Design-Based Methods for the Analysis of Modern Randomized Experiments

by

Edward Wu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2021

Doctoral Committee:

Assistant Professor Johann A Gagnon-Bartsch, Chair
Associate Professor Ben Hansen
Professor Xuming He
Assistant Professor Zhenke Wu

Edward Wu

jameswu@umich.edu

ORCID: 0000-0001-8647-2567

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Randomized experiments are increasingly prevalent across a variety of fields, particularly in the social sciences and medicine. This is due in part to their reputation as the "gold standard" for establishing causal relationships. The proliferation of randomized experiments has resulted in a variety of challenges in a time where large data sets are becoming more common. For some experiments, a large number of pretreatment covariates are available for each participant. It is common to make adjustments for small imbalances in these baseline covariates when analyzing the results of a randomized experiment. Traditional covariate adjustment methods such as linear regression can perform poorly or fail entirely when the number of covariates is large. This can be solved by first performing model selection, which may lead to concerns about data snooping and the validity of post-selection inferences. Several authors have suggested specifying the statistical analysis in advance to address this issue. However, it may not be clear ahead of time which covariates to use for making adjustments, or if covariate adjustment will even be helpful. To address this concern, we propose a flexible covariate adjustment method, the LOOP ("Leave-One-Out Potential outcomes") estimator. This method allows for automatic variable selection, so that we do not need to know ahead of time which variables to use. In addition, the method is unbiased under the Neyman-Rubin model and generally performs at least as well as the unadjusted estimator. This alleviates concerns that the adjustment could harm the performance of the treatment effect estimate.

Covariate imbalance can also be addressed using study design. In paired experiments, participants are grouped into pairs with similar characteristics, and one observation from each pair is randomly assigned to treatment. While this study design is often successful in balancing the treatment and control groups, it may still be possible to improve precision

using covariate adjustment. We build on the LOOP estimator and propose a design-based covariate adjustment method for paired experiments. This method addresses a unique trade-off that exists for paired experiments, where it can be unclear the extent to which account for the paired structure. By addressing this trade-off, the method has the potential to improve over existing methods.

Modern randomized experiments may be accompanied by a large amount of auxiliary data, such as related observational data. Sample sizes of randomized experiments are often limited due to practical constraints. However, sample sizes for the auxiliary data can be large. We propose a covariate adjustment method that allows us to use observational data sets to make adjustments to the experimental data without bias from confounding variables leaking into our analysis. Our method also adjusts for the covariates within the randomized experiment itself, and automatically interpolates between the adjustment made using the experimental covariates and the observational data set.

Finally, we propose a method for high-dimensional classification. In this method, we have the predictors in a data set compete in a "tournament" until they have been combined into single predictor. From a computation perspective, this method is a natural fit to be used within the LOOP estimator when the outcome is binary; however, it can also be used more generally. The method shares several of the features used within the covariate adjustment methods, such as the use of a leave-one-out procedure to improve performance and interpolation between competing predictors.

CHAPTER I

Introduction

Randomized experiments are an increasingly common tool used in many fields. Randomized controlled trials have long been a fixture in biomedical and pharmaceutical research, and are frequently employed across a variety of social sciences. Large technology companies, such as Microsoft, Amazon, and Facebook, perform tens of thousands randomized experiments each day.

Throughout this dissertation, we consider several examples of recent randomized experiments and the challenges associated with the analysis of these experiments. We then develop methods to address these challenges. In Chapter II, we discuss an example studying the effects of certain interventions on reducing the amount of time in juvenile detention for at-risk youth, and another where researchers study the effects of cash transfer programs on education outcomes for students in Bogota, Colombia. In Chapters III and IV, we discuss examples from the field of education, including both an impact study involving schools in Texas and an example involving educational technology.

Many of these experiments are characterized by small sample sizes and a large number of pre-treatment covariates. For example, consider the ASSISTments TestBed (see Heffernan and Heffernan (2014) and Ostrow et al. (2016)), which we discuss in Chapter IV. ASSISTments is a computer-based learning platform used by over 50,000 students throughout the United States each year, and the TestBed is a program designed for conducting randomized

experiments within ASSISTments. Researchers can propose experiments to be run within the TestBed, and students working on a specific assignment are individually randomly assigned to treatments in the proposed experiment. For a specific experiment in the TestBed, the sample size is limited to the students working on the given assignment while the experiment is being run. Despite the small sample size of an individual experiment, the ASSISTments TestBed provides rich data sets and unique challenges. Not only are there many covariates available for each student (such as performance on all prior assignments), there is also a large amount of auxiliary data. This includes randomized experiments run on similar assignments, as well student performance data for the same assignment outside of the experiment itself.

One reason for the increased prevalence of randomized experiments is that they are often considered the "gold standard" for establishing causal relationships between variables. Because participants are randomly assigned to treatment and control, we would generally expect that the only difference between the two groups are the treatment itself. As a result, randomized experiments are free from bias due to confounding variables, and it can be reasonably inferred that an observed difference between the two groups is attributable to the treatment.

Randomized experiments also allow for design-based inference; that is, the act of randomization largely justifies the statistical assumptions made (for a discussion, see Imai et al. (2009) and Imbens (2010)). Fisher (1935) proposed the use of permutation inference for testing the sharp null hypothesis (*i.e.*, that the effect of treatment is zero for all participants), showing that exact inferences can be made with no assumptions beyond randomization itself. It is possible to invert tests of the sharp null hypothesis to obtain an estimate for the treatment effect (*e.g.*, Rosenbaum (2002)); however, an analyst may instead wish to estimate the effect of treatment directly.

In this dissertation, we focus on design-based methods for estimating treatment effects. In particular, we work under a potential outcomes framework often referred to as the Neyman-Rubin model, a non-parametric model which was first introduced by Neyman

(Splawa-Neyman et al. (1990); translation of the original 1923 paper) and further developed by Rubin (1974). Under this model, each participant is assumed to have two potential outcomes: these represent the outcomes that the participant would experience if assigned to treatment and control. Each participant is randomly assigned to either treatment or control, and we observe one of the two potential outcomes based on the treatment assignment. One advantage of this potential outcomes framework is that it provides a clear definition of the treatment effect. Individual treatment effects are defined as the difference between the potential outcomes for each participant, and the average treatment effect is the mean of the individual treatment effects.

While it is impossible to observe both potential outcomes for every participant, randomization allows us to obtain an unbiased estimate for the average treatment effect. As noted above, we can simply compare the observed outcomes for the treated and control units, and attribute any differences to the treatment. For example, Neyman shows that taking a difference between the means of the outcomes for the treated and control groups results in unbiased estimate for the average treatment effect. While this simple difference in means is unbiased, it does not account for information from pretreatment covariates (variables that were measured prior to treatment assignment). By randomizing the participants, we would expect the covariates to be balanced between the treatment groups. However, small imbalances will still occur. For example, a researcher may observe after randomization that the treatment group is older on average than the control group. It may be possible improve the performance of the treatment effect estimate by adjusting for these imbalances.

Modern experiments often have a large number of covariates to choose from. This provides the potential for a substantial improvement in precision by using covariate adjustment, but also presents new challenges. In Chapter II, we present a flexible covariate adjustment method, the LOOP ("Leave-One-Out Potential outcomes") estimator, to address these concerns. Specifically, it can be unclear which covariates to use, and an overly aggressive adjustment may actually harm the performance relative to the unadjusted estimator. When the

number of covariates exceeds the sample size, commonly used covariate adjustment methods like linear regression may fail. An analyst looking to use such methods would be required to perform variable selection first, leading to concerns of data snooping or about the validity of post-selection inferences. In some cases, such as in medical trials, the analysis protocol must be pre-specified to avoid data snooping. We discuss another example where the analyses were required to be specified in advance in Chapter II. In these cases, the researchers need to either choose specific variables ahead of time or to use a covariate adjustment method that allows for automatic variable selection.

Our proposed method is design-based, yet still allows for the use of models to improve precision. We leave out each observation and impute its potential outcomes using the remaining observations. This imputation can be done via any prediction algorithm, such as random forests or linear regression. Importantly, the prediction algorithm chosen does not need to be "correct." So long as the prediction method improves over mean imputation, the LOOP estimator will improve performance over the unadjusted estimator. In addition, model selection occurs in a "black box," so any post-selection inference remains valid. In particular, the method allows for automatic variable selection, so one need not know which covariates to use ahead of time. The method is also unbiased, and it generally performs no worse than the simple difference-in-means estimator, but can often substantially improve performance.

In Chapter III, we propose a covariate adjustment method for paired experiments. While covariate adjustment is one approach for addressing covariate imbalance, researchers may also choose to use study design. In paired experiments, participants are organized into pairs prior to treatment assignment, and then one participant in each pair is randomly assigned to treatment. Ideally, the two participants in each pair would be as similar as possible. However, even if the paired design is effective at balancing covariates between the treatment and control groups, it may be helpful to make adjustments for any remaining imbalances.

The LOOP estimator assumes that the treatment assignments of the participants are

independent (*i.e.*, Bernoulli randomization). We propose extensions to the standard LOOP estimator for completely randomized and block randomized experiments. This approach also works for paired experiments; however, it does not fully take advantage of the paired structure. We therefore propose another method specifically for covariate adjustment in paired experiments. Like the LOOP estimator, this method uses sample splitting; we leave each pair out and impute its potential outcomes using the remaining pairs. This method retains the advantage of the LOOP estimator, while also addressing an issue specific to paired experiments. More specifically, it can be unclear the extent to which we should factor in the paired structure when making adjustments. We address this issue by imputing two sets of potential outcomes, one set that accounts for the pair assignments for the left out pairs and one that does not. We then interpolate between these two sets of potential outcomes to obtain a final estimate.

Another feature of modern randomized experiments is that they can be accompanied by a large amount of auxiliary data. For example, medical trials may be supplemented by health care data for other patients, and experiments within the ASSISTments TestBed are supplemented by data from the remainder of the ASSISTments platform. While randomized experiments are often limited in sample size due to practical constraints, these auxiliary data sets can be quite large. However, unlike randomized experiments, observational data suffer from confounding bias. In Chapter IV, we propose a design-based method that builds on the LOOP estimator for making adjustments to a randomized experiment using an external data set. The goal of this method is to take advantage of the larger sample size of the external data to improve precision, while ensuring that confounding bias from the observational data does not leak into the treatment effect estimate for the randomized experiment. However, it may not be clear whether the external adjustment would improve precision over an adjustment using only the covariates within the experiment. To address this concern, we take a similar approach to the method for pairs. We impute a set of potential outcomes using the external data set and another using the experimental covariates, and interpolate between the two sets

of imputed outcomes.

In Chapter V, we introduce a flexible algorithm for high-dimensional classification, the tournament classifier. This method is only tangentially related to the methods introduced in the rest of the dissertation. However, it does share some similarities with the other methods, such as the use of sample splitting to improve performance and a similar interpolation method to the one used in Chapters III and IV. We build a classifier by having the predictors within a data set compete. In each round of the tournament, we form groups of predictors then combine the predictors within each group into a single predictor. This process continues until all of the predictors have been combined into a single predictor. We suggest a specific approach for the tournament in this dissertation that is particularly suited for sparse high-dimensional data. Finally, while the method can be used as a general classification method, it is also well suited for use within the LOOP estimator computationally. As we will discuss in Chapter II, random forests are a natural fit for the LOOP estimator due to both performance and computational efficiency. Rather than fitting a separate random forest for each left out observation, we can use the out-of-bag predictions instead. The tournament classifier can be modified to take a similar approach without sacrificing computational efficiency.

CHAPTER II

The LOOP Estimator: Adjusting for Covariates in Randomized Experiments

2.1 Introduction

It is common when analyzing randomized controlled trials to adjust for small imbalances in baseline covariates in order to improve the precision of the treatment effect estimate.[1] To avoid the possibility of data snooping, and to ensure the validity of statistical inference, several authors have advocated that the statistical methods be fully specified in advance and reported in the trial protocol (*e.g.*, Begg et al. (1996) and Schulz et al. (2010)).[2] However, in cases where the analysis methods must be pre-specified, it can be unclear which covariates should be used and if covariate adjustment will even be helpful. An overly aggressive adjustment that adjusts for too many covariates can hurt precision more than it helps (*e.g.*, Freedman (2008) and Miratrix et al. (2013)).

[1]Depending on the statistical model being used, these adjustments may also be viewed as adjusting for conditional bias (*i.e.*, bias due to the realized allocation of treatment and resulting covariate imbalance). However, in the model we will consider, the treatment assignment vector T is the only source of randomness. The experimental units, their covariates, and their potential outcomes are all modeled as fixed. Conditioning on T therefore removes all randomness and fixes the treatment effect estimate. For this reason, although the covariate adjustment method we present may be viewed in spirit as adjusting for conditional bias, our discussion will be in terms of improved precision.

[2]Other authors have suggested different ways to ensure valid post-selection inferences; for example, Berk et al. (2013a) introduce a method for valid post-selection confidence intervals and Lee et al. (2016) propose a general framework for valid inference after model selection. For a further discussion on data snooping when analyzing experimental data, see Mutz et al. (2018).

A second concern when adjusting for baseline covariates is bias. Statisticians often allow for biased estimates in order to reduce the overall mean squared error, and many common methods for covariate adjustment do introduce a small amount of bias. However, in some cases practitioners may find exact unbiasedness inherently desirable for various reasons. We discuss one such example in Section 3.4.1. Spiess (2018) presents another argument for unbiasedness when analyzing randomized experiments.

In this chapter, we propose a covariate adjustment method, the LOOP ("Leave-One-Out Potential outcomes") estimator, to simultaneously address both the concerns discussed above. The method is unbiased and model selection occurs in a "black box," so any post-selection inference remains valid. In particular, the method allows for automatic variable selection, so one need not know which covariates to use ahead of time. This method is also design-based, meaning that the experimental randomization largely justifies the statistical assumptions, and it generally performs no worse than the simple difference-in-means estimator, but can often substantially improve performance.

The chapter is organized as follows. Section 2.2 reviews the covariate adjustment literature and relates our method to other estimators. Section 2.3 discusses the randomized trial that motivates our work; in this example, both model selection and bias were concerns. Section 2.4 introduces notation and assumptions and discusses the simple difference-in-means and LOOP estimators. Section 2.5 relates the LOOP estimator to post-stratification and the simple difference-in-means estimator. In Section 2.6, we provide an estimate of the variance. Section 2.7 discusses how to modify the procedures to account for different experimental designs such as block designs. In Section 2.8, we apply the LOOP estimator to examples using simulated data and real experimental data. Section 2.9 concludes.

2.2 Relation to Prior Literature

One of the virtues of randomized experiments is that the physical act of randomization largely justifies the statistical assumptions of the Neyman-Rubin model, a non-parametric

model which was first introduced by Neyman (Splawa-Neyman et al. (1990); translation of the original 1923 paper) and further developed by Rubin (1974). Covariate adjustment is often done through linear regression; however, the standard OLS model is quite different from the Neyman-Rubin model and randomization fails to justify the standard assumptions of OLS. In fact, the OLS estimate is biased under the Neyman-Rubin model; see Freedman (2008) and Lin (2013) for further discussion on OLS adjustments. Other types of regression adjustments can be used: Berk et al. (2013b) build on the work of Freedman (2008) and Lin (2013), while Bloniarz et al. (2016) propose the use of lasso adjustments when the number of covariates is large, especially when the number of covariates exceeds the number of experimental units. In addition, regression adjustments can be used to analyze randomized experiments besides treatment-control studies (*e.g.*, Lu (2016)).

Various other covariate adjustment methods have been proposed, including several that are explicitly design-based. For example, post-stratification (Holt and Smith, 1979) is an adjustment made by stratifying on a pretreatment variable, estimating the treatment effect within each stratum, and taking the weighted average over all strata. Miratrix et al. (2013) explore the properties of the post-stratified estimator under the Neyman-Rubin model. Koch et al. (1982, 1998) propose a method that tests Fisher's sharp null hypothesis (*i.e.*, that all individual treatment effects are zero). They compute the covariance matrix of the treatment and covariates under the sharp null and note that a quadratic form involving this covariance matrix has an approximate χ^2 distribution, which they use to obtain a p-value. Rosenbaum (2002) introduces a similar covariate adjustment method that involves inverting hypothesis tests of the sharp null to obtain an estimate of the treatment effect. Rosenbaum's method is quite flexible and allows for automatic variable selection; however, it assumes a constant treatment effect across units. In this chapter, we propose the LOOP estimator, which is also design-based and allows for automatic variable selection. Unlike Rosenbaum, we do not assume a constant treatment effect.

Aronow and Middleton (2013) introduce another design-based estimator, which is re-

lated to the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). This estimator involves the estimation of a function of the covariates such that the function is predictive of the outcome, resulting in a reduction in variance. In addition, so long as this function is independent of the treatment assignment, the resulting estimate of the average treatment effect will be unbiased. Following a result from Williams (1961), Aronow and Middleton (2013) suggest sample splitting to ensure independence when estimating the function of the covariates. However, many of their calculations assume that the function is a constant fixed in advance and not estimated using a sample splitting procedure. In this chapter, we propose a special case of Aronow and Middleton's estimator with a sample splitting approach. We successively leave out each observation and then impute that observation's treatment and control potential outcomes using a prediction algorithm, such as a random forest (Breiman, 2001).

Our work is similar to that of Wager et al. (2016), who also propose a set of estimators that build on the work of Aronow and Middleton. Wager et al. propose the use of sample splitting and machine learning methods to impute potential outcomes. They also provide a variance estimate, but work under a model in which they assume that the experimental units are drawn from a superpopulation and focus primarily on the population average treatment effect. In this chapter, we assume that the potential outcomes and the covariates are fixed and that the only source of randomness is in the treatment assignment. While the point estimate for the average treatment effect need not change under this model, variance estimation is different, and we derive an estimate for the variance of the LOOP estimator under this framework. Note that we focus specifically on the case where the sample splitting is a leave-one-out procedure. As we will show later, this allows for direct comparison to traditional estimators such as a simple difference-in-means and post-stratification.

Our method is also related to the augmented inverse probability weighted ("AIPW") estimator, which was proposed and developed in Robins et al. (1994), Robins (2000), and in Scharfstein et al. (1999) to estimate treatment effects in observational studies with missing

data. Like the estimator proposed by Aronow and Middleton (2013), AIPW can be considered an extension of the Horvitz-Thompson estimator: it involves a difference in means (inversely weighted by the propensity score) and a regression adjustment based on the expectation of the outcome conditional on the covariates and treatment assignment. See also Chernozhukov et al. (2018) for a related estimator, which employs both sample splitting and machine learning methods to estimate the treatment effect in a high-dimensional setting.

Several other methods use an AIPW-like estimator specifically in randomized experiments (for example, Tsiatis et al. (2008), Spiess (2018), and Rothe (2018)). Tsiatis et al. separate the modeling of covariate-outcome relationships and the evaluation of the treatment effect in order to ensure valid inference after variable selection. Other methods have been proposed to ensure valid post-selection inferences. For example, Moore and van der Laan (2009) use targeted maximum likelihood estimation to make covariate adjustments when the outcome is binary. This method involves modeling the probability that the outcome will be 0 or 1 conditional upon the covariates and the treatment assignment. One can use any procedure to model these conditional probabilities, including methods with automatic variable selection. Steingrimsson et al. (2017) give recommendations for the use of targeted maximum likelihood estimation in practice.

2.3 Motivation

Our work is motivated by a so-called "pay for success" program in the state of Illinois. In brief, a pay for success program is one in which a government contracts an outside organization to provide needed services, but only pays the organization if the services are shown to be effective, typically in a randomized controlled experiment. In our example, the contracted organization is to provide special social services to at-risk youth, and one metric for success (among others) is a reduction in the number of days spent in juvenile detention. Success of the program will be evaluated according to the results of a six year experiment in which eligible youth are randomly selected to receive either the special services or ordinary

care. The evaluation will be conducted by researchers in the School of Social Work at the University of Michigan, and we assisted the evaluators in planning the design and analysis of the experiment.

Several hundred youth are expected to take part in the program. Eligible participants are independently randomized to treatment or control, each with probability 1/2. More elaborate designs were considered, but were too logistically challenging. A key difficulty is the fact that the participants enter into the experiment continually over time, making designs such as blocking infeasible.

Several baseline covariates will be available, at least some of which (*e.g.*, age) are known to be highly predictive of outcome. The interested parties (the state, the outside organization providing the services, and the evaluators) agreed that some form of adjustment for these covariates would be desirable. However, there was initially no clear consensus on which adjustment procedure to use.

One concern was bias. Unbiasedness was felt to be desirable, perhaps more so in this example than in many others, because the state's payment rate will be directly proportional to the estimated size of the treatment effect. Any bias in the estimator therefore effectively results in a bias in the payment. Indeed, one high ranking state official was opposed to any amount of bias, even if it might reduce the mean squared error. To paraphrase, the magnitude of the error was not so much a concern, as long as it was a fair bet. Other officials were open to using a biased estimator, so long as the bias was negligible. Critically, however, it was felt that the bias should still be quantified, and in the case of biased estimators, it was unclear how to produce a concrete number for the bias. For this reason as well, an unbiased estimator was preferred. Ultimately, it was decided to use post-stratification.

A second concern was which covariates to adjust for. It was required to fully specify the analysis protocol in advance. Many potential covariates were available; however, adjusting for too many covariates could result in overadjustment, leading to inflated variance. Post-stratification is especially sensitive to overadjustment and considerable discussion was

required to come to a consensus on both the number of covariates and which specific covariates to be used.

The challenges outlined above motivate our work: we wish to produce a method that provides automatic variable selection in order to eliminate the guesswork in deciding which covariates to use, while remaining exactly unbiased under the Neyman-Rubin model.

2.4 The LOOP Estimator

In this section, we introduce the LOOP ("Leave-One-Out Potential outcomes") estimator, which we can use to obtain an unbiased estimate of the average treatment effect while adjusting for covariates.

2.4.1 Model and Notation

Consider a randomized controlled experiment in which there are N participants, indexed by $i = 1, 2, ..., N$. Each participant is randomly assigned to either treatment or control, and we let T_i denote the i-th participant's treatment assignment, such that $T_i = 1$ if the i-th participant is assigned to treatment and $T_i = 0$ if the i-th participant is assigned to control. For each participant, we observe (in addition to the treatment assignment T_i) a response variable Y_i and a q-dimensional vector of baseline covariates Z_i. We assume Bernoulli treatment assignments, *i.e.*,

$$T_i \perp\!\!\!\perp T_j$$

for $i \neq j$. We let p_i denote the i-th participant's probability of being assigned to treatment, *i.e.*,

$$p_i = P(T_i = 1)$$

and assume $0 < p_i < 1$. In some parts of this chapter, we assume for simplicity (and without much loss of generality) that $p_i = p$ for all i and for some fixed constant p, but for now we

explicitly let p_i vary from subject to subject.

Associated with each of the N participants are two fixed (non-random) potential outcomes, t_i and c_i. We assume that we observe t_i if participant i is assigned to treatment and c_i if participant i is assigned to control. That is, the observed outcome Y_i for participant i is

$$Y_i = T_i t_i + (1 - T_i) c_i.$$

We define the individual treatment effect τ_i as

$$\tau_i = t_i - c_i$$

and the average treatment effect $\bar{\tau}$ as

$$\bar{\tau} = \frac{1}{N} \sum_{i=1}^{N} \tau_i \tag{2.1}$$

which is our primary parameter of interest.

Lastly, some additional notation. Let $\mathcal{T} = \{i : T_i = 1\}$ and $\mathcal{C} = \{i : T_i = 0\}$. Let n be the (random) number of participants assigned to treatment and $N - n$ be the number assigned to control. For each participant, we define the important quantity m_i as

$$m_i = (1 - p_i) t_i + p_i c_i.$$

Note that when $p_i = \frac{1}{2}$, this is simply the mean of t_i and c_i. We will use the notation $\hat{m}_i$ to denote an estimate of m_i. Finally, we define the (signed) inverse probability weights U_i as

$$U_i = \begin{cases} 1/p_i, & T_i = 1 \\ -1/(1 - p_i), & T_i = 0 \end{cases}$$

and note that U_i has expectation 0.

