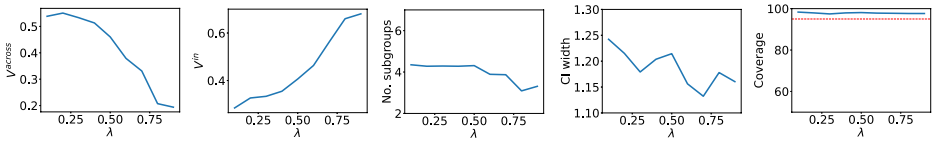
The impact of the hyperparameter γ in (5.7) is illustrated in Figure 5.5. As γ increases from 0 to 1, the number of subgroups converges to one since no partition is accepted. The performance of R2P degrades in the following aspects: V^{across} decreases, V^{in} increases, and the confidence interval width increases generally. Thus, a smaller γ may be a better choice. However, if γ is too small, the subgroups constructed by R2P overfit the data. This overfitting issue results in the loss of generalization ability on the unseen data and leads to a large number of non-informative subgroups. The impact of the hyperparameter λ is illustrated in Figure 5.6. The hyperparameter $\lambda \in [0, 1]$ determines the importance of W_l and S_l in the objective function (5.5) for partitioning. With smaller λ, the homogeneity within each subgroup is more emphasized in the objective. Then R2P constructs a larger number of subgroups, which results in larger V^{across} and smaller V^{in}. On the other hand, with a larger λ, S_l is weighted higher in the objective, then the confidence interval width decreases generally. In practice, we should search for the hyperparameters γ and λ that achieve desirable intra-subgroup homogeneity and inter-subgroup heterogeneity.



(a) Results on Synthetic dataset A.

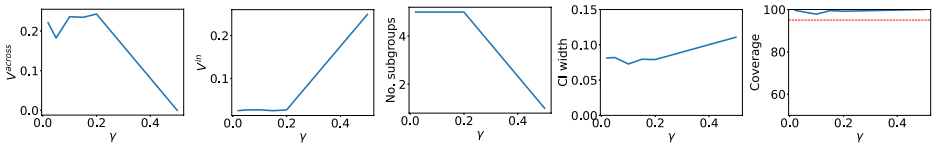


(b) Results on Synthetic dataset B.

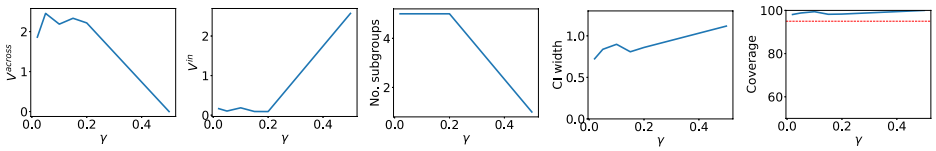


(c) Results on the IHDP dataset.

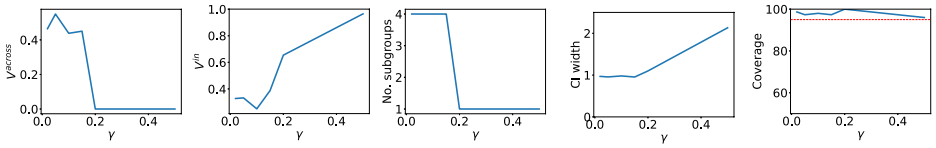
Figure 5.6 Results on different λ. (the red line indicates the target coverage rate).



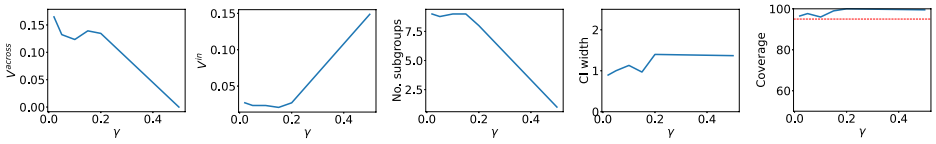
(a) Results on Synthetic dataset A.



(b) Results on Synthetic dataset B.



(c) Results on the IHDP dataset.



(d) Results on the CPP dataset.

Figure 5.5 Results on different γ. (the red line indicates the target coverage rate).

Bibliography

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abraham, S. and Sun, L. (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Available at SSRN 3158747*.

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

Alaa, A. and Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138.

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432.

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16.

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340.

Banerjee, A., Chassang, S., and Snowberg, E. (2017). Decision theoretic approaches to experiment design and external validity. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 141–174. North-Holland.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Barber, R. F. (2020). Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487–3524.