2.4.2 Average and Individual Treatment Effects

It is not possible to observe any single participant's treatment effect τ_i, because for each participant we are only able to observe the treatment response t_i or the control response c_i. However, it is well known that the average treatment effect $\bar{\tau}$ can be estimated. We define the *simple difference estimator* $\hat{\tau}_{sd}$ to be the difference of the average of the observed treatment responses and the average of the observed control responses:

$$\hat{\tau}_{sd} = \frac{1}{n}\sum_{i\in\mathcal{T}} Y_i - \frac{1}{N-n}\sum_{i\in\mathcal{C}} Y_i. \tag{2.2}$$

This provides an unbiased estimate of the average treatment effect (conditional on $0 < n < N$).

It is also possible to provide an unbiased estimate of an individual participant's treatment effect τ_i. For example, Y_iU_i is one such estimator:

$$Y_iU_i = \begin{cases} t_i/p_i, & T_i = 1 \\ -c_i/(1-p_i), & T_i = 0 \end{cases}$$

and thus

$$\begin{aligned} \mathbb{E}(Y_iU_i) &= \frac{t_i}{p_i}P(T_i = 1) + \frac{-c_i}{1-p_i}P(T_i = 0) \\ &= t_i - c_i. \end{aligned}$$

Although this is an unbiased estimator of τ_i, it generally has very high variance and is therefore not useful for practical purposes. Suppose, for example, that $p_i = 1/2$. Then if participant i is assigned to treatment we would estimate his treatment effect as $2Y_i$, and if he was assigned to control we would estimate his treatment effect as $-2Y_i$.

As an alternative estimator of τ_i, consider

$$\hat{\tau}_i = (Y_i - \hat{m}_i)U_i. \tag{2.3}$$

If $\hat{m}_i$ is independent of U_i — that is, if $\hat{m}_i$ is independent of the i-th participant's treatment assignment — then $\hat{\tau}_i$ is an unbiased estimator of τ_i:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_i) &= \mathbb{E}\left[(Y_i - \hat{m}_i)U_i\right] \\
&= \mathbb{E}(Y_i U_i) - \mathbb{E}(\hat{m}_i)\mathbb{E}(U_i) \\
&= \tau_i
\end{aligned}$$

where in the last line we use the fact that $\mathbb{E}(U_i) = 0$. The advantage of this estimator is that it will have a low variance as long as $\hat{m}_i \approx m_i$. To see why, suppose that $\hat{m}_i = m_i$ exactly. Then

$$(Y_i - m_i)U_i = \begin{cases} (t_i - m_i)/p_i, & T_i = 1 \\ (-c_i + m_i)/(1 - p_i), & T_i = 0 \end{cases}$$

but both $(t_i - m_i)/p_i$ and $(-c_i + m_i)/(1 - p_i)$ work out to be τ_i, and thus $\hat{\tau}_i$ is not only unbiased but also has zero variance. When $\hat{m}_i$ only approximately equals m_i, then the variance of $\hat{\tau}_i$ is no longer zero but is small. More precisely, in Section 2.6 we show that

$$\mathrm{Var}(\hat{\tau}_i) = \frac{1}{p_i(1 - p_i)}\mathbb{E}\left[(\hat{m}_i - m_i)^2\right].$$

To summarize, $\hat{\tau}_i$ will be unbiased and have low variance as long as: (a) $\hat{m}_i$ is independent of T_i; and (b) $\hat{m}_i$ is a good estimator of m_i.

Finally, note that $\hat{m}_i$ in equation (2.3) plays the same role as the "augmented" portion of the AIPW estimator as described by Lunceford and Davidian (2004) and the function of the covariates in the estimator of Aronow and Middleton (2013).

2.4.3 Leave-One-Out Imputation

We now define the LOOP estimator of the average treatment effect $\bar{\tau}$ as:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_i \tag{2.4}$$

where $\hat{\tau}_i$ is defined as in (2.3) and where $\hat{m}_i$ is obtained as follows. For each i, we drop observation i and use the remaining $N-1$ observations to impute t_i and c_i, using any method of our choosing (*e.g.*, linear regression, random forests, etc.). Having obtained estimates $\hat{t}_i$ and $\hat{c}_i$, we then set

$$\hat{m}_i = (1 - p_i)\hat{t}_i + p_i\hat{c}_i. \tag{2.5}$$

As an example, suppose we wish to estimate $\hat{m}_i$ using linear regression. For each i, we would drop observation i and then regress Y on T and Z using only the remaining $N - 1$ observations. We would then calculate $\hat{t}_i$ and $\hat{c}_i$ using the fitted model, plugging in Z_i for the covariates, and compute $\hat{m}_i$ as in (2.5).

Because we leave out the i-th observation when we compute $\hat{m}_i$, it follows that T_i and $\hat{m}_i$ are independent and thus that $\hat{\tau}_i$ is unbiased. It immediately follows that $\hat{\tau}$ is also unbiased. This will be true no matter how we estimate t_i and c_i, as long as we leave out observation i so that $\hat{t}_i$ and $\hat{c}_i$ are independent of T_i. Importantly, note that we impute both t_i and c_i, even though one of them is actually observed and therefore known. If we were to use the true observed value, then $\hat{m}_i$ would no longer be independent of T_i.

It is worth noting that although we use the individual treatment effect estimates $\hat{\tau}_i$ in this chapter simply as an intermediate step in the estimation of the average treatment effect $\bar{\tau}$, these individual treatment effect estimates may be useful for other purposes as well, such as in estimating treatment effect heterogeneity. Athey and Imbens (2016) and Nie and Wager (2017) use similar formulations for estimating heterogeneous treatment effects. With this in mind, we summarize below three useful facts about $\hat{\tau}_i$, the latter two of which we show in

Section 2.6:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_i) &= \tau_i \\
\text{Var}(\hat{\tau}_i) &= \frac{1}{p_i(1-p_i)}\mathbb{E}\left[(\hat{m}_i - m_i)^2\right] \\
\text{Cov}(\hat{\tau}_i, \hat{\tau}_j) &= \text{Cov}(\hat{m}_i U_i, \hat{m}_j U_j). \qquad (2.6)
\end{aligned}$$

The covariance term $\text{Cov}(\hat{m}_i U_i, \hat{m}_j U_j)$ is usually negligible and can be ignored in most applications (note that U_i and U_j are independent).

2.5 Imputing the Potential Outcomes

In the subsequent sections, we propose several methods for imputing the potential outcomes in order to estimate m_i. First, we impute the potential outcomes without making use of covariates, simply taking the mean of the observed outcomes in each treatment group. When we do this, we see that the LOOP estimator is exactly equal to the simple difference estimator. We also impute the potential outcomes using decision trees and discuss the connection between post-stratification and the LOOP estimator. Finally, we propose the use of random forests, which may provide an improvement over post-stratification and allow us to take advantage of automatic variable selection.

2.5.1 Imputing Potential Outcomes Ignoring Covariates: LOOP equals the Simple Difference Estimator

In this section, we impute the potential outcomes without making use of covariates. We simply take the mean of the observed outcomes in the treatment group (excluding observation i) to estimate t_i and the mean of the observed outcomes in the control group (excluding

observation i) to estimate c_i. That is, we estimate t_i and c_i as:

$$\hat{t}_i = \frac{\sum_{k \in \mathcal{T} \setminus i} Y_k}{n - T_i} \tag{2.7}$$

$$\hat{c}_i = \frac{\sum_{k \in \mathcal{C} \setminus i} Y_k}{(N - n) - (1 - T_i)}. \tag{2.8}$$

If the assignment probabilities are all equal, *i.e.*, if $p_i = p$ for all i and for some fixed p, then the LOOP estimator is exactly equivalent to the simple difference estimator, as we show in Appendix A. As a result of this equivalence, we conclude that in practice the LOOP estimator will typically perform no worse, or at least not that much worse, than the simple difference estimator. More precisely, in Section 2.6 we show that the variance of the LOOP estimator is directly related to the mean squared error of the $\hat{m}_i$ terms. Thus the LOOP estimator will outperform the simple difference estimator as long as we improve the imputation of the potential outcomes beyond this baseline approach (mean imputation). In addition, the equivalence between the LOOP estimator and the simple difference estimator provides us with some reassurance that the leave-one-out procedure does not inherently introduce extra variance.

2.5.2 Imputing Potential Outcomes using Decision Trees: LOOP equals Post-stratification

In this section, we discuss the connection between the LOOP estimator and post-stratification. Post-stratification is a covariate adjustment method made by stratifying on pretreatment variables, estimating the treatment effect within each stratum by taking a simple difference in means, and then taking the weighted average over all strata (Holt and Smith, 1979). We argue that when we impute potential outcomes using a decision tree (see James et al. (2013) for a summary of decision trees), the LOOP estimator is equivalent to post-stratification.

Given a single decision tree (fixed in advance), we impute the potential outcomes as follows. First, we assign each observation i to a group; this is done by applying the decision

tree to observation i's covariates. (This group may be viewed as a "leaf" or a "stratum.") For each i, we then impute t_i using the average observed outcome of the treated units within the same group (excluding observation i itself). We impute c_i similarly. Thus, using the same argument given above in Section 2.5.1, it is simple to show that the average of the $\hat{\tau}_i$ within a group is equal to the simple difference within that group. Thus, the average of all the $\hat{\tau}_i$ is a weighted average of the within-group simple differences, *i.e.*, it is a post-stratification estimator.

2.5.3 Imputing Potential Outcomes using Random Forests

In their analysis of post-stratification, Miratrix et al. (2013) show that post-stratification is nearly as efficient as blocking. However, one disadvantage of post-stratification is that we must be parsimonious in the number of variables selected. If we include too many covariates, we end up partitioning our data too finely. We can overcome this limitation and also improve on the post-stratified estimate using the LOOP estimator. One advantage of the LOOP estimator is that estimation of m_i is very flexible. One can impute the potential outcomes using any method, so long as $\hat{m}_i$ and T_i are independent. In particular, we can use ensemble methods such as boosting or bagging to improve our estimates over a single decision tree.

One such method is the random forest algorithm, and random forests will be our method of choice for imputing the potential outcomes for the remainder of the chapter. For a description of tree-based methods, including random forests, see James et al. (2013). In order to impute the potential outcomes using random forests, we could first omit observation i, and then create a random forest using the remaining $N-1$ observations, which we could use to impute t_i and c_i. However, doing this for each i would be computationally demanding. Fortunately, it is also unnecessary. Because we are using a leave-one-out procedure, and because out-of-bag predictions are essentially leave-one-out predictions, we can simply make use of the out-of-bag predictions. To clarify, random forests are an ensemble of many decision

trees, each of which is constructed using a bootstrap sample. In fitting any given tree, some number of observations will be left out. The out-of-bag prediction for the i-th observation is the prediction made using the trees that do not include observation i and is effectively a leave-one-out prediction. We can therefore fit just two random forests (one on the treatment units and one on the control units) and impute the potential outcomes using the out-of-bag predictions. By contrast, when imputing the potential outcomes using many other methods, such as OLS, we do need to create a separate model for each i. As a result, imputing the potential outcomes with random forests can be relatively computationally efficient.

Because random forests are typically an improvement over individual decision trees, they allow us to obtain a more precise estimate of the average treatment effect $\bar{\tau}$. By using random forests to effectively improve upon post-stratification, we might even hope to obtain an estimate of $\bar{\tau}$ that works as well as or better than if we had used a blocked experimental design. Moreover, random forests essentially provide automatic variable selection, making it unnecessary to decide in advance which covariates should be used. Biau (2012) shows that the rate of convergence of the random forest algorithm depends on the number of important variables present, rather than how many noise variables there are. Given these properties and the computational efficiency of random forests, we see that random forests are naturally suited for the LOOP estimator.

2.6 Variance Estimation

Aronow and Middleton (2013) give a conservative estimate of the variance of the Horvitz-Thompson estimator. They also provide an estimate for the variance of their own estimator; however, this estimate is derived under the assumption that the function of the covariates (*i.e.*, our $\hat{m}_i$) is a constant fixed in advance, not computed from the data. Wager et al. (2016) provide a variance estimate for their method, but assume that the experimental units are drawn from a superpopulation. Under a superpopulation model, an estimate for the sample average treatment effect is used to estimate the population average treatment effect,

and there are two sources of variation: the random treatment assignment and the random sampling of the experimental units from the superpopulation. The target variance in this case is the unconditional variance for the treatment effect estimate. Wager et al. (2016) use a jackknife approach to estimate the variance in this setting.

In this section, we derive an estimate for the variance of the LOOP estimator working under a finite population model and assuming that the treatment assignment is the only source of randomness. In this case, we estimate the variance for the treatment effect estimate conditional on the specific experimental sample. That is, the variance does not include a component for the random sampling of units from a superpopulation. In addition, consistent estimation of the variance of the treatment effect estimate is not generally possible due to the non-identifiability of the covariance of the potential outcomes (for example, see Splawa-Neyman et al. (1990) and Aronow et al. (2014)). We instead focus on obtaining a conservative estimate for the variance. Because it estimates an additional source of variation, the jackknife approach used by Wager et al. (2016) will be excessively conservative in some cases (for example, in the presence of treatment effect heterogeneity). Here we provide a different estimate for the variance of our estimator. In Section 2.6.1 we calculate the true variance of $\hat{\tau}$, and then in Section 2.6.2, we produce an estimate.

2.6.1 Variance of $\hat{\tau}$

In Appendix B.1, we show that:

$$\mathrm{Var}(\hat{\tau}_i) = \frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i) \tag{2.9}$$

and that

$$\gamma_{ij} = \mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \rho_{ij}\sqrt{\frac{\mathrm{Var}(\hat{m}_i)\mathrm{Var}(\hat{m}_j)}{p_i p_j (1-p_i)(1-p_j)}} \tag{2.10}$$

where

$$\rho_{ij} = \text{Corr}(\hat{m}_i U_i, \hat{m}_j U_j).$$

Combining (2.9) and (2.10) yields:

$$\text{Var}(\hat{\tau}) = \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{1}{p_i(1-p_i)}\text{MSE}(\hat{m}_i) + \sum_{i\neq j}\gamma_{ij}\right]. \tag{2.11}$$

Limiting our attention to the special case that $p_i = p$ for all i,

$$\begin{aligned}\text{Var}(\hat{\tau}) &= \frac{\overline{\text{MSE}}}{Np(1-p)} + \frac{\sum_{i\neq j}\gamma_{ij}}{N^2} \\ &= \frac{\overline{\text{MSE}}}{Np(1-p)} + \frac{(N-1)\bar{\gamma}}{N}\end{aligned} \tag{2.12}$$

where

$$\overline{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{m}_i)$$

and

$$\bar{\gamma} = \frac{1}{N(N-1)}\sum_{i\neq j}\gamma_{ij}.$$

2.6.2 Estimating the Variance

In Appendix B.2, we show that when $p_i = p$ for all i,

$$\frac{\overline{\text{MSE}}}{Np(1-p)} \leq \frac{1}{N}\left[\frac{1-p}{p}M_t + \frac{p}{1-p}M_c + 2\sqrt{M_t M_c}\right] \tag{2.13}$$

where

$$M_t = \frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{t}_i)$$

and

$$M_c = \frac{1}{N}\sum_{i=1}^{N} \mathrm{MSE}(\hat{c}_i).$$

We estimate M_t and M_c by leave-one-out cross validation:

$$\hat{M}_t = \frac{1}{Np}\sum_{i\in\mathcal{T}}(\hat{t}_i - t_i)^2 \tag{2.14}$$

$$\hat{M}_c = \frac{1}{N(1-p)}\sum_{i\in\mathcal{C}}(\hat{c}_i - c_i)^2. \tag{2.15}$$

In Appendix B.3, we show that these estimates are unbiased. We plug (2.14) and (2.15) into the bound (2.13) to obtain an estimate for the first term in (2.12):

$$\frac{1}{N}\left[\frac{1-p}{p}\hat{M}_t + \frac{p}{1-p}\hat{M}_c + 2\sqrt{\hat{M}_t\hat{M}_c}\right]. \tag{2.16}$$

Next, we provide an unbiased estimator of γ_{ij} (and thus, $\bar{\gamma}$) in Appendix B.4. Specifically, we have:

$$\hat{\gamma}_{ij} = \begin{cases} \frac{(1-p)^2}{p^2}(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}), & T_i = T_j = 1 \\ (\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}), & T_i = 0, T_j = 1 \\ (\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}), & T_i = 1, T_j = 0 \\ \frac{p^2}{(1-p)^2}(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}), & T_i = T_j = 0 \end{cases} \tag{2.17}$$

where $\hat{t}_i^{-j}$ is an estimate of t_i excluding the j-th observation (in addition to the i-th observation). We let $\hat{t}_i^{+j}$ denote an estimate of t_i including the j-th observation and assuming $T_j = 1$. Note that this is only calculable when $T_j = 1$, in which case $\hat{t}_i^{+j} = \hat{t}_i$. We define $\hat{c}_i^{-j}$ and $\hat{c}_i^{+j}$ similarly. We plug this estimate, $\hat{\gamma}_{ij}$, into the second term of (2.12) and add to the

plug-in estimator (2.16) for the bound (2.13) to obtain an estimate of the variance of $\hat{\tau}$:

$$\widehat{\mathrm{Var}}(\hat{\tau}) = \frac{1}{N}\left[\frac{1-p}{p}\hat{M}_t + \frac{p}{1-p}\hat{M}_c + 2\sqrt{\hat{M}_t\hat{M}_c}\right] + \frac{1}{N^2}\sum_{i\neq j}\hat{\gamma}_{ij} \tag{2.18}$$

2.6.3 Estimating the Variance in Practice

In practice, we recommend making two modifications when estimating the variance. First, we recommend estimating M_t and M_c as

$$\tilde{M}_t = \frac{1}{n}\sum_{i\in\mathcal{T}}(\hat{t}_i - t_i)^2$$

and

$$\tilde{M}_c = \frac{1}{N-n}\sum_{i\in\mathcal{C}}(\hat{c}_i - c_i)^2,$$

particularly when N is small. Note that these approximations require that $0 < n < N$.

Second, we recommend omitting the second term in (2.18) for computational efficiency. While we estimate m_i using a leave-one-out procedure, γ_{ij} is estimated using a leave-two-out procedure. As a result, estimating γ_{ij} requires us to increase the number of models fit by a factor of N. In addition, the γ_{ij} terms are often negligible, in the sense that $\overline{\mathrm{MSE}}/N$ goes to zero at a rate $1/N$, while $\bar{\gamma}$ goes to zero at a faster rate. In Section 4.4, we provide conditions under which the LOOP estimator is asymptotically normally distributed. By applying an argument used for paired experiments in Section 3.3.2, these conditions also imply that

$$\frac{\sum_{i\neq j}\gamma_{ij}}{\sum_{i=1}^{N}\mathrm{MSE}(\hat{m}_i)} \longrightarrow 0,$$

and thus that $\bar{\gamma}$ is asymptotically negligible.

For example, suppose that under suitable regularity conditions $\mathrm{Var}(\hat{m}_i)$ and $\mathrm{Var}(\hat{m}_j)$ go to zero at rate $1/N$. Then if ρ_{ij} goes to zero (at any rate), γ_{ij} will go to zero faster than $1/N$. Appendix C gives a more formal argument. We also provide simulation results in Section

2.8.3 to demonstrate empirically that $N\bar{\gamma}$ goes to 0 as N increases.

To see why we might expect ρ_{ij} (and likewise $\bar{\rho}$) to go to zero, recall that U_i and U_j are independent. Thus, even if $\hat{m}_i$ and $\hat{m}_j$ are correlated (which they typically will be), ρ_{ij} may still be negligible. Indeed, if $\hat{m}_i$ and $\hat{m}_j$ are perfectly correlated, then $\rho_{ij} = 0$. The only reason for $\hat{m}_i U_i$ and $\hat{m}_j U_j$ to be correlated would be through the dependence of $\hat{m}_i$ on U_j, and of $\hat{m}_j$ on U_i. These dependencies will typically decay as N grows. As an illustrative example, suppose that for all i, $\hat{m}_i$ is a linear estimator, *i.e.*, for some constants $a_{i.k}$

$$\hat{m}_i = a_{i.0} + \sum_{k \neq i} a_{i.k} U_k.$$

In this case, it can be shown (see Appendix C.1) that $\bar{\rho}$ goes to 0 at rate $1/N$; more specifically, we show $\bar{\rho} \leq 1/(N-1)$. Indeed, we further show that if $\hat{m}_i$ is a polynomial function of degree D for all i, then $\bar{\rho} \leq D/(N-1)$. Note that there do exist certain pathological cases where $\bar{\rho}$ can be large. For example, suppose that for all i, $\hat{m}_i = \prod_{k \neq i} U_k$. Then $\hat{m}_i U_i = \prod_{k=1}^{N} U_k$ for all i, so the correlation between $\hat{m}_i U_i$ and $\hat{m}_j U_j$ is exactly 1.

The two modifications discussed in this section yield the following estimate for the variance of $\hat{\tau}$:

$$\widetilde{\mathrm{Var}}(\hat{\tau}) = \frac{1}{N}\left[\frac{1-p}{p}\tilde{M}_t + \frac{p}{1-p}\tilde{M}_c + 2\sqrt{\tilde{M}_t \tilde{M}_c}\right]. \tag{2.19}$$

If there is concern that in a particular application $\bar{\gamma}$ is not negligible — either due to concern that $\bar{\gamma}$ may not go to zero faster than $1/N$ or simply due to concern that N is not large enough — we can instead use (2.18) to estimate the variance of $\hat{\tau}$.

2.6.4 Relationship between $\widetilde{\text{Var}}(\hat{\tau})$ and the Sample Variance

We show in Appendix D that when we impute potential outcomes ignoring covariates (*i.e.*, we calculate $\hat{c}_i$ and $\hat{t}_i$ as in (2.7) and (2.8)),

$$\tilde{M}_t = \frac{n}{n-1} s_t^2 \tag{2.20}$$

and

$$\tilde{M}_c = \frac{N-n}{N-n-1} s_c^2 \tag{2.21}$$

where s_t^2 and s_c^2 are the standard sample variances (of the treated and control units). We show in Appendix D that plugging (2.20) and (2.21) into (2.19) yields the following inequality:

$$\begin{aligned} \widetilde{\text{Var}}(\hat{\tau}) &\leq \left(\frac{n}{Np}\right) \frac{s_t^2}{n-1} + \left(\frac{N-n}{N(1-p)}\right) \frac{s_c^2}{N-n-1} \\ &\approx \frac{s_t^2}{n-1} + \frac{s_c^2}{N-n-1} \end{aligned} \tag{2.22}$$

with equality in (2.22) when $\tilde{M}_t$ and $\tilde{M}_c$ are equal. Thus, our variance estimate provides a result roughly equal to or slightly better than if we had performed a t-test. For a related discussion, see Aronow et al. (2014).

2.6.5 Inference for the Average Treatment Effect

In this section, we have focused primarily on estimating the variance for the LOOP estimator. Ultimately, we wish to use the variance estimate to conduct inference. In this dissertation, we generally rely on normal approximations to obtain p-values and construct confidence intervals. In Section 4.4, we provide conditions that ensure the LOOP estimator will be asymptotically normally distributed. Generally speaking, these conditions state that both the data and the imputation method are sufficiently well-behaved. We demonstrate

that the conditions hold for the case of imputation using simple linear regression in Section 4.4. However, it may be difficult to verify these conditions in other cases (*e.g.*, for random forest imputation).

Another approach would be to do a permutation test on $\hat{\tau}$, which would allow us to obtain exact p-values under Fisher's sharp null hypothesis without any additional assumptions. This approach would be computationally intensive. In addition, we may instead wish to construct confidence intervals for $\bar{\tau}$. In this case, we could use resampling methods, such as the bootstrap (Efron, 1979) or subsampling (Politis et al., 1999). These approaches often assume that the observations are independent and identically distributed, along with some form of convergence for the distribution of the estimator. Subsampling also imposes conditions on the size of the subsamples. Subsampling has several properties that could make it suitable for constructing confidence intervals for the average treatment effect. It requires a weaker form of convergence than bootstrap, and does not require the variance estimate for $\hat{\tau}$ to be consistent. Subsampling can be used to construct confidence intervals while requiring minimal assumptions; however, one downside is that like in the case of permutation inference, resampling methods are computationally intensive. In addition, when using subsampling for inference, the issue of whether we are performing inference conditional on the experimental sample (*i.e.*, targeting the sample or population average treatment effect) again becomes relevant.

2.7 Dependent Treatment Assignments

In the preceding sections, we assumed that the treatment assignments are independent of each other. In this section, we consider study designs in which the treatment assignments are not independent. For example, it is common for researchers to randomly assign a fixed number n of participants to treatment and leave the remaining $N - n$ as controls (*i.e.*, complete randomization). In such cases, treatment assignments are not independent. However, we can ensure the independence of T_i and $\hat{m}_i$ as follows: if the i-th observation is assigned

to treatment, we randomly pick one of the control observations and drop that observation as well as observation i when fitting our imputation model. Conversely, if the i-th observation is control, we randomly drop one of the treatment observations. Thus, regardless of whether T_i is equal to 0 or 1, when we estimate $\hat{m}_i$, we use $N-2$ of the remaining $N-1$ observations. Of these $N-2$ observations, $n-1$ will be assigned to treatment, $N-n-1$ will be assigned to control, and the specific allocation will be independent of T_i.

We give a numerical example to illustrate this "random drop" procedure in Appendix E.1. More specifically, we consider an example with $N = 5$, where $n = 2$ participants are to be assigned to treatment and the remaining 3 to control. For a given observation i, the random drop procedure will result in 1 treated observation and 2 control observations being selected from the remaining observations to calculate $\hat{m}_i$. We show that for an arbitrary set of 1 treated and 2 control observations, the probability the set is selected to calculate $\hat{m}_i$ is the same regardless of the treatment assignment for observation i. Thus the random drop procedure ensures that T_i is independent of $\hat{m}_i$.

We can generalize this argument to show why the random drop procedure ensures the independence of T_i and $\hat{m}_i$ for arbitrary values of N and n. Let $\mathcal{T}_i$ and $\mathcal{C}_i$ denote the indices of the treated and control observations that are used to calculate $\hat{m}_i$. We show that

$$\Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}}, \mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 1\right) = \Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}}, \mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 0\right)$$

for any disjoint subsets $\tilde{\mathcal{T}}$ and $\tilde{\mathcal{C}}$ (of size $n-1$ and $N-n-1$) of $\{1, \ldots, N\} \backslash i$. Consider the case where $T_i = 1$. There are $\binom{N-1}{n-1}$ ways to select the specific set $\tilde{\mathcal{T}}$ from $\{1, \ldots, N\} \backslash i$. Thus

$$\Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}} | T_i = 1\right) = \frac{1}{\binom{N-1}{n-1}}.$$

Conditional on $T_i = 1$ and having selected $\tilde{\mathcal{T}}$, the probability of selecting the set $\tilde{\mathcal{C}}$ is

$1/(N-n)$, as we randomly choose one of the control observations to drop. That is,

$$\Pr\left(\mathcal{C}_i = \tilde{\mathcal{C}} | \mathcal{T}_i = \tilde{\mathcal{T}}, T_i = 1\right) = \frac{1}{N-n}.$$

We then have

$$\begin{aligned}
\Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}}, \mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 1\right) &= \Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}} | T_i = 1\right) \times \Pr\left(\mathcal{C}_i = \tilde{\mathcal{C}} | \mathcal{T}_i = \tilde{\mathcal{T}}, T_i = 1\right) \\
&= \frac{1}{\binom{N-1}{n-1}} \frac{1}{N-n} \\
&= \frac{(n-1)!(N-n)!}{(N-1)!(N-n)} = \frac{(n-1)!(N-n-1)!}{(N-1)!}.
\end{aligned}$$

A similar argument can be used to show that

$$\begin{aligned}
\Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}}, \mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 0\right) &= \Pr\left(\mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 0\right) \times \Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}} | \mathcal{C}_i = \tilde{\mathcal{C}}, T_i = 0\right) \\
&= \frac{1}{\binom{N-1}{N-n-1}} \frac{1}{n} \\
&= \frac{(N-n-1)!(n-1)!}{(N-1)!}.
\end{aligned}$$

Thus

$$\Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}}, \mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 1\right) = \Pr\left(\mathcal{T}_i = \tilde{\mathcal{T}}, \mathcal{C}_i = \tilde{\mathcal{C}} | T_i = 0\right)$$

as desired. It follows that $\hat{m}_i$ is independent of T_i, as the value of $\hat{m}_i$ is determined by the indices in $\mathcal{T}_i$ and $\mathcal{C}_i$.

Because this procedure ensures that $\hat{m}_i$ and T_i are independent, $\hat{\tau}_i$ will remain unbiased. By dropping an extra observation we are losing some information. However, we could repeat this entire procedure many times, producing an unbiased estimate of $\hat{\tau}_i$ each time, which we could then average. In the aggregate, we would then make use of all remaining $N-1$ observations. Note that in practice, the use of the random drop procedure would not change our estimates much. For example, if we use the random drop procedure with a decision tree,

we would still obtain the post-stratified estimate. (See Appendix E.2 for further discussion.)

Note that a similar procedure could be used in a block-randomized experiment, in which a fixed number of participants within each block are assigned to treatment, and the rest to control. In this case, when computing $\hat{m}_i$, we would need to drop an observation that is in the same block as i. This procedure could even be extended to paired designs. In a paired design, both observation i and observation i's pair would need to be dropped. However, all of the remaining observations from the experiment could still be used to produce an estimate of m_i.

Finally, we note that the independence of $\hat{m}_i$ and T_i implies that (2.9) continues to hold. However, (2.10) is no longer valid, due to the dependence of U_i and U_j. Variance estimation in this context may therefore require a modified approach.

2.8 Results

Below, we apply the LOOP estimator (with random forests) to both simulated and actual data. In our first simulation, we compare methods when the treatment effects are either homogeneous or heterogeneous, and also demonstrate the bias of the point estimate and standard error for the OLS estimator. Next, we consider a simulation in which we examine the performance of the LOOP estimator when many of the covariates are not predictive. In our third simulation, we empirically demonstrate that the covariance terms discussed in Section 2.6 are negligible. Finally, we apply the LOOP estimator to the experiment conducted by Barrera-Osorio et al. (2011) on the effects of various cash transfer programs on educational outcomes in Colombia.

2.8.1 Simulation 1: Heterogeneous and Homogeneous Treatment Effects

Consider a randomized experiment in which there are N subjects and there is a single covariate, Z, with three possible values: 0, 1, and 2. For each value of Z, there are $N/3$ subjects and each subject has potential outcomes that are generated from a normal distribution with

means given in Table 2.1 and standard deviation 0.1. We consider two scenarios, one where the treatment effects are heterogeneous and the other with homogeneous treatment effects. We consider four cases: (a) $N = 30$ and heterogeneous treatment effects; (b) $N = 100$ and heterogeneous treatment effects; (c) $N = 30$ and homogeneous treatment effects; (d) and $N = 100$ and homogeneous treatment effects.

Table 2.1: Simulation 1: Potential Outcome Values

Treatment Effects	**Z Value**	**Mean** c_i	**Mean** t_i
Heterogeneous	0	0	1
	1	1	1
	2	1	2
Homogeneous	0	0	1
	1	1	2
	2	1	2

For each of the four cases we do the following. We generate a single set of treatment and control potential outcomes for the N subjects. We then create 100,000 random assignment vectors (T), where the treatment assignments are independent Bernoulli random variables with probability 1/2. For each of these 100,000 treatment assignment vectors, we compute the observed outcomes (Y) and estimate the average treatment effect and nominal standard error.

We compare the results using OLS, the LOOP estimator with random forests, LOOP with OLS imputation, and cross estimation with random forests (Wager et al., 2016). Note that for cross estimation, we use the code provided on GitHub; however, we remove the specified node size parameter. This modification improves performance in the context of this simulation. The bias is estimated as the mean point estimate minus the true ATE. We also show the mean nominal standard error and estimate the true standard error using the standard deviation of the 100,000 point estimates. Similarly, we estimate the mean squared error as the mean squared error (relative to the true average treatment effect) of the 100,000 point estimates. The nominal standard errors for the LOOP estimator are calculated using

Table 2.2: Simulation 1 Results

Method	Bias Est.	Nominal SE	True SE	MSE	Coverage
(a) $N = 30$, Heterogeneous Treatment Effects					
LOOP (RF)	-0.00004	0.043	0.035	0.00183	99.13%
LOOP (OLS)	0.00000	0.100	0.041	0.01000	99.59%
Cross Estimation	0.00050	0.095	0.034	0.00903	99.70%
OLS	-0.01240	0.095	0.035	0.00919	99.94%
(b) $N = 100$, Heterogeneous Treatment Effects					
LOOP (RF)	0.00002	0.021	0.015	0.00021	99.42%
LOOP (OLS)	0.00001	0.054	0.016	0.00026	100.00%
Cross Estimation	0.00002	0.053	0.015	0.00021	100.00%
OLS	-0.00353	0.053	0.016	0.00027	100.00%
(c) $N = 30$, Homogeneous Treatment Effects					
LOOP (RF)	-0.00003	0.045	0.040	0.0016	98.53%
LOOP (OLS)	-0.00040	0.090	0.083	0.0070	96.22%
Cross Estimation	0.00007	0.045	0.039	0.0015	98.43%
OLS	-0.00152	0.086	0.083	0.0069	95.72%
(d) $N = 100$, Homogeneous Treatment Effects					
LOOP (RF)	-0.00001	0.021	0.014	0.00021	99.43%
LOOP (OLS)	0.00036	0.051	0.048	0.00023	96.07%
Cross Estimation	-0.00001	0.021	0.014	0.00021	99.43%
OLS	0.00026	0.051	0.048	0.00023	95.95%

the method of Section 2.6.3, while the nominal standard errors for cross estimation are calculated using the estimator provided by Wager et al. (2016). For OLS, the point estimate is obtained by regressing Y on T and Z (without any interaction terms), while the nominal standard errors are calculated using the usual formulas (not robust standard errors). Z is treated as a continuous variable in the regression (not as a factor). We also compute the coverage probabilities at a confidence level of 95%. We show the results in Table 2.2. Finally, note that in Table 2.2, the nominal standard error refers to the mean nominal standard error over the 100,000 trials, while the true standard error refers the estimate for the true standard error described above. We continue this practice throughout the remainder of the chapter.

We can see that LOOP and cross estimation are unbiased, while the OLS estimate is biased. This bias is smaller for homogeneous treatment effects and when N is larger. LOOP

(with OLS) performs very similarly to the standard OLS estimator. Although the OLS estimator is slightly biased, it tends to have slightly lower standard errors in these examples, so the mean squared error for the two methods are very similar. However, it is not always the case that LOOP simply shifts the error from bias to standard error. In the case of heterogeneous treatment effects with $N = 30$, we see that the standard OLS estimator has a smaller MSE than LOOP with OLS. We can also see that the true standard errors are similar for LOOP (with random forests) and cross estimation. However, in the case of heterogeneous treatment effects, the nominal standard error of cross estimation is quite conservative, even when N increases. The nominal standard error for LOOP (with random forests) is also conservative, but less so. In the case of cross estimation, this conservative bias is partially because Wager et al. assume that the experimental units are drawn from a superpopulation and must account for this additional uncertainty. For LOOP, the conservative bias is related to the inequality (2.13). For a discussion on a related inequality for the simple difference estimator, see Aronow et al. (2014).

Technical note: cross estimation is slightly biased as implemented. This is due to the difference between the out-of-bag and the leave-one-out estimates of the potential outcomes. This issue can easily be fixed by reducing (by one) the size of the bootstrap sample used in the random forest when making out-of-bag predictions of the potential outcomes.

2.8.2 Simulation 2: Estimating the Treatment Effect for a Binary Response

In our second simulation, we consider a randomized experiment in which the response is either zero or one. Each of the N subjects has one of three sets of potential outcomes: (a) zero regardless of treatment assignment, (b) zero if control and one if treatment, and (c) one regardless of treatment assignment. Like in Section 2.8.1, the treatment assignments are independent Bernoulli random variables with probability 1/2. We also have one covariate (Z_1) that is predictive of the outcome. Higher values of this covariate indicate that the participant is more likely to be in groups (b) or (c) than group (a). Finally, we assume there

are k noise covariates (Z_k).

We generate Z_1 from a standard normal distribution. For each subject i, the probabilities that the subject ends up in each group is determined as follows: we calculate $w_{i1} = 1$, $w_{i2} = \exp(0.5c \times Z_{i1})$, and $w_{i3} = \exp(c \times Z_{i1})$, where c is a positive constant. The probability that observation i is assigned to group j is $p_{ij} = w_{ij}/(w_{i1} + w_{i2} + w_{i3})$. Thus, higher values of c indicate Z_1 is more predictive of outcome. In addition, observation i is most likely to be in the third group (and least likely to be in the first group) if Z_{i1} is positive.

Under this framework, we consider three sets of simulations. First, we assume that both the number of subjects ($N = 200$) and the predictive power of Z_1 ($c = 3$) are constant, and vary the number of noise covariates (from $k = 5$ to $k = 100$ in increments of 5). Next, we fix the predictive power of Z_1 ($c = 3$) and the number of noise covariates ($k = 50$), and vary the number of subjects from 100 to 1000 in increments of 50. Finally, we fix the number of subjects ($N = 200$) and noise covariates ($k = 50$), and vary the predictive power of Z_1 (from $c = 0$ to $c = 5.5$ in increments of 0.5). We run 10,000 trials for each simulation. For each set of simulations, we index the results to the true standard error for the simple difference estimator. We show the results in Figure 2.1.

We observe that while the performance of OLS declines as the number of noise covariates increases, the performance of LOOP remains constant relative to the simple difference estimator. Similarly, OLS performs worse than the simple difference estimator when the number of subjects is small, while the LOOP estimator outperforms the simple difference estimator for all sample sizes. Finally, it is important to note that covariate adjustment does not help when the covariates are not useful for predicting the outcomes. When Z_1 is predictive of the outcome, LOOP outperforms the simple difference estimator. However, we note that even when Z_1 is not predictive of outcome, the performance of the LOOP estimator is still comparable to that of the simple difference estimator. We discuss this further in Section 2.8.4, where we apply the LOOP estimator to actual experimental data.

Technical note: we slightly modify the procedure described in Section 2.8.1. This is

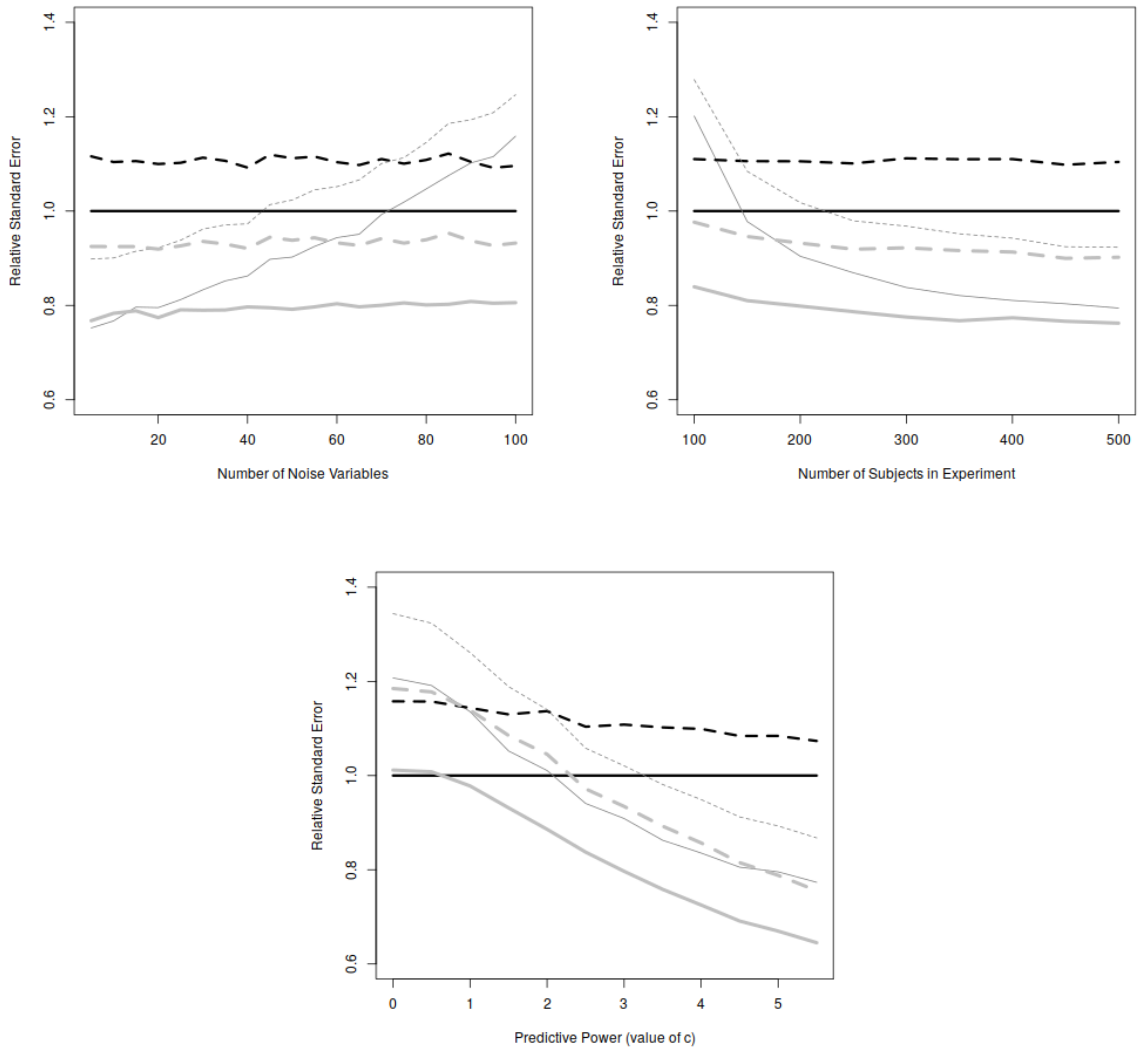

Figure 2.1: Comparison of standard errors for Simulation 2. All standard errors are relative. That is, each value has been divided by the standard error for the simple difference estimator. We use solid lines to denote the the true standard error and dotted lines to denote the nominal standard error. Method used is shown by the color and width of the lines: (a) simple difference estimator, black lines; (b) OLS, thin gray lines; and (c) LOOP, bold light gray lines.

because we compare different simulations in each chart with varying parameter values, and we wish to avoid the variability associated with using a single set of potential outcomes for each simulation. For each of the 10,000 trials, we generate new covariates and potential outcomes and obtain a point estimate and a nominal standard error. We then calculate the nominal standard error as the average of the 10,000 nominal standard errors, and the true

standard error by taking the standard deviation of the 10,000 differences between each point estimate and the true $\bar{\tau}$ for that trial (*i.e.*, the standard deviation of the 10,000 values for $\hat{\tau} - \bar{\tau}$).

2.8.3 Simulation 3: Negligibility of $\bar{\gamma}$

In Section 2.6.3, we argue that $\bar{\gamma}$ is typically negligible and can often be ignored when estimating the variance of $\hat{\tau}$. To support this argument, we show via simulation that $N\bar{\gamma}$ goes to zero as N increases. For this simulation, we generate a single set of $N = 100$ subjects using the setup of Simulation 1 (with heterogeneous treatment effects) in Section 2.8.1 for the covariates and potential outcomes. We then estimate

$$\bar{\gamma} = \frac{\sum_{i \neq j} \gamma_{ij}}{N(N-1)}$$

for each of the first $N = 10, 20, ..., 100$ observations. For each N, we generate 100,000 treatment assignment vectors, calculate the $\hat{\tau}_i$'s for each treatment assignment vector, and use the results to obtain a simulation estimate $\tilde{\gamma}$ of $\bar{\gamma}$, along with a standard error for this estimate. In Figure 2.2, we plot $|N\tilde{\gamma}|$ on a log scale. We can see that the value of $|N\tilde{\gamma}|$ declines as N increases. For a table of the values and standard errors of $\tilde{\gamma}$, see Appendix F.1.

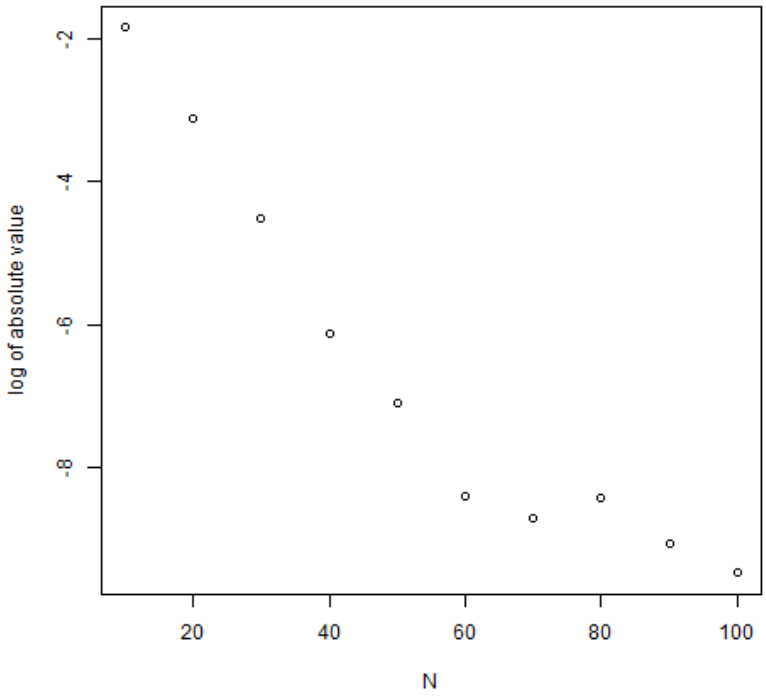

Figure 2.2: Estimate of $|N\bar{\gamma}|$ for different values of N; values are plotted on a log scale. Note that the estimates begin to taper at around $N = 70$. This is due to the standard error of our estimate $\tilde{\gamma}$ of $\bar{\gamma}$. See Appendix F.1 for more details.

2.8.4 Cash Transfer Programs and Enrollment

In their experiment in 2005, Barrera-Osorio et al. studied the effects of several conditional cash transfer programs on educational outcomes for students in Bogota, Colombia. They conducted experiments in two localities of Bogota, San Cristobal and Suba. For our analysis, we focus on the San Cristobal experiment. The San Cristobal experiment involved 10,907 students from grades 6 to 11. These students were selected by lottery to be assigned to one of two treatments or to control: 3,427 students were assigned to the "basic" treatment, 3,424 to the "savings" treatment, and the remaining 4,056 were assigned to control. In the basic treatment, each student received a bi-monthly payment of roughly 15 USD so long as the student attended school at least 80% of days that month. In the savings treatment, each student received a bi-monthly payment of roughly 10 USD so long as they met the

attendance threshold. The remaining third was held in a bank account and paid to the students' families when it was time to re-enroll for the subsequent year. Barrera-Osorio et al. use the following covariates. For each student, they use age, age squared, gender, grade, years behind (or ahead) relative to their grade, and indicator for whether the student is over age for their grade. They also record the marital status, age, and years of education for the head of household, as well as several household characteristics: whether or not the residence is rented or owned, income, total number of people, number of children, and an indicator for single parent household. Finally, they include household values for indices that relate to access to utilities, possession of durable goods, the physical infrastructure of the house, and poverty.