Basse, G., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.

Basu, D. (1980). Randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75(371):575–582.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bates, S., Sesia, M., Sabatti, C., and Candès, E. (2020). Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences*, 117(39):24117–24126.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., and Jacobsen, J.-H. (2019). Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197.

Berrevoets, J., Jordon, J., Bica, I., Gimson, A., and van der Schaar, M. (2020). OrganITE: Optimal transplant donor organ offering using an individual treatment effect. In *Advances in Neural Information Processing Systems*, volume 33, pages 20037–20050. Curran Associates, Inc.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Biewald, L. (2020). Experiment tracking with weights and biases. Software available from wandb.com.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563.

Bose, A., Ling, H., and Cao, Y. (2018). Adversarial contrastive estimation. In *ACL*.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, 6(1):1–9.

Cai, T. T., Low, M., and Ma, Z. (2014). Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, 109(507):1054–1070.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Caughey, D., Dafoe, A., Li, X., and Miratrix, L. (2021). Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *arXiv preprint arXiv:2101.09195*.

Ceylan, C. and Gutmann, M. (2018). Conditional noise-contrastive estimation of unnormalised models. In *ICML*.

Cheng, D., Chakrabortty, A., Ananthakrishnan, A. N., and Cai, T. (2020a). Estimating average treatment effects with a double-index propensity score. *Biometrics*, 76(3):767–777.

Cheng, D., Li, J., Liu, L., and Liu, J. (2020b). Sufficient dimension reduction for average causal effect estimation. *arXiv preprint arXiv:2009.06444*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Cohen, P. L. and Fogarty, C. B. (2021+). Gaussian prepivoting for finite population causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (to appear)*.

Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics*, volume 482. John Wiley & Sons.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450.

Crump, R., Hotz, V. J., Imbens, G., and Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical science*, pages 331–345.

Ding, P., Feller, A., and Miratrix, L. (2015). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671.

Dorie, V. (2016). Npci: Non-parametrics for causal inference. *URL: https://github.com/vdorie/npci*.

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pages 181–187.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):nil.

Everitt, B. and Skrondal, A. (2002). *The Cambridge dictionary of statistics*, volume 106. Cambridge University Press Cambridge.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh and London.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513.

Fisher, R. A. (1935). *Design of experiments*. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner Publishing Co.

Fogarty, C. B. (2021). Prepivoted permutation tests. *arXiv preprint arXiv:2102.04423*.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Freedman, D. A. (2006). Statistical models for causation: what inferential leverage do they provide? *Evaluation review*, 30(6):691–713.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The annals of applied statistics*, 2(1):176–196.

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Gao, R., Nijkamp, E., Kingma, D. P., Xu, Z., Dai, A. M., and Wu, Y. N. (2020). Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7518–7528.

Garthwaite, P. H. (1996). Confidence intervals from randomization tests. *Biometrics*, pages 1387–1393.

Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *The Annals of Statistics*, 36(2):875–905.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2013). A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pages 738–746.

Ghosh, T., Ma, Y., and de Luna, X. (2021). Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Statistica Sinica*, 31(2):821–842.

Gill, R. D. and Robins, J. M. (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, pages 1785–1811.

Greenewald, K., Shanmugam, K., and Katz, D. (2021). High-dimensional feature selection for sample efficient treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2224–2232. PMLR.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Grimmer, J., Messing, S., and Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434.

Group, G. H. S. et al. (1987). The gambia hepatitis intervention study. *Cancer Research*, 47(21):5782–5787.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(2).

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. *Biometrika*, 105(1):239–246.

Heckman, J. J. and Karapakula, G. (2019). The perry preschoolers at late midlife: A study in design-specific inference. Technical report, National Bureau of Economic Research.

Hemming, K., Taljaard, M., McKenzie, J. E., Hooper, R., Copas, A., Thompson, J. A., Dixon-Woods, M., Aldcroft, A., Doussau, A., Grayling, M., Kristunas, C., Goldstein, C. E., Campbell, M. K., Girling, A., Eldridge, S., Campbell, M. J., Lilford, R. J., Weijer, C., Forbes, A. B., and Grimshaw, J. M. (2018). Reporting of stepped wedge cluster randomised trials: Extension of the consort 2010 statement with explanation and elaboration. *BMJ*, 363:k1614.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., and Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4(1):61–80.

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

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Hu, X. and Lei, J. (2020). A distribution-free test of covariate shift using conformal prediction. *CoRR*.