In their experiment, Barrera-Osorio et al. collected re-enrollment status from administrative records. However, they were unable to obtain re-enrollment status for approximately 10% of the observations. In our analysis, we consider both re-enrollment status itself and whether the re-enrollment status is missing as outcome variables. For each outcome variable, we estimate the average treatment effect for the basic treatment compared to the savings treatment, the basic treatment compared to control, and the savings treatment compared to control. We use the same covariates and restrict our analysis to students in grades 6 through 10 as in Barrera-Osorio et al. (2011). We compare the standard errors using LOOP (with random forests), the simple difference estimator, OLS, and cross estimation (with random forests) in Table 2.3. See Appendix F.2 for the full results, including additional methods (LOOP with OLS and OLS with interaction terms) and the point estimates for the treatment effect.

As we can see, OLS, cross estimation, and LOOP provide improvement over the simple difference estimator when missing status is the outcome variable of interest. We can also see that even in this traditional setting (*i.e.*, a large sample size with relatively few covariates), LOOP performs at least as well as OLS. Finally, covariate adjustment does not help when re-enrollment status is the outcome variable, as the covariates are less predictive of outcome.

Table 2.3: Comparison of Standard Errors with Missing and Re-enrollment Status as Outcomes

Treatments	**Method**	**Missing Status** ($\times 10^{-3}$)	**Re-enrollment Status** ($\times 10^{-3}$)
Basic vs. Savings	LOOP	6.0	11.8
	Simple Difference	7.4	11.8
	OLS	6.3	11.6
	Cross Estimation	6.0	11.6
Basic vs. Control	LOOP	5.8	11.6
	Simple Difference	7.1	11.6
	OLS	6.1	11.5
	Cross Estimation	5.7	11.5
Saving vs. Control	LOOP	5.7	11.3
	Simple Difference	7.0	11.4
	OLS	6.1	11.2
	Cross Estimation	5.7	11.2

2.9 Discussion

While methods of covariate adjustment can improve the precision of the estimate of the average treatment effect, they often require the researchers to perform variable selection. For example, when using post-stratification, we must be careful not to use too many covariates otherwise we partition the data set too finely. Over-adjustment can result in poorer performance with linear regression as well: OLS performs poorly when the sample size is small relative to the number of covariates or as the number of noise covariates increases.

The LOOP estimator is an unbiased estimate of the average treatment effect and randomization justifies the assumptions made. One advantage of the LOOP estimator is that estimation of m_i is very flexible. One can impute the potential outcomes using any method, so long as $\hat{m}_i$ and T_i are independent. One baseline approach is to estimate m_i without making use of covariates, simply taking the mean of the observed outcomes in each treatment group. In this case, the LOOP estimator is exactly equal to the simple difference estimator. This suggests that the LOOP estimator will generally outperform the simple difference estimator, so long as we use a sensible method for imputing the potential outcomes. It is

possible to harm precision in certain cases: we might have a small number of observations and an overly flexible imputation method, which could result in overfitting. However, if we were to use a sufficiently regularized imputation method, we would generally expect that the LOOP estimator would perform at least as well as, or at least not much worse than, the simple difference estimator. For example, we might use an ensemble method that includes mean imputation within the ensemble. While we have not explored such an imputation method in this chapter, we expect that it would likely help guard against overfitting.

In this chapter, we suggest the use of random forests to impute the potential outcomes, as they are computationally efficient relative to other methods, likely improve performance over a post-stratified estimate, and allow for automatic variable selection. Because of the automatic variable selection, we can adjust for covariates without knowing ahead of time which covariates we wish to use, and any post-selection inference is still valid. Finally, as with any covariate adjustment method, the LOOP estimator only improves precision over the unadjusted estimator if the covariates are predictive of outcome. However, we see that even when the covariates are not predictive of outcome, the LOOP estimator generally performs as well as the simple difference estimator.

CHAPTER III

Design-Based Covariate Adjustments in Paired Experiments

3.1 Introduction

In randomized controlled trials, we expect the pretreatment covariates of the treatment and control groups to be similar except for the treatment itself. However, there will often be small imbalances in baseline covariates due to chance variation in treatment assignment, which can be addressed in multiple ways. One way to improve the precision of the treatment effect estimate would be to adjust for these imbalances during the analysis. Alternatively, it might be possible to balance covariates through the design of the experiment. For example, in paired experiments, participants are organized into pairs prior to treatment assignment, and then one participant in each pair is randomly assigned to treatment. Ideally, the two participants in each pair would be as similar as possible.

Paired designs are commonly used when the sample size is small. For example, Pane et al. (2014) discuss a randomized trial involving schools in Texas testing the effectiveness of a computer program, the Cognitive Tutor Algebra 1 curriculum. In this trial, schools were organized into 22 pairs and then pair randomized.

While a paired design is often effective at balancing covariates between the treatment and control groups, it may still be helpful to make adjustments for remaining covariate

imbalances. Similar situations can occur with other study designs; for example, covariate adjustments may be helpful in rerandomized trials (see Li and Ding (2020)). Perhaps in part because covariate balance is addressed through experimental design, covariate adjustment methods in paired experiments are relatively understudied. Covariate adjustment methods can be model-based or design-based (for a discussion, see Imai et al. (2009) and Imbens (2010)). Model-based estimators have the potential to improve efficiency; however, incorrect modeling assumptions can result in bias and increased mean squared error. Design-based estimators rely only on randomization as the basis for inference, diminishing the concern of model misspecification. Hierarchical linear models (see Raudenbush and Bryk (2002) and Woltman et al. (2012)) are an example of a model-based approach for blocked experiments, including paired experiments. Pinheiro and Bates (2000) and Dixon (2016) note that hierarchical linear models are a common way to analyze blocked experiments. However, the use of such models requires one to make various modeling decisions, potentially raising concerns about model misspecification. For example, Dixon (2016) notes that there is some debate as to whether block effects should be modeled as fixed or random.

As noted above, covariate adjustments in paired experiments are relatively understudied, and design-based methods are even more so. Imbens and Rubin (2015) and Fogarty (2018) discuss regression-based adjustments. Imbens and Rubin work under a superpopulation model, assuming that the pairs within the experiment are drawn at random from an infinite population, and focus on the population average treatment effect. Fogarty examines the use of regression adjustments in paired experiments under a design-based framework, building on the work of Freedman (2008) and Lin (2013), who discuss regression adjustments in completely randomized experiments. More recently, covariate adjustment methods have been proposed for completely randomized and Bernoulli randomized experiments that involve the use of sample splitting and machine learning methods to impute potential outcomes. These include Aronow and Middleton (2013), Wager et al. (2016), Chernozhukov et al. (2018), Spiess (2018), and Rothe (2018), as well as the work of Chapter II. Some of these methods can

be used in more general designs including blocked experiments, *e.g.*, Aronow and Middleton (2013). However, unlike the case of regression adjustments, there is not currently an analogue to these methods that is specifically for paired experiments.

In this chapter, we present an analogous approach to these machine learning methods for paired experiments. The method is design-based; however, it also allows for the use of models to improve performance. We leave out each pair and impute the potential outcomes using information from the remaining observations. This imputation can be done with any prediction method, such as linear regression or random forests. Regardless of the imputation method, the resulting treatment effect estimate is unbiased and randomization is the basis for inference. This flexibility has several advantages. For example, one issue when making covariate adjustments is choosing which and how many covariates to use. We can address this issue by choosing an imputation method that allows for automatic variable selection. An alternative approach is to use targeted maximum likelihood estimation, which Moore and van der Laan (2009) note allows for automatic variable selection when making covariate adjustments. Balzer et al. (2016b) and Balzer et al. (2016a) propose the use of targeted maximum likelihood estimation in paired experiments.

Our method also addresses an issue that is specific to paired experiments, which we will call the pair inclusion trade-off. In paired experiments, the performance of a covariate adjustment method can suffer if it fails to properly account for the pair assignments. If the relationship between the covariates and outcome within pairs is the opposite of the relationship overall, *i.e.*, a Simpson's paradox occurs, then omitting the pair assignments will hurt precision relative to the unadjusted estimator. However, in cases where the pair assignments are not predictive of the outcome, it is better to ignore the pairing. Both Aronow and Middleton (2013) and Wu and Gagnon-Bartsch (2018) present versions of their methods that allow for block randomizations; however, neither of these methods directly address the pair inclusion trade-off. We discuss the pair inclusion trade-off further in Section 3.4. The framework we present allows us to address the trade-off. We impute two sets of

potential outcomes, one in which we account for and the other where we ignore the pair assignments. Having two sets of imputed potential outcomes, we then interpolate between them by minimizing the cross validated mean squared error. By addressing this trade-off, we protect against the Simpson's paradox, but retain the potential for improvements in precision if the pairing is not informative.

Covariate adjustment methods have also been proposed for matched-pair cluster randomized trials. For example, Small et al. (2008) propose a design-based estimator, while Wu et al. (2014) propose a method that assumes a superpopulation.

This chapter is organized as follows. In Section 3.2, we discuss the model and introduce notation. In Section 3.3, we present the estimator and derive a variance estimate. We discuss the pair inclusion trade-off further and present an imputation method to address it in Section 3.4. In Section 3.5, we apply the estimator to simulated data. In Section 3.6, we use the method to estimate the effect of the Cognitive Tutor Algebra 1 curriculum mentioned above. Section 3.7 concludes.

3.2 Background and Notation

3.2.1 Estimating the Average Treatment Effect

In this chapter, we once again work under the Neyman-Rubin model. Our notation largely follows that of Chapter II. In particular, consider a randomized experiment in which there are $2N$ individuals, indexed by $i = 1, 2, ..., 2N$. We let $T_i = 1$ if the participant is assigned to treatment and $T_i = 0$ if control. Each of the $2N$ participants has two fixed (non-random) potential outcomes, t_i and c_i. We observe t_i if participant i is assigned to treatment and c_i otherwise. That is, the observed outcome Y_i for participant i is

$$Y_i = T_i t_i + (1 - T_i) c_i.$$

We define the individual treatment effect for each participant as $t_i - c_i$, and the average treatment effect as

$$\bar{\tau} = \frac{1}{2N} \sum_{i=1}^{2N} (t_i - c_i).$$

We first consider a case where the treatment assignments are not pair randomized. Suppose the T_i are independent Bernoulli random variables with probability $p = 0.5$, and that we wish to estimate the average treatment effect. One estimate is obtained by taking the average observed outcome of the treatment group and subtracting the average observed outcome of the control group (the "simple difference estimator"). This estimator is unbiased, conditional on both the treatment and control groups containing at least one participant. However, for each participant, suppose we observe a q-dimensional vector of baseline covariates Z_i prior to treatment assignment. It may be possible to use these covariates to improve the precision of the estimate over the simple difference estimator. For example, we could estimate the average treatment effect as

$$\frac{1}{2N} \sum_{i=1}^{2N} \{2(Y_i - \hat{m}_i)T_i - 2(Y_i - \hat{m}_i)(1 - T_i)\}, \tag{3.1}$$

where $\hat{m}_i$ is a function of Z_i. Several authors have noted an estimator of this form can be used to incorporate covariate information (for example, see Robins et al. (1994), Scharfstein et al. (1999), Robins (2000), and Tsiatis et al. (2008)). Aronow and Middleton (2013) use this estimator in a design-based framework, and note that if $\hat{m}_i$ is predictive of the observed outcome Y_i, then the resulting estimate will improve over the unadjusted estimator. In Chapter II, we build on this work and suggest estimating the quantity $m_i = (t_i + c_i)/2$. In addition, Aronow and Middleton (2013) note that this estimate is unbiased if T_i and $\hat{m}_i$ are independent. One way to ensure this independence is by obtaining $\hat{m}_i$ through a sample splitting procedure. For example, one could leave out the i-th observation and calculate $\hat{m}_i$ using the remaining observations. As noted by Aronow and Middleton (2013), sample splitting is especially natural in the case of block randomized experiments, where treatment

assignments in one block are independent of treatment assignments in the remaining blocks. See Wager et al. (2016), Chernozhukov et al. (2018), Spiess (2018), and Rothe (2018) for similar estimators.

3.2.2 Notation for Paired Experiments

We now consider the case where the participants are pair randomized. Suppose that the $2N$ participants are organized into N pairs. We index the pairs by $i = 1, 2, ..., N$, each with two participants indexed by $j = 1, 2$, and the quantities defined in Section 3.2.1 are re-indexed by i and j. For example, for participant j in pair i, we denote the potential outcomes as t_{ij} and c_{ij}, and define the observed outcome, treatment indicator, and covariates as Y_{ij}, T_{ij}, and Z_{ij}, respectively.

For each pair, one of the two participants is randomly chosen to be assigned to treatment and the other is assigned to control. That is, $T_{i1} \sim \text{Bern}(0.5)$, and $T_{i2} = 1 - T_{i1}$. The T_{ij}'s are not mutually independent because exactly one participant in each pair must be assigned to treatment. However, we assume the T_{i1}'s are mutually independent. We can therefore essentially convert our paired experiment to a Bernoulli randomized experiment by treating each pair as an experimental unit, as we describe next.

When treating each pair as a unit, we can draw direct analogues between the notation of paired and Bernoulli randomized experiments. We denote each pair's treatment assignment by T_i, where $T_i = T_{i1}$. For each pair, we also observe a response variable W_i defined below and a $2q$-dimensional vector of baseline covariates (Z_{i1}, Z_{i2}). As with a Bernoulli randomized experiment, each pair has two "potential outcomes": we observe $a_i = t_{i1} - c_{i2}$ if $T_i = 1$ and $b_i = t_{i2} - c_{i1}$ if $T_i = 0$. To differentiate these outcomes from those of the individual participants, we will refer to a_i and b_i as "potential differences." We define the observed difference W_i as:

$$W_i = T_i a_i + (1 - T_i) b_i.$$

We define the pair-level treatment effect τ_i as

$$\tau_i = \frac{(t_{i1} - c_{i1}) + (t_{i2} - c_{i2})}{2} = \frac{1}{2}(a_i + b_i)$$

and the average treatment effect $\bar{\tau}$ as

$$\bar{\tau} = \frac{1}{N} \sum_{i=1}^{N} \tau_i$$

which is our primary parameter of interest.

We can obtain an unbiased estimate for the average treatment effect in paired experiments by averaging the observed differences

$$\hat{\tau}_{sd} = \frac{1}{N} \sum_{i=1}^{N} W_i.$$

We will refer to this estimator as the simple difference estimator for paired experiments, as it is exactly equal to the difference in means between the treatment and control groups. However, the variance estimation of the simple difference estimator will be different under a paired design than it is in completely or Bernoulli randomized experiments. For more details, see Imai (2008), who analyzes $\hat{\tau}_{sd}$ under the Neyman-Rubin model in a paired design.

As in the case of completely or Bernoulli randomized experiments, it may be possible to use covariates to improve precision over the simple difference estimator. We propose such a covariate adjustment method for paired experiments in the next section.

3.3 A Design-Based Covariate Adjustment Procedure

3.3.1 Estimating the Average Treatment Effect

We now present an estimator that is analogous to the estimator given in (3.1), but for paired experiments. Define the quantity

$$\begin{aligned} d_i &= m_{i1} - m_{i2} \\ &= \frac{1}{2}(a_i - b_i) \end{aligned}$$

where $m_{ij} = (t_{ij} + c_{ij})/2$, and let

$$\hat{\tau}_i = \left(W_i - \hat{d}_i\right) T_i + \left(W_i + \hat{d}_i\right) (1 - T_i)$$

where $\hat{d}_i$ is an estimate for d_i. This estimator differs from (3.1) as d_i involves a difference of potential differences, while m_i in (3.1) involves a sum of potential outcomes.

Recall that for Bernoulli randomized experiments, (3.1) is an unbiased estimate of the average treatment effect if $\hat{m}_i$ and T_i are independent. An identical argument can be used for paired experiments to show that $\hat{\tau}_i$ will be unbiased if $\hat{d}_i$ and T_i are independent.

We define an estimate of the average treatment effect as

$$\begin{aligned} \hat{\tau} &= \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_i \\ &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \left(W_i - \hat{d}_i\right) T_i + \left(W_i + \hat{d}_i\right) (1 - T_i) \right\} \end{aligned} \tag{3.2}$$

in which we estimate d_i by using a leave-one-out procedure. We will refer to this sample splitting estimator as the P-LOOP (paired leave-one-out potential outcomes) estimator. For each pair i, we drop both observations and use the remaining $N-1$ pairs to impute a_i and b_i using any method (such as a random forest or linear regression). We then set $\hat{d}_i = \frac{1}{2}(\hat{a}_i - \hat{b}_i)$

and repeat this procedure for all N pairs to obtain $\hat{\tau}$. This leave-one-out procedure ensures that the estimate will be unbiased, as $\hat{d}_i$ and T_i are independent.

To see why this estimator is generally an improvement over the simple difference estimator, we consider the baseline approach where we set $\hat{d}_i = 0$ for all i. In this case, (3.2) will exactly equal the simple difference estimator. As we will show in Section 3.3.3, the variance of (3.2) depends directly on how well one estimates d_i. So long as one estimates the values d_i better than setting $\hat{d}_i = 0$, the estimator will perform better than the simple difference estimator.

3.3.2 Asymptotic Normality

In this section, we demonstrate that (3.2) is asymptotically normally distributed under certain regularity conditions. Consider an infinite sequence of pairs $i = 1, 2, 3, \ldots$. As before, the potential outcomes and covariates for all pairs are fixed quantities. For a given sample size N, we observe the first N pairs in the sequence, and we will consider the behavior of (3.2) as N increases.

We first define some additional notation. Let $U_i = 2T_i - 1$ (*i.e.*, $U_i = 1$ if $T_i = 1$ and -1 if $T_i = 0$) and note that U_i has expectation 0. For a given sample size N, let $\hat{d}_i^{(N)}$ be the estimate for d_i as calculated using the remaining $N - 1$ observations in the sample and define the quantities $d_{0i}^{(N)} = \mathrm{E}(\hat{d}_i^{(N)})$ and $\tilde{d}_i^{(N)} = \hat{d}_i^{(N)} - d_{0i}^{(N)}$. For simplicity, we will often suppress the superscript (N) within an equation.

For some intuition as to why (3.2) converges to a normal distribution, consider the following decomposition:

$$\begin{aligned}\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}\left(W_i - \hat{d}_i U_i\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(W_i - d_{0i}U_i\right) - \frac{1}{N}\sum_{i=1}^{N}\tilde{d}_i U_i\end{aligned}$$

We generally expect $\tilde{d}_i$ to shrink to zero as the sample size increases, while the U_i are either

-1 or 1. We might therefore expect the average of the "remainder terms" $\tilde{d}_i U_i$ to converge to zero. In fact, we later show that

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{d}_i U_i$$

converges in probability to zero. The $W_i - d_{0i}U_i$ terms are independent random variables and are generally of order 1 (*i.e.*, they will not shrink to zero as N increases). So long as none of the variances of the $W_i - d_{0i}U_i$ dominate the variances of the remaining terms, we would expect the appropriately centered and scaled mean of these terms to converge to a normal distribution. If the first term converges in distribution to a normal distribution and the second terms converges in probability to zero, it would follow that $\hat{\tau}$ is asymptotically normally distributed.

In order for this intuition to hold, the data and the imputation method used must be sufficiently well-behaved. We next present general assumptions regarding their behavior. Note that we do not necessarily prove that these assumptions hold for any specific prediction algorithm. However, we do show that the conditions hold for the case of simple linear regression imputation in Bernoulli randomized experiments in Section 4.4. In some cases, verifying the conditions may be difficult. For example, the mathematical properties of random forests are difficult to study, and analyzing the method often requires a simplification of the algorithm (for a further discussion, see Scornet et al. (2015)).

Assumption 1. *There exists some* $0 < C < \infty$ *and* $q > 0$ *such that for all* i*,*

$$Var(\tilde{d}_i) = Var(\hat{d}_i) \leq C/N^q$$

That is, as we observe more units, the variation of $\hat{d}_i$ across randomizations will shrink to zero. For example, suppose we were imputing the potential outcomes using OLS. Under the regularity conditions of Freedman (2008) combined with an assumption of bounded covariates, we would have that $\text{Var}(\hat{d}_i)$ goes to zero at a rate C/N for all i. Note that it should be possible to relax this condition and allow C to vary across pairs (see Wu

and Gagnon-Bartsch (2018) for an analogous condition), but we assume a fixed C here for simplicity.

Assumption 2. *Let ρ_{ij} be the correlation of $\tilde{d}_i U_i$ and $\tilde{d}_j U_j$, and $\bar{\rho} = \frac{\sum_{i \neq j} \rho_{ij}}{N(N-1)}$. We assume that*

$$N^{1-q}\bar{\rho} \longrightarrow 0$$

In the case where $q = 1$, the average correlation would only need to go to zero at any rate for assumption 2 to hold. We would expect the correlation between $\tilde{d}_i U_i$ and $\tilde{d}_j U_j$ to be weak, as the only dependence between these terms comes from the inclusion of U_i in $\tilde{d}_j$ (and U_j in $\tilde{d}_i$). That is, if U_i and $\tilde{d}_j$ were independent, then we would have

$$\begin{aligned} \mathrm{Cov}(\tilde{d}_i U_i, \tilde{d}_j U_j) &= \mathrm{E}(\tilde{d}_i U_i \tilde{d}_j U_j) \\ &= \mathrm{E}(\tilde{d}_i \tilde{d}_j U_j)\mathrm{E}(U_i) = 0. \end{aligned}$$

Moreover, the influence of pair j on $\tilde{d}_i$ (and hence the correlation) should decrease as the number of pairs increases. Even if $\tilde{d}_i$ and $\tilde{d}_j$ are themselves highly correlated, we would expect the correlation between $\tilde{d}_i U_i$ and $\tilde{d}_j U_j$ to be weak. As an extreme example, suppose $\tilde{d}_i = \tilde{d}_j$ exactly. Then

$$\begin{aligned} \mathrm{Cov}(\tilde{d}_i U_i, \tilde{d}_j U_j) &= \mathrm{E}(\tilde{d}_i U_i \tilde{d}_j U_j) \\ &= \mathrm{E}(\tilde{d}_i^2 U_i U_j) \\ &= \mathrm{E}(\tilde{d}_i^2 U_i)\mathrm{E}(U_j) = 0. \end{aligned}$$

Assumption 3. *Recall that $d_{0i}^{(N)} = E(\hat{d}_i^{(N)})$ for some fixed N. For each pair i, we assume that the limit of $d_{0i}^{(N)}$ exists and denote the limit as $d_{\infty i}$. We also assume*

$$\frac{1}{N}\sum_{i=1}^{N}\left(d_{0i}^{(N)} - d_{\infty i}\right)^2 \longrightarrow 0.$$

In other words, the expected value of the $\hat{d}_i$ converges pointwise to some limit and the mean square of the d_{0i} converge as well. This does not necessarily mean that d_{0i} will converge to the true value of d_i. In most cases, we would expect that the imputation method will not be able to perfectly estimate d_i on average and we characterize this in the next assumption.

Assumption 4. *Let* $V_N = \sum_{i=1}^{N}(d_i - d_{\infty i})^2$. *There exists* $0 < K < \infty$ *such that*

$$\frac{V_N}{N} \longrightarrow K,$$

and

$$\max_{i=1,\ldots,N} \frac{(d_i - d_{\infty i})^2}{V_N} \longrightarrow 0.$$

That is, the mean squared error of the imputation method converges to a value K. In addition, no single term of the mean squared error dominates the remaining terms.

When assumptions 1 through 4 hold $N(\hat{\tau}-\tau)/\sqrt{V_N}$ converges in distribution to a standard normal random variable. For a proof, see Appendix G.

3.3.3 Variance

We now estimate the variance of (3.2). Let $\hat{W}_i = \hat{a}_i T_i + \hat{b}_i(1 - T_i)$, and define the mean squared errors of $\hat{d}_i$ and $\hat{W}_i$ as $\text{MSE}(\hat{d}_i) = \text{E}\{(d_i - \hat{d}_i)^2\}$ and $\text{MSE}(\hat{W}_i) = \text{E}\{(W_i - \hat{W}_i)^2\}$. In Appendix H, we show

$$\text{Var}(\hat{\tau}_i) = \text{MSE}(\hat{d}_i)$$

and thus that the variance is

$$\text{Var}(\hat{\tau}) = \frac{1}{N^2}\left\{\sum_{i=1}^{N} \text{MSE}(\hat{d}_i) + \sum_{i \neq j} \gamma_{ij}\right\} \tag{3.3}$$

where $\gamma_{ij} = \text{Cov}(\hat{\tau}_i, \hat{\tau}_j)$. In Appendix I, we show that

$$\frac{\sum_{i \neq j} \gamma_{ij}}{\sum_{i=1}^{N} \text{MSE}(\hat{d}_i)} \longrightarrow 0$$

under the conditions outlined in Section 3.3.2. Because $\sum_{i \neq j} \gamma_{ij}$ is negligible relative to $\sum_{i=1}^{N} \text{MSE}(\hat{d}_i)$, we suggest that the variance be estimated without the covariance terms in practice. For this reason, we focus on estimating $\text{MSE}(\hat{d}_i)$.

In Appendix J, we show that the mean squared error of $\hat{d}_i$ is less than the mean squared error of $\hat{W}_i$ and thus that

$$\frac{1}{N^2} \sum_{i=1}^{N} \text{MSE}(\hat{d}_i) \leq \frac{1}{N^2} \sum_{i=1}^{N} \text{MSE}(\hat{W}_i)$$

We can obtain an unbiased estimate for this upper bound, which we use to estimate the variance of $\hat{\tau}$:

$$\widehat{\text{Var}}(\hat{\tau}) = \frac{1}{N^2} \sum_{i=1}^{N} (W_i - \hat{W}_i)^2. \tag{3.4}$$

To compare this variance estimator to the variance estimator for the simple difference estimator, consider a special case where we estimate the average treatment effect without using covariates. In absence of any covariate information, it would be logical to set $\hat{a}_i = \hat{b}_i = \bar{W}^{(-i)}$ where $\bar{W}^{(-i)} = \sum_{j \neq i} W_j/(N-1)$. In this baseline approach, the P-LOOP estimator would exactly equal the simple difference estimator, as $\hat{d}_i = 0.5(\bar{W}^{(-i)} - \bar{W}^{(-i)}) = 0$ for all i. In addition, we show in Appendix K that the variance estimate for the P-LOOP estimator would equal

$$\frac{1}{(N-1)^2} \sum_{i=1}^{N} (W_i - \hat{\tau}_{sd})^2,$$

which is equal to $N/(N-1)$ times the standard variance estimate in a paired t-test (for

example, see Imai (2008)).

Importantly, because $\text{Var}(\hat{\tau}_i) = \text{MSE}(\hat{d}_i)$, the performance of the estimator depends directly on how well we estimate d_i. If we improve the estimate of d_i over setting $\hat{a}_i = \hat{b}_i = \bar{W}^{(-i)}$, we will be able to improve precision relative to the simple difference estimator. However, improving the estimate of d_i is not necessarily trivial. Because we are interested in estimating the difference between m_{i1} and m_{i2}, it does not suffice to reduce the mean squared error for the imputed potential outcomes as in the estimator of Wu and Gagnon-Bartsch (2018). For example, it is possible to obtain estimates of the potential outcomes (the t's and c's) that are reasonably close to the true values while having $\hat{d}_i$ of the incorrect sign. On the other hand, we could have estimates for the potential outcomes that are far from the true values that result in $\hat{d}_i$ being close to the true d_i. We discuss imputation methods to address this concern in the next section.

3.4 Imputation Methods of Potential Differences in Paired Experiments

3.4.1 The Pair Inclusion Trade-Off

We next present an imputation method to address the pair inclusion trade-off discussed in Section 3.1. We first discuss this trade-off further and then propose a method for addressing the trade-off within the P-LOOP estimator. The pair inclusion trade-off is perhaps easiest to understand in the context of a linear model, rather than the Neyman-Rubin model. Consider the following standard linear regression model

$$Y = \alpha + T\tau + P\beta + Z\gamma + \epsilon$$

where Y is the observed outcome, T is the treatment assignment vector, Z is a covariate, and P is a $2N \times (N-1)$ matrix of indicator variables that encodes the pair assignments.

Suppose that there are pair effects (that is, $\beta \neq 0$), and that Z is correlated with both P and T. If we were to omit P and regress Y onto T and Z, then we would bias the estimate of τ. On the other hand, suppose that the pairing is not informative ($\beta = 0$). In this case, including P in the regression would inflate the variance for $\hat{\tau}$, and it would be preferable to omit P from the regression.

Several authors have compared the variance of the simple difference estimator for completely and pair randomized designs under the Neyman-Rubin model (for example, see Imai (2008) and Pashley and Miratrix (2017)). The difference in variance under these designs can be either positive or negative. Similarly, it may be possible to reduce the variance of our estimate when making covariate adjustments by ignoring the pair assignments. However, Imai (2008) cautions against analyzing paired experiments as if they were completely randomized, noting that this can result in biased confidence intervals and hypothesis tests. Fortunately, this is not an issue with the P-LOOP estimator, as we always account for the paired design. We always drop both observations in each pair when estimating d_i, and the decision to ignore or include the paired structure for the remaining observations only affects the adjustment term $\hat{d}_i$.

When we discuss the inclusion or exclusion of the paired structure when imputing potential outcomes, we refer specifically to how we treat the remaining pairs when building a prediction model. If we ignore the paired structure when imputing potential outcomes, this means we fit a model to the remaining observations as individual units. If we include the paired structure when imputing potential outcomes, this means we fit a model to the remaining observations, treating each pair as a unit. Regardless of which approach we choose, the estimator remains design-based. For a given pair i, we always leave out both observations, and we wish to use the remaining observations such that we obtain the best estimate for d_i.

Suppose we ignore the paired structure of the data when we train our imputation model for the potential outcomes. In this case, we model the relationship between the covariates and the outcome overall, rather than the relationship within pairs. However, if the relationship

between the covariates and outcome within pairs is sufficiently different from the relationship overall, we could obtain a $\hat{d}_i$ that is far from the truth. One situation where this could happen is when a Simpson's paradox occurs, and the relationship between the covariates and outcome within pairs is the opposite of the overall relationship.

Consider a hypothetical experiment in which a blood pressure medication is being tested on pairs of twins, and each pair belongs to either ethnicity A or ethnicity B. For each participant, we record a single covariate, an indicator for the presence of a genetic mutation. On average, participants with this mutation have blood pressure that is 5 units lower. Suppose this mutation is common in ethnicity A and rare in ethnicity B. However, for reasons unrelated to the mutation, ethnicity A has a baseline blood pressure that is on average 10 units higher than the baseline for ethnicity B. In this case, the presence of the mutation would be associated with higher blood pressure as ethnicity A is more likely to have the mutation and also has a higher baseline blood pressure. However, within pairs, the presence of the mutation will be associated with lower blood pressure. If we ignore pair assignments when estimating d_i, we would infer that the presence of the mutation is associated with a higher value of blood pressure. For a given pair, we would want the presence of the mutation to predict lower blood pressure. Thus, the prediction of the difference $d_i = m_{i1} - m_{i2}$ would be of the wrong sign, resulting in poorer performance relative to the simple difference estimator. On the other hand, if the paired structure is not predictive of the outcome, then it would be better to omit the pair assignments when imputing the potential differences.

It can be unclear whether we should account for the pair assignments when imputing the potential differences. To avoid data snooping, we propose an imputation method in the rest of this section that automatically addresses the trade-off. We first propose methods for calculating $\hat{a}_i$ and $\hat{b}_i$ that do and do not account for the pair assignments in the prediction model, producing two sets of potential differences. Having produced two estimates for each a_i and b_i, we propose a method to automatically interpolate between them.

3.4.2 Estimating d_i when Pairs are not Predictive: Impute Potential Outcomes Separately

We first estimate d_i without accounting for the pair assignments for the observations outside of pair i. To do this, we drop both observations in pair i, then fit a model on the individual observations for the remaining pairs and separately impute all four potential outcomes (*i.e.*, t_{i1}, c_{i1}, t_{i2}, and c_{i2}) for pair i. Although we ignore pair assignments for the observations outside of pair i, we must drop both observations in the pair when estimating d_i to ensure that the treatment effect estimate is unbiased.

More specifically, for each pair i, we drop both observations in the pair. We then fit a prediction algorithm on the remaining observations, ignoring the pair assignments and treating each individual as a unit. For example, we could regress Y_{kj} onto T_{kj} and Z_{kj} for $k \neq i$. We then use this model to impute t_{i1}, c_{i1}, t_{i2}, and c_{i2}. To obtain $\hat{t}_{i1}$, we would plug in the covariates for the first observation in pair i and a treatment indicator of 1. We would obtain estimates for the remaining potential outcomes similarly and set

$$\hat{d}_i = \frac{1}{2}(\hat{t}_{i1} + \hat{c}_{i1}) - \frac{1}{2}(\hat{t}_{i2} + \hat{c}_{i2}).$$

3.4.3 Estimating d_i when Pairs are Predictive: Impute Potential Differences Directly

Next, we propose a method that accounts for pair assignments when estimating d_i. Rather than imputing the potential outcomes (t_{i1}, c_{i1}, t_{i2}, and c_{i2}), we impute a_i and b_i directly, treating each pair as an observational unit. Recall from Section 3.3 that a_i and b_i are analogous to the potential outcomes in an experiment with Bernoulli randomization. We can therefore apply a procedure to the paired units that is similar to the leave-one-out procedure described earlier for estimating m_i in equation (3.1). For Bernoulli experiments, we would only use the control units when imputing c_i and the treatment units when imputing t_i. However, for paired experiments a_i and b_i are determined by which unit is arbitrarily

labeled $j = 1$ and are therefore effectively interchangeable. As an example, for the i-th pair, we have $a_i = t_{i1} - c_{i2}$. However, if we had instead recorded the second unit in the pair first, then the values of a_i and b_i would be switched and a_i would be $t_{i2} - c_{i1}$. We can take advantage of this fact to use all observations (except those in pair i) when imputing each potential difference.

When treating the pairs as units, we have $2q$ covariates rather than q covariates for each unit. We start by leaving out pair i. We then wish to impute a_i and b_i using the $2q$ covariates for the remaining pairs. One way to do this would be to simply concatenate the covariate vectors for the two observations in each pair. In this case, we define Z_i^a as the vector of covariates where the covariates for the treated units come first. That is, $Z_i^a = (Z_{i1}, Z_{i2})$ if $T_i = 1$, and $Z_i^a = (Z_{i2}, Z_{i1})$ if $T_i = 0$. For example, suppose $Z_{i1} = (1, 2)$ and $Z_{i2} = (3, 4)$. Then Z_i^a would be $(3, 4, 1, 2)$ if $T_i = 0$, and $(1, 2, 3, 4)$ if $T_i = 1$. In other words, Z_i is the concatenated vector of covariates as it is ordered in the original data, while Z_i^a is the concatenated vector where the covariates for the treated unit come first.

Alternatively, we may wish to transform the covariates in some way; for example, we could take the means and differences of the covariates. This is similar to the approaches used by Imbens and Rubin (2015) and Fogarty (2018). In this case, define Z_i as

$$\left(\frac{Z_{i1} + Z_{i2}}{2}, Z_{i1} - Z_{i2}\right).$$

That is, Z_i is the vector where the first q entries are the averages of each covariate for the pair, and the second q entries are the differences (observation 1 minus observation 2). In analogy to the concatenation example, we define Z_i^a to be the means and the treatment minus control differences.

We can now estimate d_i using these combined covariates and the observed differences. After leaving out pair i, we impute a_i by creating a model using the observed outcomes W_k (for $k \neq i$) as our response variable and the covariates Z_k^a as our predictors. This model

incorporates all of the remaining $N-1$ pairs and predicts the value of a for a given set of covariates. We plug the covariates Z_i into this model to obtain $\hat{a}_i$. The same model can be used to impute b_i. If we had labeled the second participant in the pair as the first participant, then a_i and b_i would be reversed. We therefore use the same model to impute b_i, but reverse the order of the covariates for pair i. In the concatenation example, we would plug (Z_{i2}, Z_{i1}) into the model instead of (Z_{i1}, Z_{i2}). In the transformation example, we would plug

$$\left(\frac{Z_{i1}+Z_{i2}}{2}, Z_{i2}-Z_{i1}\right)$$

into the model. Having obtained estimates $\hat{a}_i$ and $\hat{b}_i$, we set

$$\hat{d}_i = \frac{1}{2}(\hat{a}_i - \hat{b}_i).$$

3.4.4 Interpolating between Imputation Methods

We have proposed two methods for imputing potential outcomes. However, we often do not know ahead of time which method will perform better. We therefore adaptively interpolate between the two methods.

For each pair i, we have two estimates of a_i obtained using the two imputation methods described above. We refer to these estimates as $\hat{a}_i^{(1)}$ and $\hat{a}_i^{(2)}$. We wish to obtain the value α_i that minimizes the distance between a_i and the interpolation $\hat{a}_i = \alpha_i \hat{a}_i^{(1)} + (1-\alpha_i)\hat{a}_i^{(2)}$. However, we want $\hat{a}_i$ to be independent of T_i. We therefore use a leave-one-out procedure to calculate α_i. For each i, we leave out pair i and set α_i to the value that minimizes the mean squared error for the remaining observations. In other words, we have

$$\alpha_i = \underset{x\in[0,1]}{\operatorname{argmin}} \sum_{k\in\mathcal{A}\setminus i} \left\{a_k - \left(x\hat{a}_k^{(1)} + (1-x)\hat{a}_k^{(2)}\right)\right\}^2.$$

Taking the derivative with respect to x and setting equal to 0, we have

$$\alpha_i = \frac{\sum_{k \in \mathcal{A} \setminus i} (a_k - \hat{a}_k^{(2)})(\hat{a}_k^{(1)} - \hat{a}_k^{(2)})}{\sum_{k \in \mathcal{A} \setminus i} (\hat{a}_k^{(1)} - \hat{a}_k^{(2)})^2}.$$

which we then restrict to be in the interval $[0, 1]$. We then set our final estimate of a_i to be $\hat{a}_i = \alpha_i \hat{a}_i^{(1)} + (1 - \alpha_i)\hat{a}_i^{(2)}$. We use a similar procedure for $\hat{b}_i$.

3.5 Simulation Results

We present two simulations in the next two subsections. The first simulation illustrates the pair inclusion trade-off, while the second considers a scenario with a non-linear relationship between the covariate and potential outcomes. In both cases, we compare the performance of P-LOOP with the simple difference estimator and the estimators discussed in Fogarty (2018), which we will refer to as Regression 1 and Regression 2. Regression 1 involves the treatment minus control outcomes regressed onto the treatment minus control covariates, while Regression 2 is the same regression with the addition of the mean of the covariates in each pair. For P-LOOP, recall from earlier that we are excluding the pair assignments in our imputation method if we impute the potential outcomes (t_{i1}, c_{i1}, t_{i2}, and c_{i2}) separately, while we are including the pair assignments if we impute the potential differences (a_i and b_i) directly. We show results using each of these imputation strategies as well as the interpolation method. We use both random forests and ordinary least squares as prediction methods.

For each of the scenarios described below, we generate a single set of potential outcomes. Next, we generate 10,000 treatment assignment vectors. For each of these, we obtain a treatment effect estimate and the nominal variance (*i.e.*, the estimated variance) using each estimator. This results in 10,000 point estimates and 10,000 variance estimates for each method, which we can use to estimate the true variance and the expectation of nominal variance for that method. We estimate the true variance as the variance of the 10,000 point

estimates, and the expectation of the nominal variance as the mean of the 10,000 nominal variances.

3.5.1 The Pair Inclusion Trade-Off

We consider a hypothetical experiment based off the scenario described in Section 3.4.1, where we are interested in the effect of a blood pressure medication. We generate $N = 50$ pairs of twins, half of which are of ethnicity $E_i = 0$ and the other half $E_i = 1$. We randomly assign one participant in each pair to treatment and assign the other to control. That is, $T_{i1} \sim \text{Bern}(0.5)$ and $T_{i2} = 1 - T_{i1}$. Next, suppose there exists a genetic mutation Z_{ij}. For each participant, we set $Z_{ij} \sim \text{Bernoulli}(p_k)$ for $E_i = k$. We set $p_1 = 0.9$ and $p_0 = 0.5$. That is, participants of ethnicity $E_i = 1$ are more likely to have the mutation. We assume that only the observed outcome Y_{ij}, as well as T_{ij} and Z_{ij}, are recorded. Suppose that ethnicity 1 has a higher baseline blood pressure than ethnicity 0 (for reasons unrelated to the mutation), but that the presence of the mutation is causally associated with lower blood pressure. We generate the outcome as:

$$Y_{ij} = 80 - 10T_{ij} - 5Z_{ij} + 10E_i + \epsilon_{ij}$$

where ϵ_{ij} are independent $\text{N}(0, 4)$ random variables. Because participants for ethnicity $E_i = 1$ have higher baseline blood pressure, Z_{ij} is positively correlated with blood pressure across all participants. Thus a Simpson's paradox occurs: overall, Z_{ij} has a positive association with blood pressure, while within pairs, Z_{ij} has a negative association with blood pressure. We summarize the results of this simulation in Table 3.1 under the column Simpson's paradox.

We also generate a set of potential outcomes in which the pairs contain no additional information (beyond its association with covariate Z_{ij}). We generate the observed outcome

Table 3.1: Simulation Results

	Simpson's paradox			Uninformative pairs		
Method	True var	E(Nom)	Cov pr	True var	E(Nom)	Cov pr
Simple Difference	0.343	0.342	0.943	0.361	0.365	0.947
P-LOOP RF (differences)	0.154	0.167	0.951	0.151	0.168	0.952
P-LOOP RF (outcomes)	0.440	0.462	0.952	0.146	0.154	0.949
P-LOOP RF (interpolated)	0.152	0.170	0.953	0.148	0.156	0.948
P-LOOP OLS (differences)	0.152	0.160	0.950	0.148	0.160	0.950
P-LOOP OLS (outcomes)	0.442	0.462	0.953	0.146	0.154	0.949
P-LOOP OLS (interpolated)	0.152	0.164	0.952	0.148	0.156	0.949
Regression 1	0.151	0.150	0.943	0.148	0.149	0.944
Regression 2	0.153	0.148	0.942	0.149	0.148	0.940

Note. True var is the estimate for the true variance. E(Nom) refers to the estimate for the expected value of the nominal variance. For P-LOOP, these are estimates for expression (3.3) and for the expected value of (3.4) respectively. Cov pr is the estimated coverage proportion. We provide further details on how we obtain these estimates in Appendix L.1. The Monte Carlo estimates of the true variances have standard errors ranging from 0.002 to 0.007, while the Monte Carlo estimates for the expected values of the nominal variances all have standard errors below 0.0002. We provide these standard errors in Appendix L.1.

as:

$$Y_{ij} = 80 - 10T_{ij} + 5Z_{ij} + \epsilon_{ij}$$

where ϵ_{ij} are independent $\mathrm{N}(0, 4)$ random variables. In this case, E_i is associated with outcome because it is associated with Z_{ij}, but otherwise has no effect on outcome. We summarize the results of this simulation in Table 3.1 under the column Uninformative pairs.

We see that in the Simpson's paradox case, imputing the potential outcomes separately (not accounting for pairs when estimating a_i and b_i) causes inflated variance relative to

the simple difference estimator, while imputing potential differences directly (accounting for pairs) results in improved performance. However, in the case where the pair assignments are uninformative, it is better to impute the potential outcomes separately. The gains in this example are relatively minor; however, we show in the later sections that the improvements can be more substantial.

3.5.2 A Non-Linear Scenario

In the previous example, the potential outcomes were generated from a linear model with independent, normally distributed noise. We examine a more complex scenario in this section. Consider a hypothetical experiment in which we are testing the effect of a drug on recovery time for an illness. We generate $N = 50$ pairs. For each participant, we observe a single covariate, Z, corresponding to the baseline health score for that participant. To obtain this health score, we generate $Z_{0i} \sim \text{Unif}(0, 10)$ for each pair i. We then set $Z_{ij} = Z_{0i} + \epsilon_{ij}$, where ϵ_{ij} are independent $\text{N}(0, 1)$ random variables. The outcome in this example will be time to recovery.

The mean recovery time under treatment and control will be determined by the following logistic functions:

$$\mu_c(Z_{ij}) = 3 + \frac{10}{1 + \exp\left(2Z_{ij} + 12\right)}$$

and

$$\mu_t(Z_{ij}) = 3 + \frac{10}{1 + \exp\left(2Z_{ij} + 8\right)}.$$

We then generate the control potential outcomes using gamma random variables with shape parameter 30 and rate parameter $30/\mu_c(Z_{ij})$. We generate the treatment potential outcomes analogously. A higher health score is associated with quicker recovery under both treatment and control; however, this recovery is expected to occur more quickly for treated units. We show the results of this simulation in Table 3.2. P-LOOP with random forests outperforms the other methods. This is not surprising, as the potential outcomes are obtained using a

Table 3.2: Simulation Results

Method	True var	E(Nom var)	Cov pr
Simple Difference	0.094	0.373	1
P-LOOP RF (differences)	0.069	0.160	0.996
P-LOOP RF (outcomes)	0.046	0.097	0.991
P-LOOP RF (interpolated)	0.046	0.097	0.992
P-LOOP OLS (differences)	0.068	0.371	1
P-LOOP OLS (outcomes)	0.062	0.364	1
P-LOOP OLS (interpolated)	0.065	0.363	1
Regression 1	0.066	0.351	1
Regression 2	0.066	0.358	1

Note. True var is the estimate for the true variance. E(Nom var) refers to the estimate for the expected value of the nominal variance. For P-LOOP, these are estimates for expression (3.3) and for the expected value of (3.4) respectively. Cov pr is the estimated coverage proportion. We provide further details on how we obtain these estimates in Appendix L.1. The Monte Carlo estimates of the true variances have standard errors ranging from 0.0006 to 0.0013, while the Monte Carlo estimates for the expected values of the nominal variances all have standard errors below 0.0004. We provide these standard errors in Appendix L.1.

non-linear data generating process. In addition, all of the methods are conservative, although P-LOOP with random forests is much less conservative than the other methods. We also observe that there is considerable benefit from excluding the pair assignments when imputing potential outcomes when using random forests as the imputation method.

3.5.3 Remainder Terms

In this subsection, we investigate the quantity

$$\frac{1}{N}\mathrm{E}\left\{\left(\sum_{i=1}^{N}\tilde{d}_i U_i\right)^2\right\} \tag{3.5}$$

for each of the data generating procedures used in the simulations discussed above. This quantity is of interest for several reasons. The convergence of (3.5) to zero plays an important role in proving the central limit theorem discussed in Section 3.3.2. This convergence also implies that $\sum_{i \neq j} \gamma_{ij}$ is negligible relative to $\sum_{i=1}^{N} \text{MSE}(\hat{d}_i)$, as discussed in Section 3.3.3. In Appendix L.2, we show that

$$\left| \frac{1}{N} \sum_{i \neq j} \gamma_{ij} \right| \leq \frac{1}{N} \text{E} \left\{ \left(\sum_{i=1}^{N} \tilde{d}_i U_i \right)^2 \right\}.$$

It follows that the convergence of (3.5) to zero implies the convergence of $\sum_{i \neq j} \gamma_{ij} / N$ to zero.