Huang, M.-Y. and Yang, S. (2022). Robust inference of conditional average treatment effects using dimension reduction. *Statistica Sinica*, 32:1–21.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Hughes, J. P., Heagerty, P. J., Xia, F., and Ren, Y. (2020). Robust inference for the stepped wedge design. *Biometrics*, 76(1):119–130.

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary clinical trials*, 28(2):182–191.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jacob, D. (2021). Cate meets ml-the conditional average treatment effect and machine learning. *arXiv preprint arXiv:2104.09935*.

Jean, N., Xie, S. M., and Ermon, S. (2018). Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *Advances in Neural Information Processing Systems*, pages 5322–5333.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Ji, X., Fink, G., Robyn, P. J., Small, D. S., et al. (2017). Randomization inference for stepped-wedge cluster-randomized trials: an application to community-based health insurance. *The Annals of Applied Statistics*, 11(1):1–20.

Jockel, K.-H. (1986). Finite sample properties and asymptotic efficiency of monte carlo tests. *The annals of Statistics*, pages 336–347.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR.

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. (2018a). Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.

Johansson, U., Linusson, H., Löfström, T., and Boström, H. (2018b). Interpretable regression trees using conformal prediction. *Expert systems with applications*, 97:394–404.

Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. M., and Wu, Y. (2016). Exploring the limits of language modeling. *ArXiv*, abs/1602.02410.

Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*, 73(361):167–170.

Kallus, N. (2020). Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR.

Kallus, N., Mao, X., and Udell, M. (2018). Causal inference with noisy and missing covariates via matrix factorization. In *NeurIPS*.

Karmakar, B., French, B., and Small, D. (2019). Integrating the evidence from evidence factors in observational studies. *Biometrika*, 106(2):353–367.

Katsevich, E. and Ramdas, A. (2020). A theoretical treatment of conditional independence testing under model-x. *arXiv preprint arXiv:2005.05506*, 4.

Kempthorne, O. (1952). *The design and analysis of experiments.* Wiley.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer.

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

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245.

Kenny, A., Voldal, E., Xia, F., Heagerty, P. J., and Hughes, J. P. (2021). Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *arXiv preprint arXiv:2111.07190*.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020a). Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. (2020b). Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33.

Kim, I., Balakrishnan, S., and Wasserman, L. (2021+). Minimax optimality of permutation tests. *Annals of Statistics (to appear)*.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Lada, A., Peysakhovich, A., Aparicio, D., and Bailey, M. (2019). Observational data for heterogeneous treatment effects with application to recommender systems. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 199–213.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Lee, H.-S., Zhang, Y., Zame, W., Shen, C., Lee, J.-W., and van der Schaar, M. (2020). Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. *Advances in Neural Information Processing Systems*.

Lee, K.-Y., Li, B., Chiaromonte, F., et al. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Annals of Statistics*, 41(1):221–249.

Lee, Y. and Barber, R. F. (2021). Distribution-free inference for regression: discrete, continuous, and in between. *arXiv preprint arXiv:2105.14075*.

Lehmann, E. L. (1975). *Nonparametrics: statistical methods based on ranks.* Holden-day, Inc.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018a). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018b). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):911–938.

Levy, J. (2019). Tutorial: Deriving the efficient influence curve for large models. *arXiv preprint arXiv:1903.01706*.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903.

Li, F., Hughes, J. P., Hemming, K., Taljaard, M., Melnick, E. R., and Heagerty, P. J. (2021). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. *Statistical Methods in Medical Research*, 30(2):612–639.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2020). Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*.

Loh, W. W., Richardson, T. S., and Robins, J. M. (2017). An apparent paradox explained. *Statistical Science*, 32(3):356–361.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554.

Lunneborg, C. (2000). Random assignment of available cases: Let the inference fit the design. *Unpublished work, University of Washington, Seattle, WA*.

Luo, W., Wu, W., and Zhu, Y. (2019). Learning heterogeneity in causal inference using sufficient dimension reduction. *Journal of Causal Inference*, 7(1).

Ma, S., Zhu, L., Zhang, Z., Tsai, C.-L., and Carroll, R. J. (2019). A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *Annals of statistics*, 47(3):1505.

Ma, Z. and Collins, M. (2018). Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *EMNLP*.

MacDorman, M. F. and Atkinson, J. O. (1999). Infant mortality statistics from the 1997 period linked birth/infant death data set. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 47:1–23.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

McAllester, D. A. (1999). Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363.

Microsoft Research (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.11.1.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6(1-2):39–75.

Mita, G., Filippone, M., and Michiardi, P. (2021). An identifiable double vae for disentangled representations. In *International Conference on Machine Learning*, pages 7769–7779. PMLR.