For each of the three data generating processes discussed in Sections 3.5.1 and 3.5.2, we generate potential outcomes and covariates for 1000 pairs. We then consider each of the first $N = 50, 100, \ldots, 1000$ of these pairs. For a given N, we generate 1000 treatment assignment vectors, which we use to estimate (3.5) for both random forest and OLS imputation. For more details on the simulation procedure, see Appendix L.2.

In Figure 3.1, we plot the estimated values of (3.5) against the sample size N (both on a log base 10 scale). For each of the data generating procedures (and for both imputation methods), we can see that the estimated values of (3.5) shrink to zero as N increases. For the non-linear data generating process, this decrease occurs more slowly when using random forest imputation. Note that (3.5) contains terms relating to both the variances and covariances of the $\tilde{d}_i U_i$. With this particular data generating process, the variance of $\tilde{d}_i$ shrinks more slowly with random forest imputation. For a further discussion, see Appendix 6.

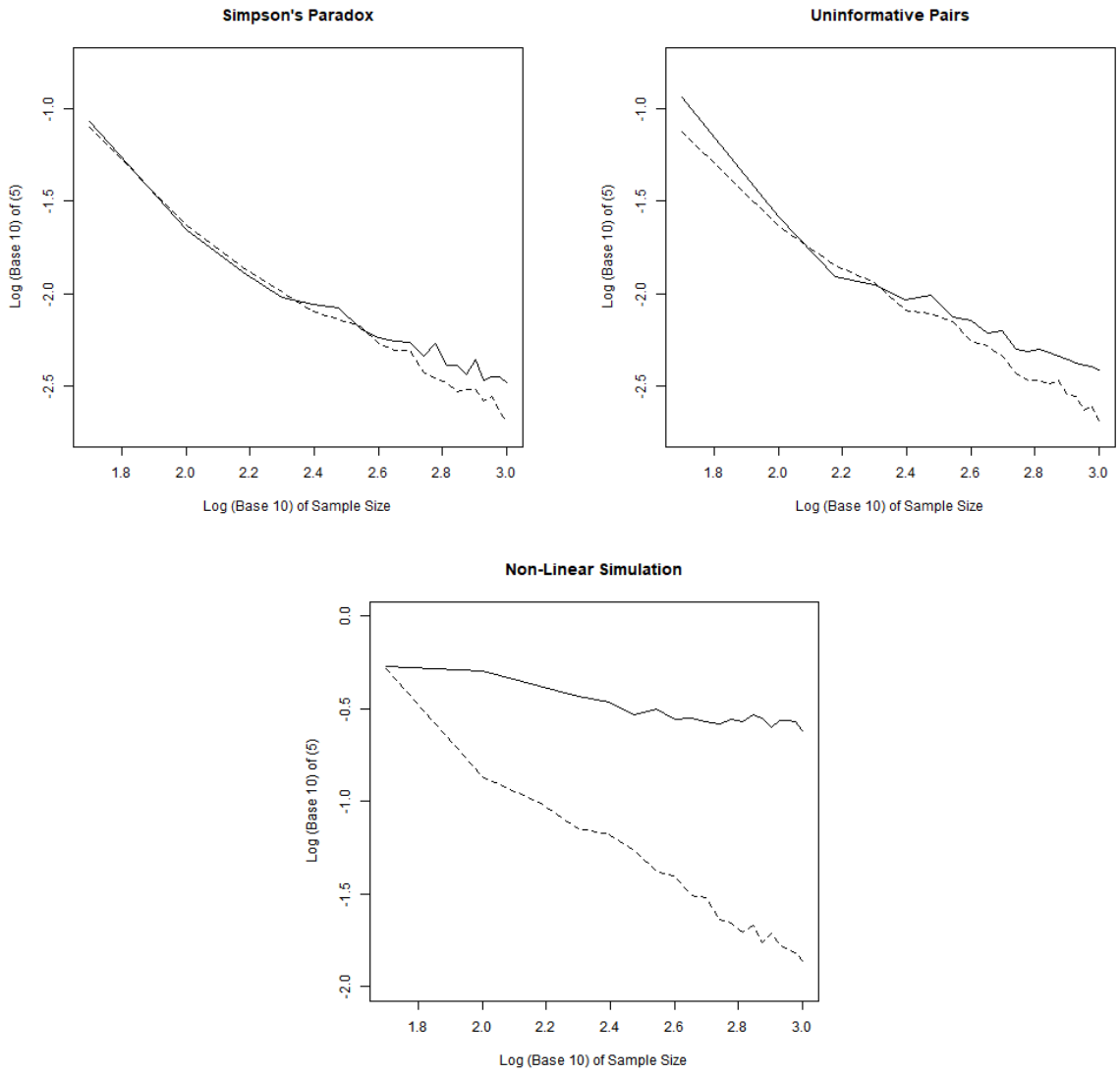

Figure 3.1: We plot the estimated values of quantity (3.5) (*i.e.*, $\mathrm{E}\{(\sum_{i=1}^{N} \tilde{d}_i U_i)^2\}/N$) against the sample size N. Both values are plotted on a log base 10 scale. The top two charts show the estimates of (3.5) corresponding to the data generating procedures in Section 3.5.1. The bottom chart shows the estimates corresponding to Section 3.5.2. The values of (3.5) are estimated for both random forest imputation (solid line) and OLS imputation (dashed line).

3.6 Cognitive Tutor Impact Study

We apply our method to estimate the effect of an intervention in a randomized trial involving schools in Texas. This trial (discussed in Pane et al. (2014)) tested the effectiveness of a computer program, the Cognitive Tutor Algebra 1 curriculum, and included 22 pairs of schools. The outcome of interest is the passing rate of the schools on the math section of the Texas Assessment of Knowledge and Skills (TAKS) in 2008. Available covariates included the school type (middle or high school) and a pretest score, the passing rate from 2007. We estimate the average treatment effect using either just the pretest score or both the pretest score and school type as covariates. In Table 3.3, we compare the performance of P-LOOP with the simple difference estimator and the estimators discussed in Fogarty (2018). We use random forests and linear regression as imputation methods in the P-LOOP estimator. As in the case of the simulations, we show the results imputing potential differences (accounting for pairs), imputing potential outcomes separately (ignoring the pair assignments), and the interpolation between the two. Note that P-LOOP imputing potential differences with OLS most closely matches the Regression 2 method, as both methods account for pairing and use the differences and averages of the covariates for making adjustments.

Both P-LOOP and the methods of Fogarty (2018) have smaller nominal variance than the unadjusted estimator. Regression 1 has lower variance than Regression 2 when the pretest score is the only covariate, but Regression 2 has lower variance when the school type is included. Both regression methods always account for the pair assignments. For the P-LOOP estimator, we see that it is better to impute the potential outcomes separately, and that the interpolation method imputes values closer to the potential outcomes imputation. With the interpolation method, we do not lose out on the precision gains from ignoring the pairs in our imputation, but we are still protected against a potential Simpson's paradox.

Table 3.3: Comparison of Methods

	Pretest		Pretest and school type	
Method	Point est	Nominal var	Point est	Nominal var
Simple Difference	-6.82	9.82	-6.82	9.82
P-LOOP RF (differences)	-4.41	7.06	-5.62	7.86
P-LOOP RF (outcomes)	-2.82	5.72	-4.94	5.60
P-LOOP RF (interpolated)	-3.53	6.39	-5.10	5.75
P-LOOP OLS (differences)	-2.79	6.56	-2.17	4.38
P-LOOP OLS (outcomes)	-2.04	5.66	-1.81	4.13
P-LOOP OLS (interpolated)	-2.08	5.85	-2.06	4.00
Regression 1	-2.61	6.18	-2.61	6.18
Regression 2	-2.60	6.56	-2.27	4.57

Note. Point est and nominal var refer to the point estimates and nominal variances for each method.

3.7 Discussion

In paired experiments, the design of the experiment helps to enforce covariate balance between the treatment and control groups. While this design is often effective, it can be useful to make covariate adjustments to further improve precision. Covariate adjustments in paired experiments share many of the issues in completely randomized experiments; for example, it can be unclear ahead of time which covariates to use. A unique issue to paired experiments is the pair inclusion trade-off, so we must take particular care when making adjustments in paired experiments. Failing to account for the pair assignments can harm performance (for example, when a Simpson's paradox occurs), while including the paired structure when the pair assignments are not predictive can needlessly inflate variance.

We present a design-based method for paired experiments, the P-LOOP estimator. This estimator is guaranteed to be unbiased by design. Nonetheless, the pair-inclusion trade-off is

still relevant because it affects the variance of the estimator. To the best of our knowledge, this method is the first to directly address the pair inclusion trade-off. Generally, other methods account for the pairing, which protects against Simpson's paradox and other situations where the within pair trends differ from the overall trend. However, our method imputes two sets of potential outcomes, one excluding and one including the pair assignments, and automatically interpolates between the two. As we see in the Texas Schools data, this allows for improved precision. The P-LOOP estimator is also the first method specifically for paired experiments that involves sample splitting and the use of machine learning methods to impute potential outcomes, building on the flexible approaches used in completely randomized experiments. This flexibility can be beneficial in several ways, such as allowing for automatic variable selection or high dimensional covariates. However, the leave-one-out approach can also be computationally intensive. If computation time is an issue, one can modify the procedure to leave out multiple pairs instead of single pairs at a time.

CHAPTER IV

Integrating Experimental and Observational Data

4.1 Introduction

A well known advantage of randomized experiments is that they do not suffer from confounding bias. They also allow for design-based inference; that is, the act of randomization largely justifies the statistical assumptions made. Design-based estimators are typically unbiased and their associated inference (*e.g.*, their standard errors) come with guarantees regarding accuracy, while relying only on randomization as the basis for inference. Examples of design-based methods include Schochet (2015) and Rosenbaum (2002), as well as the estimator introduced in Chapter II. However, sample sizes in randomized experiments may be limited by practical considerations and are often small, which can limit precision of treatment effect estimates. Conversely, observational studies typically offer much larger sample sizes at lower costs, but without the statistical guarantees from randomization. For example, the analysis of observational data often requires untestable assumptions, such as the assumption of no unmeasured confounding variables.

Prior literature has explored the possibility of improving precision in randomized experiment by pooling the controls from the experiment with historical controls from observational data sets or from similar experiments. This literature dates back to at least Pocock (1976); see Viele et al. (2014) and Lim et al. (2018) for more recent discussions. While much of this work uses a Bayesian framework, frequentist approaches exist as well (Yuan et al., 2019).

To the best of our knowledge, none of these methods are design-based or unbiased, and in many cases, the bias can be arbitrary large depending on the choice of historical controls. There have also been attempts to use observational data to improve the generalization of a randomized experiment to a larger population when estimating the population average treatment effect (for example, see Hartman et al. (2015), Kallus et al. (2018), and Rosenman et al. (2018)).

In this chapter, we again focus on estimating the average treatment effect within the experimental sample. However, our goal now is to use external data sets to improve precision of the treatment effect estimate within a target randomized experiment. We introduce a flexible method that allows researchers to employ machine learning algorithms to learn from the observational data, and use the resulting models to improve precision in randomized experiments. Importantly, there is no requirement that the machine learning models are "correct" in any sense. The final experimental results rely only on randomization as the basis for inference and are guaranteed to be exactly unbiased. Thus, there is no danger of confounding biases in the observational data leaking over into the experiment.

This chapter is organized as follows. In Section 4.2 we review notation and present background for covariate adjustment using external data sets. Section 4.3 presents the method. In Section 4.4, we discuss the asymptotic behavior of the estimator. In Section 4.5, we apply the method to simulated data. Section 4.6 concludes.

4.2 Background and Notation

In this chapter, we adopt the notation and setting of Chapter II. Consider a randomized experiment with N participants, indexed by $i = 1, \ldots, N$. Associated with each participant are two potential outcomes t_i and c_i, which are the outcomes we would observe for each observation if they were assigned to treatment and control respectively. Each participant is randomly and independently assigned to treatment with probability p_i, and we observe each participant's outcome Y_i, their treatment assignment indicator T_i, and set of covariates Z_i.

Let $\mathcal{T}$ and $\mathcal{C}$ be the sets of indices for the treated and control units respectively, and let n be the number of treated observations. Our primary parameter of interest is the average treatment effect

$$\tau = \frac{1}{N}\sum_{i=1}^{N}(t_i - c_i).$$

4.2.1 Randomized Controlled Trials and the Remnant

Our goal is to use an external data set to improve precision within a randomized experiment. We will refer to two different data sets: a randomized controlled trial (RCT) and the "remnant." The RCT (or randomized experiment) will be the primary data set of interest. We are interested in estimating the average treatment effect (sometimes referred to as the "sample average treatment effect" or SATE in other literature) within the RCT. The remnant (or external data) will refer to any data outside of the RCT that we will use to make adjustments to the treatment effect estimate within the RCT.

As an example, consider the ASSISTments TestBed (see Heffernan and Heffernan (2014) and Ostrow et al. (2016)). ASSISTments is a computer-based learning platform used by over 50,000 students throughout the United States each year. The TestBed is a program designed for conducting randomized experiments within ASSISTments, and has been made accessible to third-party researchers. These researchers can propose experiments to be run; for example, a researcher may wish to compare video- and text-based instructional feedback. Students working on a specific assignment are individually randomized between the two conditions, and the researcher can then compare the impact of the two conditions on an outcome variable of interest such as homework completion. In this case, the trial of interest will be the RCT. However, a given RCT is like to involve relatively few students. By contrast, the ASSISTments database includes data for hundreds of thousands of users who were not involved in the target RCT. Many of these users may have completed similar assignments (or the same homework assignment at a time period prior to when the RCT was run). These students who were not participants in the target RCT constitute the remnant for the RCT.

4.2.2 Design-Based Covariate Adjustment Using the Remnant

Although we may wish to use the remnant to improve precision when estimating the average treatment effect in the RCT, care must be taken to avoid invalidating the benefits of randomization. For example, simply pooling the data from the remnant and the RCT would undermine the randomization. However, it is possible to make adjustments using the remnant while remaining design-based. One approach is to use "remnant-based residualization" or rebar, which was first developed for matching-based observational studies (Sales et al., 2018b) and then applied to randomized experiments with auxiliary observational data (Sales et al., 2018a).

Using the remnant, one could construct a model f_{ext} to predict outcomes from covariates, which could be applied to the experimental data set to get externally imputed predictions $Z_i^r = f_{\text{ext}}(Z_i)$. So long as these predictions are constructed without using the treatment assignment, Z_i^r will function as a pretreatment covariate (*i.e.,* its value does not depend on the treatment assignment). Next define residuals $R_i = Y_i - Z_i^r$. Each observation has two residualized potential outcomes, $t_i - Z_i^r$ and $c_i - Z_i^r$, and a residualized observed outcome R_i. The externally imputed outcome Z_i^r is independent of the treatment assignment and the difference between the residualized potential outcomes is still the individual treatment effect $t_i - c_i$. We can therefore replace Y_i with R_i in any unbiased estimate for the average treatment effect and still obtain an unbiased estimate. For example, one might estimate the ATE as the simple difference between the treatment groups

$$\hat{\tau}^{\text{Rebar}} = \frac{1}{n}\sum_{i\in\mathcal{T}} R_i - \frac{1}{N-n}\sum_{i\in\mathcal{C}} R_i.$$

This is an unbiased estimate of the average treatment effect. In addition, if the model f_{ext} predicts the outcomes within the randomized experiment well, then the rebar estimator will improve precision over the simple difference estimator. However, if the remnant generalizes poorly to the randomized experiment and f_{ext} performs poorly in the RCT, then rebar could

hurt performance.

4.3 Method

In this section, we present a method that allows us to make covariate adjustments using the remnant while remaining design-based and also solving the primary issue with rebar (*i.e.,* that it can hurt precision). We first present an approach that only uses the remnant to make adjustments, then discuss ways to augment this method with information from the within RCT covariates.

Recall from Chapter II the following estimate for the average treatment effect:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{p_i}(Y - \hat{m}_i)T_i - \frac{1}{1-p_i}(Y - \hat{m}_i)(1 - T_i)\right\}, \tag{4.1}$$

where $\hat{m}_i$ is an estimate for the quantity $m_i = (1 - p_i)t_i + p_i c_i$.

So long as $\hat{m}_i$ is independent of T_i, $\hat{\tau}$ will be unbiased. We define the LOOP estimator as $\hat{\tau}$, where $\hat{m}_i$ is calculated using a leave-one-out procedure. This procedure ensures that $\hat{m}_i$ and T_i are independent. For each observation i, we omit that observation and fit a prediction algorithm to the remaining observations. We then impute t_i and c_i by plugging Z_i into the fitted model. For example, we could fit a random forest to the observations in $\mathcal{T}\backslash i$, then plug Z_i into the fitted model to get $\hat{t}_i$, and use a similar procedure to calculate $\hat{c}_i$. We then set $\hat{m}_i = (1 - p_i)\hat{t}_i + p_i\hat{c}_i$ and plug into equation (4.1). We can use any prediction algorithm to impute the potential outcomes for observation i so long as we exclude that observation when fitting the model.

4.3.1 Design-Based Adjustments Using the Remnant

One way to incorporate the external predictions Z_i^r is to run the LOOP estimator on the RCT data using only Z_i^r as a covariate and using linear regression as the imputation

method. More specifically, we have

$$\hat{t}_i = \hat{\alpha}_i^t + \hat{\beta}_i^t Z_i^r$$

where the regression coefficients are obtained by the OLS regression of Y onto Z for the observations in $\mathcal{T}\backslash i$. Similarly define

$$\hat{c}_i = \hat{\alpha}_i^c + \hat{\beta}_i^c Z_i^r.$$

We refer to this approach as ReLOOP or "remnant-based LOOP."

One advantage of ReLOOP over rebar is that for each observation i, we can use the remaining $N-1$ observations to help determine the best use of Z_i^r for calculating $\hat{m}_i$. For example, if the external predictions are highly accurate imputations of control outcome within the RCT, we might expect that $\hat{\alpha}_i^c \approx 0$ and $\hat{\beta}_i^c \approx 1$, and thus $\hat{c}_i \approx Z_i^r$. However, in the case where the external predictions are noise, then we would expect $\hat{\alpha}_i^c \approx \bar{Y}_{-i}^c$ and $\hat{\beta}_i^c \approx 0$, where $\bar{Y}_{-i}^c$ is the average of outcome of the observations in $\mathcal{C}\backslash i$. Then we would have $\hat{c}_i \approx \bar{Y}_{-i}^c$ (*i.e.,* mean imputation), and the LOOP estimator would be approximately equal to the simple difference estimator. In other words, when the external predictions are not predictive of outcome, the LOOP procedure reverts to mean imputation and performance is not harmed relative to the simple difference estimator. In fact, if we constrained $\hat{\beta}_i^c$ and $\hat{\beta}_i^t$ to be 0, then ReLOOP would be equivalent to the simple difference estimator. If we instead constrained $\hat{\beta}_i^c$ and $\hat{\beta}_i^t$ to be 1, then ReLOOP would be equal to rebar. In ReLOOP, we use the RCT data to estimate the coefficients, and therefore are typically able to make more effective use of the external predictions. Because the fitted coefficients can still be approximately equal to 0 or 1, we would also generally expect ReLOOP to perform no worse than the simple difference estimator or rebar.

Another advantage of ReLOOP is that the external predictions do necessarily need to accurately impute the potential outcomes. Because the external predictions are used as a

covariate within LOOP, it suffices for Z^r to predictive of the outcome within the RCT. Even if the RCT is systematically different from the remnant (for example, if the outcomes in the RCT differ in scale from those in the remnant), the external predictions will still be useful if they are correlated with the experimental outcomes. As an extreme example, if Z^r is anticorrelated with the experimental outcomes, then ReLOOP will improve precision over the simple difference estimator.

4.3.2 Design-Based Adjustments Using the Remnant and RCT Covariates

ReLOOP generally accomplishes our goals: it uses the remnant to make covariate adjustments, while remaining design-based and protecting against harm. However, it fails to make full use of the covariates within the RCT itself. The covariates Z_i are only used to impute potential outcomes to the extent to which they are included in $Z_i^r = f_{\text{ext}}(Z_i)$. However, if the remnant generalizes poorly to the RCT, then Z_i^r may not be predictive of the outcome in the RCT even if the covariates themselves are. Even if Z_i^r improves precision in the RCT, we may be able to further improve performance by adjusting for the RCT covariates as well. In this section, we consider strategies for incorporating the RCT covariates into ReLOOP.

We first augment Z_i with Z_i^r, *i.e.*,

$$\tilde{Z}_i = (Z_{i1}, \ldots, Z_{ip}, Z_i^r),$$

and consider strategies to use $\tilde{Z}_i$. One simple approach is to run LOOP on the RCT data, and replacing Z_i with $\tilde{Z}_i$. We refer to this approach as "ReLOOP+." The hope would be that the prediction algorithm chosen could make use of Z_i^r, while also performing covariate adjustment within the RCT itself. Recall that we can use any prediction algorithm for imputing the potential outcomes. We use random forests as suggested in Chapter [LOOP], and will refer to ReLOOP+ with random forest imputations as "ReLOOP+RF."

ReLOOP+ makes use of both the within RCT covariates and the external predictions.

While Z_i^r is a function of Z_i, the external prediction model f_{ext} is fit on the remnant rather than on the remaining RCT observations. If the remnant extrapolates poorly to the RCT, then Z^r may not be predictive of the outcome. If the external predictions are complete noise, we would generally expect ReLOOP+ to perform similarly to the standard LOOP estimator, as augmenting Z_i with Z_i^r only removes one degree of freedom. However, because the remnant is often much larger than the RCT, it may be possible that f_{ext} provides a more accurate imputation than a model fit to the RCT covariates alone. In such cases, including the external predictions could improve precision.

In cases where Z^r is a strong predictor of the outcome, we may wish to essentially pass through the external predictions directly to equation (4.1), as is done in ReLOOP. Using Z^r in a nonparametric method such as a random forest may be inefficient relative to OLS. Even if we were to use OLS as the imputation method in ReLOOP+, the inclusion of the within RCT covariates could harm performance.

It may not always be clear ahead of time whether ReLOOP or ReLOOP+ will perform better; it depends on how predictive the covariates are within the RCT and the extent to which the remnant generalizes to the RCT. To address this issue, we use an ensemble of the two methods in a manner similar to Chapter [P-LOOP]. We impute two sets of potential outcomes: we obtain $\hat{c}_i^{\text{LS}}$ and $\hat{t}_i^{\text{LS}}$ using the ReLOOP approach, and $\hat{c}_i^{\text{RF}}$ and $\hat{t}_i^{\text{RF}}$ using ReLOOP+RF. We then interpolate between the two sets of potential outcomes. For each i, define

$$\hat{c}_i^{\text{EN}} = \gamma_i^c \hat{c}_i^{\text{LS}} + (1 - \gamma_i^c)\hat{c}_i^{\text{RF}},$$

where γ_i^c is obtained by minimizing the mean squared error between the observed outcomes and the interpolated potential outcomes for the set $\mathcal{C}\backslash i$. That is, we have

$$\gamma_i^c = \underset{x\in[0,1]}{\text{argmin}} \sum_{k\in\mathcal{C}\backslash i} \left\{ Y_k - \left(x\hat{c}_i^{\text{LS}} + (1-x)\hat{c}_i^{\text{RF}} \right) \right\}^2.$$

We define γ_i^t and $\hat{t}_i^{\text{EN}}$ analogously. We refer to this ensemble approach as "ReLOOP+EN."

ReLOOP+EN allows us to combine the strengths of both ReLOOP and ReLOOP+RF. As we will see in Section 4.5, ReLOOP+EN effectively tracks the better performing method of ReLOOP and ReLOOP+RF. The result is a design-based covariate adjustment method that makes adjustments for both the external predictions and the RCT covariates, while protecting against harm relative to the unadjusted estimator.

4.4 Asymptotic Normality of ReLOOP

In Chapter II, we provide both a point estimate and a standard error estimate for the LOOP estimator. However, researchers may often be interested in constructing confidence intervals for the average treatment effect. In this section, we show that the LOOP estimator is asymptotically normally distributed under certain regularity conditions. It follows that we can construct approximate confidence intervals for the average treatment effect using a normal approximation. We first discuss the asymptotic normality of the LOOP estimator generally by adapting the proof given for paired experiments in Chapter III. We then show that LOOP with simple linear regression (and therefore ReLOOP) converges to a normal distribution.

Let $\{(c_i, t_i, Z_i), i = 1, 2, \dots\}$ be an infinite sequence of experimental observations. As before, the potential outcomes and covariates for all pairs are fixed quantities. We observe the first N units in the sequence, and we will consider the behavior of (4.1) as N increases. For a given sample size N, let $\hat{m}_i^{(N)}$ be the estimate for m_i as calculated using the remaining $N-1$ observations in the sample and define the quantities $m_{0i}^{(N)} = \mathrm{E}(\hat{m}_i^{(N)})$ and $\tilde{m}_i^{(N)} = \hat{m}_i^{(N)} - m_{0i}^{(N)}$. For simplicity, we will often suppress the superscript (N) within an equation.

In order for the LOOP estimator to converge to a normal distribution, we need the data and the imputation method to be sufficient well-behaved. In Chapter III, we present conditions (along with the intuition and reasoning for these conditions) that are sufficient for the estimator to be asymptotically normally distributed in paired experiments. The same reasoning holds for the LOOP estimator. In short, these assumptions say that our impu-

tations converge as the sample size increases and that no single observation dominates the remaining observations asymptotically. More specifically, suppose the following conditions hold:

1. There exists some $0 < C < \infty$ and $q > 0$ such that for all i,

$$\text{Var}(\tilde{m}_i) = \text{Var}(\hat{m}_i) \leq C/N^q.$$

2. Let ρ_{ij} be the correlation of $\tilde{m}_i U_i$ and $\tilde{m}_j U_j$, and $\bar{\rho} = \frac{\sum_{i \neq j} \rho_{ij}}{N(N-1)}$. We assume that

$$N^{1-q}\bar{\rho} \longrightarrow 0.$$

3. For each observation i, we assume that the limit of $m_{0i}^{(N)}$ exists and denote the limit as $m_{\infty i}$. We also assume

$$\frac{1}{N}\sum_{i=1}^{N}\left(m_{0i}^{(N)} - m_{\infty i}\right)^2 \longrightarrow 0.$$

4. There exists $0 < K < \infty$ such that

$$\frac{\sum_{i=1}^{N}(m_i - m_{\infty i})^2}{N} \longrightarrow K,$$

and

$$\max_{i=1,\ldots,N} \frac{(m_i - m_{\infty i})^2}{\sum_{k=1}^{N}(m_k - m_{\infty k})^2} \longrightarrow 0.$$

These conditions are sufficient for the LOOP estimator to be asymptotically normally distributed.

Proposition 1. *Let* $V_N = \sum_{i=1}^{N} \frac{(m_i - m_{\infty i})^2}{p_i(1-p_i)}$. *Assume conditions (1) through (4) hold. In addition, suppose there exists* $0 < \epsilon < 0.5$ *such that* $\epsilon < p_i < 1 - \epsilon$ *for all* i. *Then the quantity* $N(\hat{\tau} - \tau)/\sqrt{V_N}$ *converges in distribution to a standard normal random variable.*

See Appendix M for a proof.

Assumptions (1) through (4) describe the behavior of the imputation method and data. We generally do not prove that they hold for a specific imputation method; however, we do prove that the assumptions hold in the case of simple linear regression. Consider the case where $\hat{t}_{i,\text{slr}}$ is obtained by regressing the outcomes Y onto a single covariate Z for the observations in $\mathcal{T}\backslash i$. Similarly, $\hat{c}_{i,\text{slr}}$ is obtained using leave-one-out regression for the control units. Define $\hat{m}_{i,\text{slr}} = (1-p_i)\hat{t}_{i,\text{slr}} + p_i\hat{c}_{i,\text{slr}}$.

Proposition 2. *Let ρ^t be the limiting correlation between t and Z, and ρ^c be the limiting correlation between c and Z. Suppose the following conditions hold:*

1. *$p_i = p$ for all i*
2. *t_i, c_i, and Z_i are bounded*
3. *The quantities $\frac{1}{N}\sum_{i=1}^N c_i$, $\frac{1}{N}\sum_{i=1}^N c_i^2$, $\frac{1}{N}\sum_{i=1}^N t_i$, $\frac{1}{N}\sum_{i=1}^N t_i^2$, $\frac{1}{N}\sum_{i=1}^N Z_i$, and $\frac{1}{N}\sum_{i=1}^N Z_i^2$ converge*
4. *$-1 < \rho^t < 1$ and $-1 < \rho^c < 1$*

Then conditions (1) through (4) hold for $\hat{m}_{i,slr}$.

See Appendix N for a proof. Because ReLOOP is LOOP with a single covariate Z_i^r, it follows that ReLOOP is also asymptotically normally distributed.

4.5 Simulations

We examine the performance of ReLOOP and ReLOOP+ using various sets of simulations. We first investigate the effects of varying sample size, the predictive power of the covariates Z, and the predictive power of the remnant-based predictions Z^r. We also consider a simulation in which we vary the extent to which the remnant extrapolates to a randomized experiment.

4.5.1 Simulation 1

We generate our data using a model that is parameterized in such a way that we are able to independently vary these three quantities (sample size, the predictive power of the covariates, and the the predictive power of the remnant predictions).

We simulate a randomized experiment in which there are N subjects. For each subject i there are two covariates, Z_{i1} and Z_{i2}, which are independent and Unif$(0, 10)$. The potential outcomes are generated from the following linear model:

$$\begin{aligned} a_i &= 2Z_{i1} + Z_{i2} + \delta_i \\ c_i &= \frac{a_i}{\sigma_a} \\ t_i &= c_i + 3 \end{aligned}$$

where $\delta_i \sim \mathrm{N}(0, \sigma^2)$ and $\sigma_a^2 \equiv \mathrm{Var}(a_i) = \frac{500}{12} + \sigma^2$. By generating our potential outcomes as above, we have defined our generative model so that the potential outcomes have unit pooled variance. We can alternatively write the observed outcome as:

$$Y_i = 3T_i + \frac{2}{\sigma_a} Z_{i,1} + \frac{1}{\sigma_a} Z_{i,2} + \epsilon_i$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2/\sigma_a^2)$.

For each observation, we also simulate remnant predictions Z_i^r by taking the true c_i and adding a normally distributed noise term with mean 0 and variance σ_{rem}^2.

Again, our goal is to investigate variations in sample size, the predictive power of the covariates, and the predictive power of the remnant predictions. Sample size is directly indexed by N. We can index the predictive power of the covariates with

$$R_{cov}^2 = 1 - \frac{\sigma^2}{\sigma_a^2}.$$

Similarly, the predictive power of the remnant prediction Z_i^r is

$$R^2_{rem} = 1 - \frac{\sigma^2_{rem}}{\text{Var}(Z_i^r)} = 1 - \frac{\sigma^2_{rem}}{1 + \sigma^2_{rem}}.$$

Thus, given a desired R^2_{cov} and R^2_{rem}, the corresponding values of σ^2 and σ^2_{rem} are:

$$\sigma^2 = \frac{1 - R^2_{cov}}{R^2_{cov}} \times \frac{500}{12}$$
$$\sigma^2_{rem} = \frac{1 - R^2_{rem}}{R^2_{rem}}.$$

We perform three sets of simulations. In each, we vary one of the quantities N, R^2_{cov}, or R^2_{rem} while holding the other two fixed. For each set of simulations, we compare the following methods:

1. Simple difference estimator
2. LOOP: Includes only the covariates Z_1 and Z_2. Uses a random forest as the imputation method.
3. ReLOOP: Uses only the remnant predictions Z^r. Uses OLS as the imputation method.
4. ReLOOP+RF: Uses Z_1 and Z_2 and the remnant predictions Z^r as covariates. Uses a random forest as the imputation method.
5. ReLOOP+EN: Interpolates between the previous two methods.

We use the following simulation procedure. For a given set of N, R^2_{cov}, and R^2_{rem}, we perform $k = 1000$ trials. For each trial, we generate a set of covariates, potential outcomes, a treatment assignment vector, and remnant predictions as described above. We then produce an estimate of the variance of each method. Next, we average the estimated variance across the k trials. Finally, we plot the average variance for each of the adjustment methods relative to the variance of the simple difference estimator. That is, for each method (2) – (5) we plot (average variance of method) / (average variance of simple difference estimator).

Varying Sample Size For these simulations, we hold the predictive power of the covariates and remnant predictions constant and vary the sample size. We consider four scenarios: (1) $R^2_{rem} = 0.25, R^2_{cov} = 0.25$; (2) $R^2_{rem} = 0.75, R^2_{cov} = 0.25$; (3) $R^2_{rem} = 0.25, R^2_{cov} = 0.75$; and (4) $R^2_{rem} = 0.75, R^2_{cov} = 0.75$. In each scenario the sample sizes considered are $N = 30, 40, 50, 75, 100, 150, 200$. Results are in Figure 4.1.

We first note that all methods perform better than the simple difference estimator (the relative variances are all less than 1), suggesting that adjustment is typically helpful. In addition, we observe that both variants of ReLOOP+ typically outperform ordinary LOOP, indicating that incorporating the remnant predictions is typically beneficial. The exception is when the covariates are highly informative but the remnant predictions are not ($R^2_{rem} = 0.25, R^2_{cov} = 0.75$, lower left panel) and the sample size is small, in which case ReLOOP+RF performs slightly worse than ordinary LOOP.

We also observe that ReLOOP+EN does well at tracking its better performing component (ReLOOP or LOOP+RF). It performs reasonably well at small sample sizes, and quickly converges to near optimal at larger sample sizes. In some cases ReLOOP+EN performs better than either component individually.

Varying Predictive Power of Remnant Prediction In these simulations, we hold the predictive power of the covariates and sample size constant and vary R^2_{rem}. We again consider four scenarios, with N fixed at either 30 or 60, and R^2_{cov} fixed at either 0.25 or 0.75. The values of R^2_{rem} considered are $R^2_{rem} = 0.05, 0.15, ..., 0.85, 0.95$. Results are in Figure 4.2.

Once again, we observe that ReLOOP+EN tends to perform at least as well as either of its two components. This is particularly true for $N = 60$, where ReLOOP+EN closely follows (or drops below) the lower of the component lines. As expected, the three methods that incorporate the remnant predictions all improve as R^2_{rem} increases, while the performance of LOOP stays constant. We see that ReLOOP+EN is notably outperformed by LOOP only when R^2_{rem} is lower than R^2_{cov} and the sample size is small.

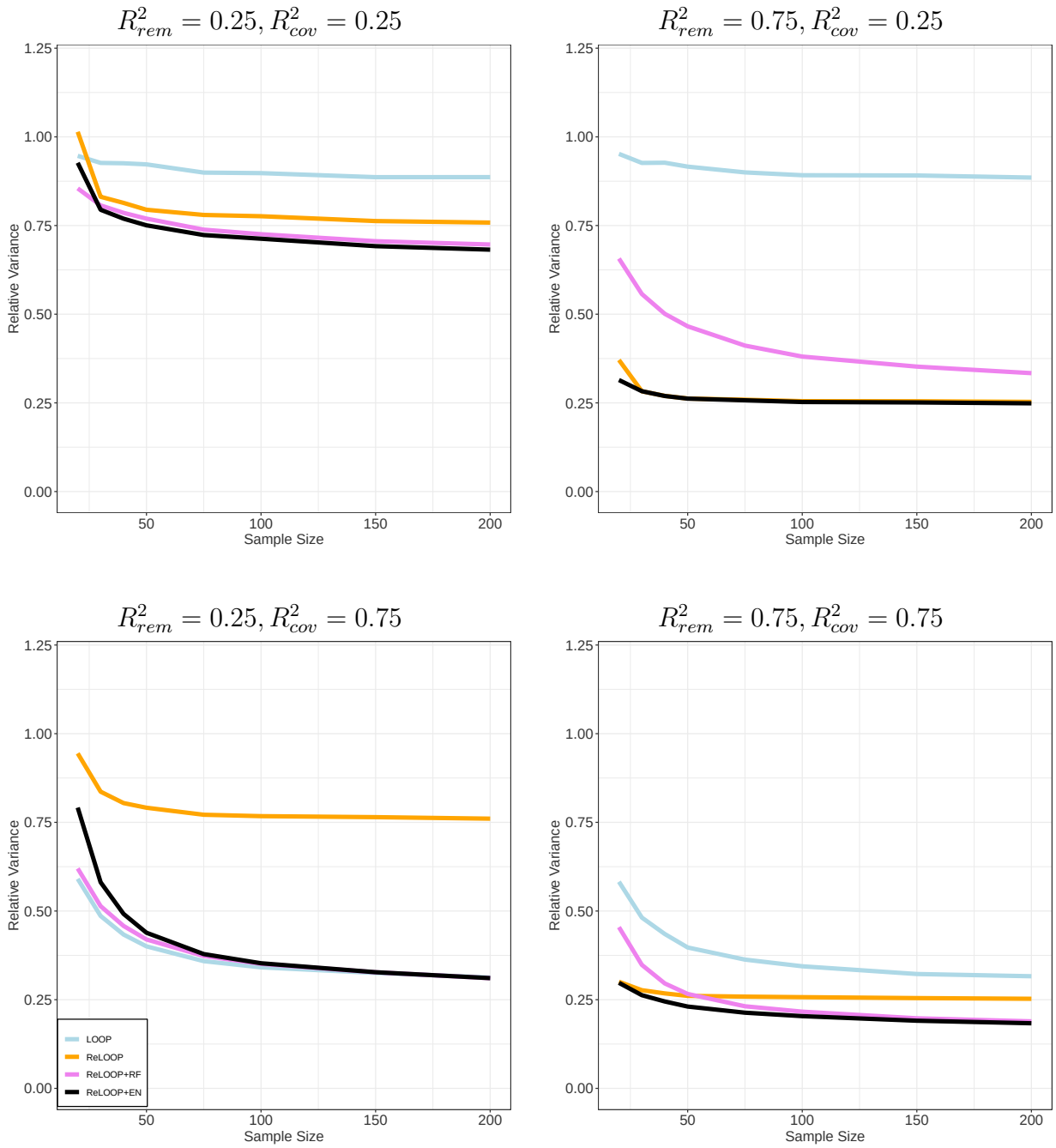

Figure 4.1: Varying sample size.

Varying Predictive Power of Covariates For this simulation, we hold the predictive power of the remnant predictions and sample size constant and vary $R^2_{cov} = 0.05, 0.15, ..., 0.85, 0.95$. We consider four scenarios: (1) $N = 30, R^2_{rem} = 0.25$; (2) $N = 30, R^2_{rem} = 0.75$; (3) $N = 60, R^2_{rem} = 0.25$; and (4) $N = 60, R^2_{rem} = 0.75$. Results are in Figure 4.3.

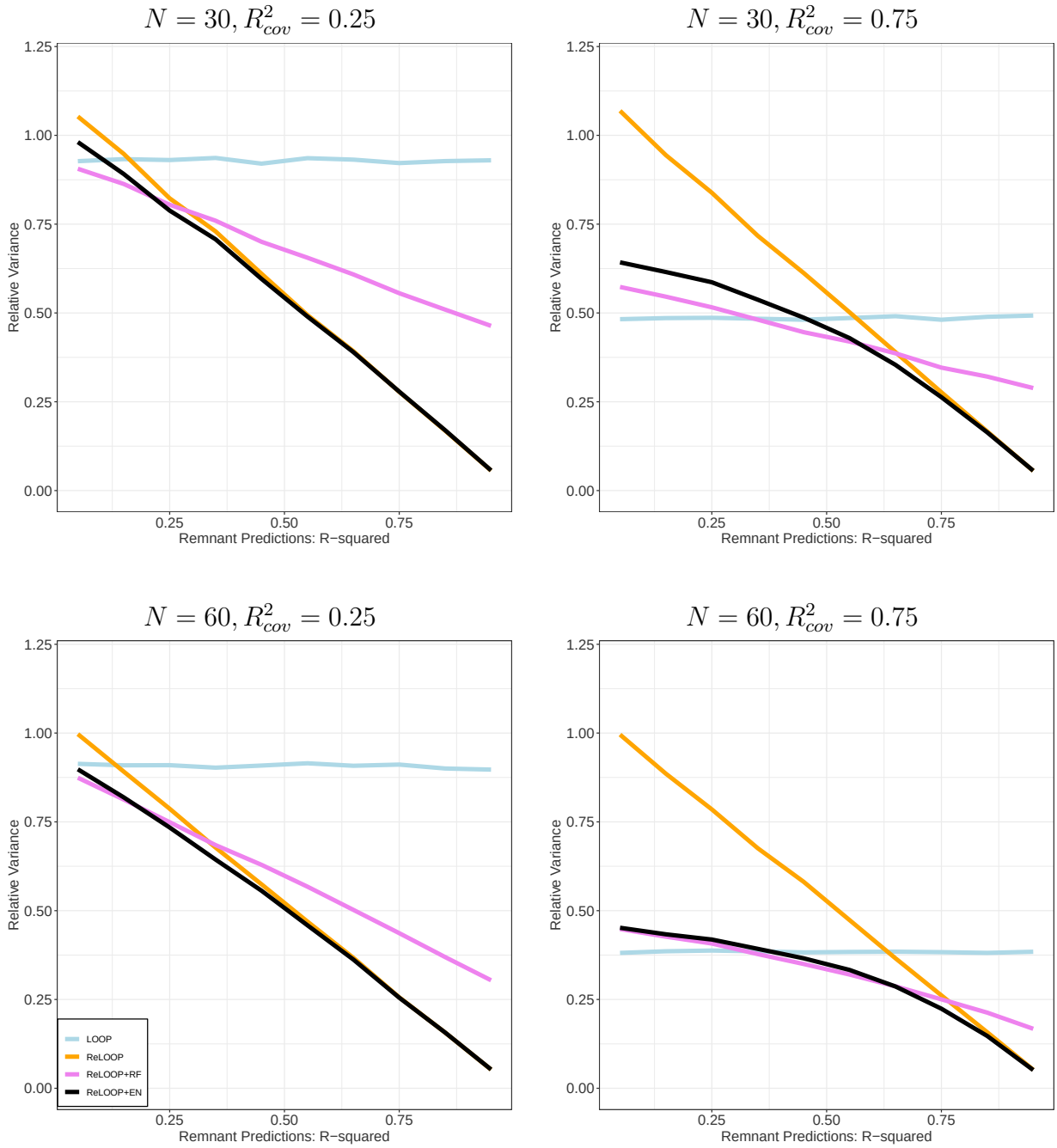

Figure 4.2: Varying R^2_{rem}

Here the performance of ReLOOP stays constant, as it makes use only of the remnant predictions, not the covariates. The remaining methods all improve as R^2_{cov} increases. As before, we can see that ReLOOP+EN tracks its better performing component reasonably well, especially when $N = 60$.

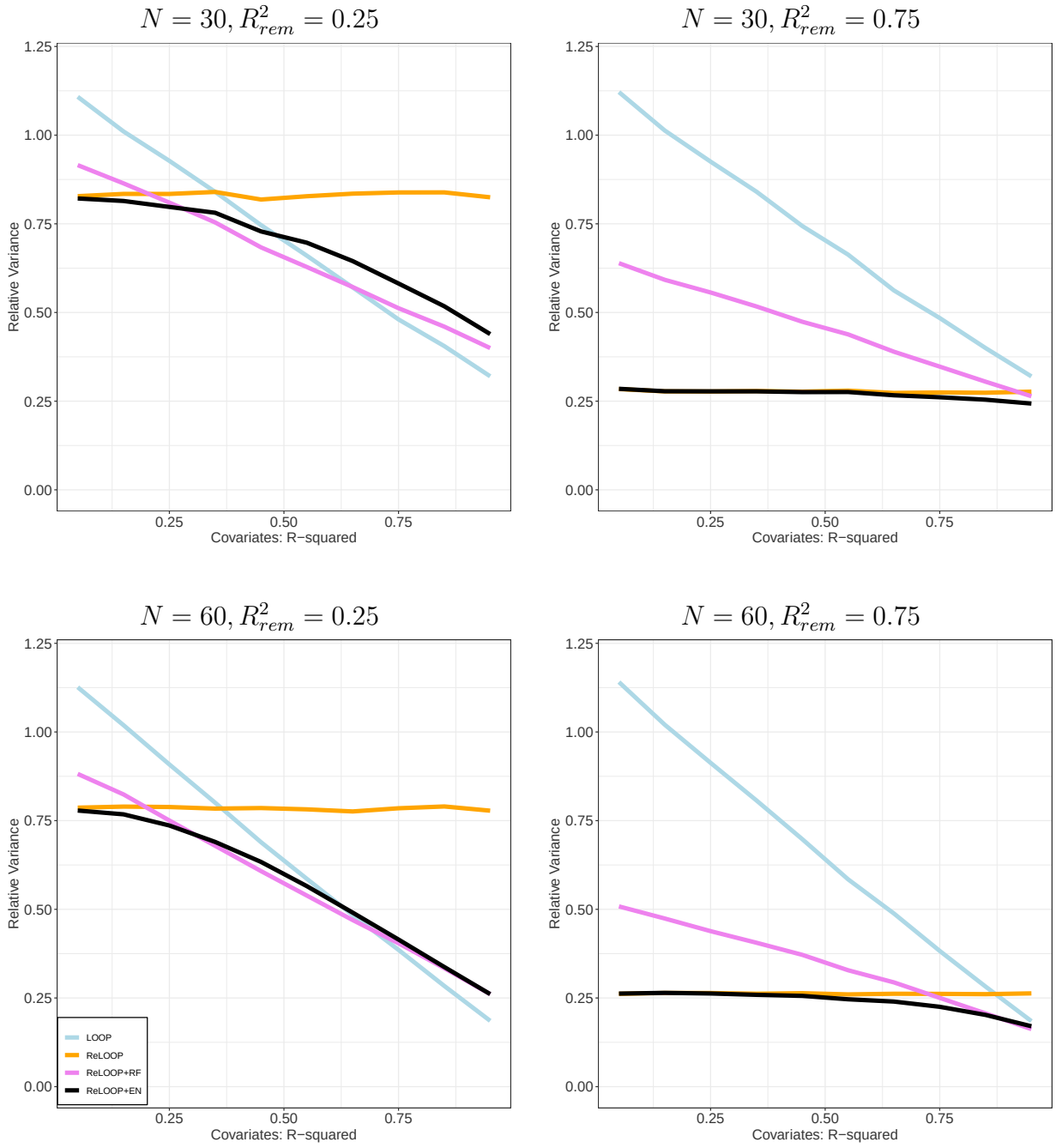

Figure 4.3: Varying R^2_{cov}

4.5.2 Simulation 2

We next consider an example where we simulate both an RCT and a remnant, while varying the extent to which the remnant extrapolates to the RCT. We first simulate a randomized experiment with $N = 30$ observations and $p = 10$ covariates. For each observation,

we generate a vector of covariates $Z_i = (Z_{i,1}, \ldots, Z_{i,10})$, where $Z_{i,j}$ are independent and Unif$(0, 5)$. The potential outcomes are generated as

$$c_i = \beta_1 Z_{i,1} + \beta_2 Z_{i,2} + \beta_3 Z_{i,3} + \epsilon_i$$
$$t_i = c_i + 3$$

where $(\beta_1, \beta_2, \beta_3) = (1.5, 1, -0.5)$ and ϵ_i are independent and normally distributed with mean 0 and standard deviation 3. We also simulate a remnant with $N = 500$ observations. The observations in the remnant are generated from the same distribution as the control potential outcomes in the randomized experiment; however, we also add noise to the coefficients. That is, we set $\beta_{j,rem} = \beta_j + \delta_j$ where $\delta_j \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2_{coef})$ for $j = 1, 2, 3$. We next generate a vector of covariates $V_i = (V_{i,1}, \ldots, V_{i,10})$ for each observation, where $V_{i,j}$ are independent and Unif$(0, 5)$. Then each observation in the remnant is generated as

$$U_i = \beta_{1,rem} V_{i,1} + \beta_{2,rem} V_{i,2} + \beta_{3,rem} V_{i,3} + \epsilon_{i,rem}$$

where $\epsilon_{i,\text{rem}}$ are independent and normally distributed with mean 0 and standard deviation 3. To obtain the remnant predictions Z^r, we regress U onto V.