Monti, R. P., Zhang, K., and Hyvärinen, A. (2020). Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282.

Muntinga, M. E., Hoogendijk, E. O., Van Leeuwen, K. M., Van Hout, H. P., Twisk, J. W., Van Der Horst, H. E., Nijpels, G., and Jansen, A. P. (2012). Implementing the chronic care model for frail older adults in the netherlands: study protocol of act (frail older adults: care in transition). *BMC geriatrics*, 12(1):1–10.

Nabi, R. and Shpitser, I. (2020). Semi-parametric causal sufficient dimension reduction of high dimensional treatments. *arXiv preprint arXiv:1710.06727*.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Annals of Agricultural Sciences*, 10:1–51. (Translated to English and edited by D. M. Dabrowska and T. P. Speed, *Statistical Science* (1990), 5, 465–480).

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer.

Park, S., Bastani, O., Matni, N., and Lee, I. (2020). PAC confidence sets for deep neural networks via calibrated prediction. In *8th International Conference on Learning Representations (ICLR)*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pentina, A. and Lampert, C. (2014). A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2):225–232.

Pollard, D. (2002). *A user's guide to measure theoretic probability*. Cambridge University Press.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787.

Puelz, D., Basse, G., Feller, A., and Toulis, P. (2021+). A graph-theoretic approach to randomization tests of causal effects under general interference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (to appear)*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Roach, J. and Valdar, W. (2018). Permutation tests of non-exchangeable null models. *arXiv preprint arXiv:1808.10483*.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.

Rosenbaum, P. R. (2002b). *Observational studies*. Springer.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rosenbaum, P. R. (2010). Evidence factors in observational studies. *Biometrika*, 97(2):333–345.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, 106(493):285–295.

Rosenbaum, P. R. (2017). The general structure of evidence factors in observational studies. *Statistical Science*, 32(4):514–530.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1980a). Comment on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75(371):591.

Rubin, D. B. (1980b). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Seibold, H., Zeileis, A., and Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. *The international journal of biostatistics*, 12(1):45–63.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421.

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

Shawe-Taylor, J. and Williamson, R. C. (1997). A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Southworth, L. K., Kim, S. K., and Owen, A. B. (2009). Properties of balanced permutations. *Journal of Computational Biology*, 16(4):625–638.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Stuart, E. A., Rubin, D. B., and Osborne, J. (2004). Matching methods for causal inference: Designing observational studies. *Harvard University Department of Statistics mimeo*.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2).

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2021). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, pages 1–37.

Tchetgen, E. J. T. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Thompson, J., Davey, C., Fielding, K., Hargreaves, J., and Hayes, R. (2018). Robust analysis of stepped wedge trials using cluster-level summaries within periods. *Statistics in medicine*, 37(16):2487–2500.

Thompson, J. A., Fielding, K. L., Davey, C., Aiken, A. M., Hargreaves, J. R., and Hayes, R. J. (2017). Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in medicine*, 36(23):3670–3682.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tran, C. and Zheleva, E. (2019). Learning triggers for heterogeneous treatment effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5183–5190.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Twisk, J. W. (2021). *Analysis of Data from Randomized Controlled Trials: A Practical Guide*. Springer Nature.

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.

Van Der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, R. and De Gruttola, V. (2017). The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in medicine*, 36(18):2831–2843.

Wang, Y., Zhang, D., Du, G., Du, R., Zhao, J., Jin, Y., Fu, S., Gao, L., Cheng, Z., Lu, Q., et al. (2020). Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Welch, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika*, 29(1/2):21–52.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wu, P. and Fukumizu, K. (2020). Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. In *International Conference on Artificial Intelligence and Statistics*, pages 1157–1167. PMLR.

Yang, C., Rangarajan, A., and Ranka, S. (2018). Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, Y., Bellot, A., and Schaar, M. (2020). Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR.

Zhang, Y., Berrevoets, J., and van der Schaar, M. (2022). Identifiable energy-based representations: An application to estimating heterogeneous causal effects. *International Conference on Artificial Intelligence and Statistics (to appear)*.

Zhang, Y. and Zhao, Q. (2021). Multiple conditional randomization tests. *arXiv preprint arXiv:2104.10618*.

Zhang, Y. and Zhao, Q. (2022). What is a randomization test? *arXiv preprint arXiv:2203.10980*.

Zhao, A. and Ding, P. (2021). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. (2019). On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR.

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1).

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187.

Zhu, J., Chen, N., and Xing, E. P. (2014). Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847.