We perform a set of simulations in which we vary $\sigma_{coef} = 0.25, 0.5, \ldots, 3.75, 4$. We use the same procedure as outlined in Section 4.5.2. For each value of σ_{coef} we perform $k = 10,000$ trials, and compare the performance of the simple difference estimator, LOOP, ReLOOP, ReLOOP+RF, and ReLOOP+EN. We plot the variance of each method relative to the variance of the simple difference estimator in Figure 4.4.

The ReLOOP methods perform well when the remnant extrapolates well to the randomized experiment (*i.e.*, σ_{coef} is low). As σ_{coef} increases, the variance of the ReLOOP methods increase and ReLOOP is eventually outperformed by LOOP. However, even for $\sigma_{coef} = 4$, ReLOOP+RF and ReLOOP+EN both outperform the standard LOOP estimator, indicating there is some value gained from incorporating the remnant.

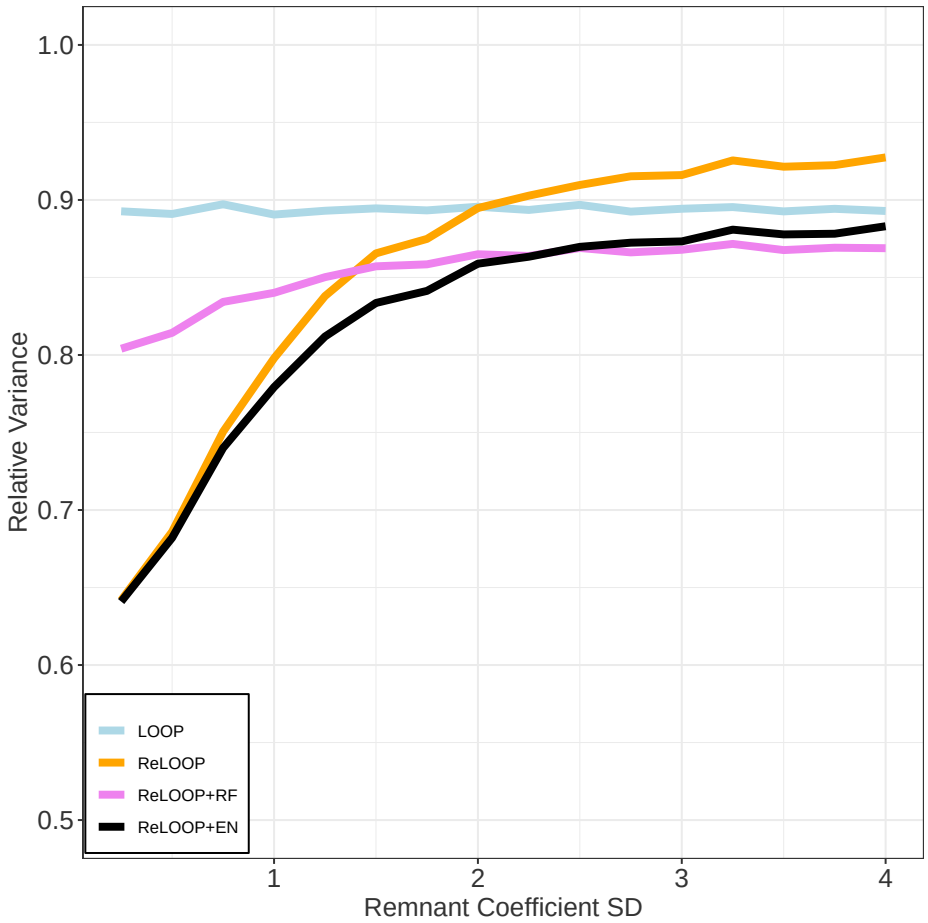

Figure 4.4: Varying σ_{coef}

4.6 Discussion

In this chapter, we have presented a method that seeks to combine the strengths of randomized experiments and observational studies. Randomized experiments allow for unbiased estimation of treatment effects with minimal assumptions. However, precision may be limited due to sample size constraints. On the other hand, observational studies may have large sample sizes, but suffer from confounding bias and require an analyst to make untestable modeling assumptions. ReLOOP also allows us to make covariate adjustments to a randomized experiment using an observational data set, while remaining design-based. The method is unbiased and will generally not harm precision relative to the simple difference estimator. In cases where the observational data set extrapolates well to the randomized experiments,

we can improve precision considerably.

In some cases, the external data set may not be helpful or may be less helpful than the RCT covariates alone. We also propose a method, ReLOOP+, that makes adjustments using both the observational data set and the RCT covariates. However, it is not necessarily clear ahead of time whether ReLOOP or ReLOOP+ will perform better. We therefore propose an ensemble method that interpolates between the two methods. In Section 4.5, we observe that this ensemble method does well at tracking the better performing method. We also see that both ReLOOP and ReLOOP+ generally perform at least as well as the simple difference estimator even when the covariates and the external predictions are both noise.

CHAPTER V

A Tournament Classifier

5.1 Introduction

High-dimensional data sets are increasingly common in many fields; however, the analysis of these data sets presents unique challenges. For example, many standard statistical methods assume or require that the sample size n exceeds the number of variables p. The number of parameters to be estimated within a model generally increases as the dimension of a data set grows, causing some methods to fail entirely. For example, it is well known that the ordinary least squares solution is not identifiable when $p > n$. Similarly, when using linear discriminant analysis (LDA) with high-dimensional data, new estimation techniques are required because the sample covariance matrix is non-invertible.

Many solutions have been proposed to address these challenges. Some methods, such as support vector machines (Hearst et al., 1998), are well suited for high dimensional data. We can also use dimension reduction or feature selection techniques (*e.g.*, Guyon and Elisseeff (2003)) to turn a high dimensional problem into a low dimensional one. In other cases, we may modify existing techniques for low dimensional data, often by putting restrictions on the fitted model. For example, we can add a penalty term to linear or logistic regression to shrink coefficients towards zero (see Tibshirani (1996), Zou and Hastie (2005), Hastie et al. (2009)). Linear discriminant analysis can be modified by adding a penalty (Witten and Tibshirani, 2011) or by putting restrictions on the covariance matrix, such as the "independence" rule

(*e.g.*, Bickel et al. (2004)).

Even in cases where we can fit a model, there may be issues with performance in high-dimensional settings due to overfitting. For example, an increasing number of features does not necessarily translate to an increase in the number of variables with a true relationship to the outcome. In many high-dimensional data sets, the true signal is sparse, and only a few of the features have a true relationship with the outcome. However, classifiers may overfit to chance variation in null features, resulting in worse accuracy when generalizing to new data. Some of the approaches outlined above either explicitly or implicitly assume that the signal from the predictors is sparse. For example, the lasso shrinks many coefficients to zero and is therefore well suited to situations where the signal is sparse. Sparse extensions also exist for both logistic regression (*e.g.*, Shevade and Keerthi (2003) and Abramovich and Grinshtein (2018)) and linear discriminant analysis (*e.g.*, Guo et al. (2007), Trendafilov and Jolliffe (2007), Shao et al. (2011), and Witten and Tibshirani (2011)).

In this chapter, we propose a framework for classification to address these concerns. We group the predictors and perform a "tournament": in each round of the tournament, we combine each group of predictors into a single predictor. We then group the combined predictors, perform another round of the tournament, and continue until we've combined all of the predictors into a single predictor. This competition results in a sort of variable selection: strong predictors are more likely to make it through the tournament, and spurious predictors are more likely to get weeded out. We also propose a specific approach within this framework. We use leave-one-out LDA models within the tournament, and shrink coefficients for weakly predictive variables to zero to promote sparsity. Methods that produce sparse coefficient or discriminant vectors can improve accuracy by reducing the amount of overfitting to null features. However, it may still be possible to overfit to the remaining predictors, which we address using the leave-one-out procedure. Our method is most similar to the sparse LDA methods, as we also use LDA to obtain a sparse classifier in a high-dimensional setting.

One advantage of this method is its flexibility. Although we propose a specific algorithm

for performing the tournament, the approach for comparing predictors can be readily generalized. Our proposed algorithm constructs a linear classifier, where we shrink coefficients for weakly predictive variables towards zero. However, we may wish to modify the method to obtain a non-linear classifier or for data sets where the signal is not sparse. The framework can also be used to take a feature selection or dimension reduction approach, or to adjust the dependence structure between variables.

This chapter is organized as follows. In Section 5.2, we provide a motivating example. Section 5.3 introduces the method. In Section 5.4, we apply the method to high-dimensional microarray data sets. Section 5.5 concludes.

5.2 Motivation

Consider a data set with p predictors $X_j \in \mathbb{R}^n, j = 1, \ldots, p$ and labels $Y_i \in \{-1, 1\}$ for observations $i = 1, \ldots, n$, and let $\mathcal{S}$ be the set of observations with class label 1 (*i.e.*, $\{i : Y_i = 1\}$). In this chapter, we will assume that the classes are balanced, but note that the method should be easily generalizable to cases where there is class imbalance.

We consider a hypothetical high-dimensional microarray data set. For example, the measured outcome Y could be positive or negative for breast cancer and the predictors X the measured gene expression levels. To illustrate suppose the values of the predictors come from a data generated process

$$(X_{i1}, \ldots, X_{ip}) = Y_i \beta + \epsilon_i,$$

where β is a fixed p-dimensional vector and ϵ_i is a multivariate normal random variable with mean 0 and covariance matrix Σ. In such data sets, often only a few genes will be differentially expressed between class labels. That is, relatively few entries of β are non-zero and are thus predictive of the outcome. In addition, the dependence between predictors may often be small after appropriate preprocessing. Expression for genes sharing the same

biological pathway may be highly correlated; however, this dependence would be for a small number of genes in the same pathway, and we would expect most genes to be uncorrelated (after appropriate preprocessing).

Classification models in such high-dimensional settings can also be particularly susceptible to overfitting. Although only a few variables may have a relationship with the outcome, we would expect some number of predictors to have a spurious relationship with the class label. That is, within the training set, these variables will be predictive of outcome by random chance even if there is no true relation. Often they will only have a weak relationship compared to predictors that have a true relationship with the outcome. However, in cases where there are many such spurious variables, the combined effect could attenuate the signal of the truly predictive variables and lower accuracy.

In this chapter, we propose a general algorithm for high-dimensional classification, which we call the tournament classifier. We then propose a specific approach to address the issues described in this section. Ideally the fitted model would ignore these spurious predictors; however, even if the model discards these predictors, it may still overfit to the signal in the remaining predictors. To deal with spurious predictors, we have all of the variables compete within a tournament, with the expectation that stronger predictors will make it through the tournament. However, spurious predictors may still make it through to the end. We therefore propose regularization that shrinks coefficients for unimportant predictors towards zero, resulting in a sparse classifier and helping to reduce overfitting. To address overfitting to the remaining predictors, we propose the use of sample splitting. We allow for dependence between the predictors; however, we add constraints to prevent overfitting.

5.3 Tournament Classifier

We form groups of predictors and perform what we call a tournament. In each round of the tournament, the predictors within each group "compete" against each other, and are combined into a single predictor. For example, we could pair the predictors, then combine

each pair by taking a linear combination of the predictors within a pair that minimizes some loss function. In the next round of the tournament, we form groups from the combined predictors and combine the predictors within each group into a single predictor. The process continues until all of the predictors have been combined into a single classifier.

The method used to combine the predictors can vary. We present one potential choice in this section, in which predictors are grouped into pairs. This method for comparing predictors can be thought of in two equivalent ways.

Perspective 1 The first way to view the comparison method is as a leave-one-out linear discriminant analysis. In the first round of the tournament, we combine each pair into a single predictor, resulting in $p/2$ predictors. For a pair of predictors X_1 and X_2, we leave out the i-th observation and for the remaining observations, fit a bivariate LDA model where the coefficients are constrained to be positive. This yields a linear classifier into which we can plug in X_{i1} and X_{i2} to obtain a value $X_{i,\{1,2\}}$. Repeating this procedure for each observation results in a combined predictor $X_{\{1,2\}} = (X_{1,\{1,2\}}, \ldots X_{n,\{1,2\}})$. We perform the same leave-one-out procedure on the remaining predictors to yield a total of $p/2$ combined predictors.

This method can be thought of as a relaxation of the independence rule with LDA. By fitting bivariate LDA models, we allow for pairwise dependence between variables. In addition, constraining the coefficients to be positive ensures that we preserve the sign of the marginal effect for each predictor. Thus we allow for some pairwise dependence, but not so much that the sign of the effect can be reversed by the model. Finally, we also employ a leave-one-out approach to help reduce overfitting.

Perspective 2 We can also view this comparison method as minimizing the mean squared error of a linear combination of the transformed predictors. Again, we leave out the i-th observation. Then for the remaining observations, we transform each predictor such that the mean of each transformed predictor will be 1 for the observations in $\mathcal{S}\backslash i$, and -1 for

observations in $\mathcal{S}^c \backslash i$. In addition, for a predictor that predicts the class label Y well, we would expect the values of the transformed predictor for observations in $\mathcal{S} \backslash i$ to be closely clustered around 1, and the values for observations in $\mathcal{S}^c \backslash i$ to be closely clustered around -1. For predictors that do not predict the class label well, the transformed predictor will fluctuate more.

We next compare transformed predictors by interpolating between each pair. For a pair X_1 and X_2, we select a weight

$$w_{i,\{1,2\}} = \underset{x \in [0,1]}{\operatorname{argmin}} \sum_{k \neq i} \{Y_k - (xZ_{k1} + (1-x)Z_{k2})\}^2$$

where Z_{kj} is the transformed value of the k-th observation of predictor j. We then set $Z_{i,\{1,2\}} = w_{i,\{1,2\}} Z_{i1} + (1 - w_{i,\{1,2\}}) Z_{i2}$, and $Z_{\{1,2\}} = (Z_{1,\{1,2\}}, Z_{2,\{1,2\}}, \ldots, Z_{n,\{1,2\}})$. Details on the transformation and the calculation of the weights are given in Appendix O. We repeat this procedure for all the pairs of predictors and for all observations, resulting in $p/2$ combined predictors.

Completing the Tournament After completing the first round of the tournament, we pair the combined predictors and apply the comparison method to these new pairs, resulting in $p/4$ predictors. We then continue the tournament until all of the predictors are combined into a single predictor $Z^f = Z_{\{1,\ldots,p\}}$. Because each pair of predictors is combined by taking n linear combinations (one for each left out observation), each term of the final predictor can be written as a linear combination

$$Z_i^f = \sum_{k=1}^{p} a_{ik} Z_{ik},$$

where the a_{ik} are determined by the weights calculated throughout the tournament.

The completed tournament yields n different linear classifiers, one corresponding to each

left out observation. We combine these by taking a simple average. That is, we set

$$a_k = \frac{1}{n} \sum_{i=1}^{n} a_{ik}$$

for $k = 1, \ldots, p$. This yields a classifier $f(z) = \text{sign}(\sum_{k=1}^{p} a_k z_k)$, where $z = (z_1, \ldots, z_p)$ is a transformed new observation.

The results of the tournament classifier depends in part on the grouping of the predictors. If we had paired the variables differently, the results would likely end up being different. To address this issue, we perform the procedure T times, randomly pairing the variables each time. For each iteration t, we have a classifier $f_t(z) = \text{sign}(\sum_{k=1}^{p} a_{k,t} z_k)$. This results in T different classifiers, which we can use to make predictions (*e.g.*, by majority vote).

5.3.1 Overfitting

In cases where the signal is sparse (*i.e.*, relatively few variables are predictive of outcome), the method can still give weights to the noise predictors when combining the predictors. While these weights will generally be small, the accumulation of small errors due to the large number of predictors can harm performance. We therefore propose a regularization parameter r_0 to shrink the weights towards zero. The goal of this regularization would be to obtain a sparse coefficient vector and improve the generalization of the classifier to new data.

Define a value $r \in [0, 0.5]$. For a particular pair of predictors, we select interpolation weights w_i using the leave one out procedure discussed above. We set $w_i' = 0$ for $w_i < r$, $w_i' = 1$ for $w_i > 1 - r$, and $w_i' = w_i$ otherwise. For example, if we have $r = 0.25$, any weights larger than 0.75 would be set to 1. We then interpolate between the two predictors using w_i' instead of w_i. This has the effect of zeroing out weaker predictors, and the value of r determines how weak a predictor must be to be zeroed out. Higher values of r result in more regularization. In the case where $r = 0.5$, only one predictor would be kept. When $r = 0$,

the regularization parameter would have no effect on the interpolation.

We vary the value of r as the tournament progresses. For example, we might expect that predictors that make it further into the tournament are more likely to have a true relationship with the outcome, and thus wish to taper the amount of regularization as the tournament proceeds. Let r_0 be some value between 0 and 0.5. We then set $r = r_0$ for the first round of the tournament, $r = 0$ for the last round of the tournament, and linearly interpolate between the two values for the rounds in between.

While this regularization parameter shrinks the coefficients for unimportant variables to zero, it is still possible to overfit to the remaining predictors. As described above, we address this in part by using sample splitting. We also limit the extent to which we model dependence between predictors, as modeling the dependence without constraints could result in overfitting as well. We do this by only modeling pairwise dependence and by constraining the coefficients to be positive. We can think of the regularization described in this section as an extension of this constraint, as it further limits the potential values of the coefficients in the model.

5.3.2 Advantages

Flexibility In this section, we have outlined a specific choice for performing the tournament. However, as we noted earlier, one advantage of the tournament classifier is its flexibility, and we can modify various aspects of the procedure to suit the specific application. For example, we could modify the size of the groups, how we process the data, the method we use to compare the processed predictors (such as changing the loss function), or how we combine the various classifiers produced by cross validation or permutation. We could also change our regularization method. In this section, we suggest choosing an initial regularization value r_0, then linearly tapering the regularization towards zero as we continue the tournament. However, we could also consider a regularization scheme in which the tapering is non-linear, or choose an entirely different method for regularization.

The tournament classifier may also be used as a tool for dimension reduction or variable selection. In the procedure described, we continue the tournament until all of the predictors have been combined into a single predictor. However, we may instead choose to stop the tournament early, such as when we have a certain number of combined predictors left. These predictors could then be passed onto a low dimensional classification algorithm. In the case where we set $r = 0.5$ for all rounds of the tournament, each comparison would necessarily choose only one of the predictors. We can also choose to select covariates by running the entire tournament for some choice of regularization parameter, and then selecting the covariates with non zero coefficients.

Computational Efficiency Another advantage of the tournament classifier is computational efficiency. Many of the steps in the algorithm are easily parallelizable. We perform several iterations of the classifier, and randomly group the predictors each time. These iterations can all be computed separately. For each iteration, we can also compute the comparisons for a given round in parallel. Thus, given enough processors, the entire model can be fit in the same amount of time it takes to do $\log_2 p$ pairwise comparisons. In addition, while the comparisons themselves are done using a leave-one-out procedure, both the processing step and the calculation of the leave-one-out weights can be done using vectorized calculations. In combination, these properties suggest that the tournament classifier model can be fit quickly.

The tournament classifier is also a good computational fit for use as an imputation method within the LOOP estimator when the outcome is binary. As discussed in Chapter II, the LOOP estimator can be computationally intensive due to the leave-one-out procedure. However, by using random forests, we can take advantage of out-of-bag predictions rather than fitting a separate random forest for each observation. We can use a similar approach using the tournament classifier. For each iteration, we randomly leave out a subset of observations rather than using all of the observations. We then fit the tournament classifier

model to the remaining observations. This results in a set of models (one for each iteration), and we would expect each observation to be out of bag in a subset of these models. To impute the potential outcomes for the i-th observation, we use the subset of models where that observation was left out.

5.4 Results

In this section, we apply the tournament classifier first to simulated data and then to 22 microarray data sets.

5.4.1 Simulation

We consider a hypothetical data set generated under the setting described in Section 5.2. Recall that we have n observations generated from the model

$$(X_{i1}, \ldots, X_{ip}) = Y_i\beta + \epsilon_i,$$

where β is a fixed p-dimensional vector and ϵ_i is a multivariate normal random variable with mean 0 and covariance matrix Σ.

We consider several scenarios. We set $\Sigma = I_p$, where I_p is the p-dimensional identity matrix. We then fix the number of observations and non-null features, and vary the number of null features. In each scenario, we generate 50 training observations and 100 test observations, and set $\beta = (6, 3, 2, 0, \ldots, 0)$. That is, the first three coefficients are 6, 3, and 2, and the remaining $p - 3$ coefficients are all 0. We then vary the sample size, setting $p = 2^4, 2^5, \ldots, 2^{11}$. We can therefore observe how the performance of various algorithms change as the dimension increases in a setting where the signal is sparse.

For each set of parameter values, we generate 100 different sets of data and calculate the training and test set error for each data set using DLDA (*i.e.*, diagonal LDA or LDA with the independence rule), lasso, and the tournament classifier. We average the prediction

accuracies across the 100 different sets of data to obtain a measure for the performance of each classification algorithm under each scenario. For DLDA, we select the top $k = 3$ predictors as ranked by largest absolute value of marginal t-test statistics. For lasso, we use the `scikit-learn` implementation in Python where the penalty term λ is automatically selected within the `LassoCV` function. For the tournament classifier, we fit the model using regularization parameters $r_0 = 0$ and 0.5, corresponding to no regularization and high amounts of regularization.

In Table 5.1, we compare the results for each scenario. The performance of all three methods is similar when the dimension is low. However, as the number of predictors increases, the performance of both DLDA and the no regularization version of the tournament classifier degrade more rapidly. In fact, DLDA and no regularization tournament classifier perform no better than random guessing when $p = 2^{11}$. For both the high regularization tournament classifier and lasso, the accuracy drops more slowly. In addition, the high regularization version of the tournament classifier outperforms the other methods for all sample sizes. We also do not necessarily select an optimal r_0, suggesting that we could improve performance further by doing so (*e.g.*, by cross validation).

Table 5.1: Simulation Results: Test Set Accuracy by Method

Sample Size	**DLDA**	**Lasso**	**TC – No Reg**	**TC – High Reg**
2^4	0.693	0.700	0.697	0.712
2^5	0.680	0.698	0.688	0.704
2^6	0.651	0.673	0.639	0.699
2^7	0.633	0.663	0.612	0.684
2^8	0.601	0.662	0.579	0.679
2^9	0.552	0.642	0.555	0.670
2^{10}	0.537	0.629	0.538	0.663
2^{11}	0.514	0.622	0.523	0.658

5.4.2 Microarray Data

These data sets were obtained from the `datamicroarray` package in R (available on GitHub at `https://github.com/ramhiser/datamicroarray`). The data sets are mostly cancer related, and have sample sizes ranging from 31 to 248 and dimension ranging from 456 to 54,613.

We compare the performance of the tournament classifier to lasso. As in Section 5.4.1, we use the `scikit-learn` implementation of lasso in Python where the penalty term λ is automatically selected within the `LassoCV` function. For the tournament classifier, we consider values of the regularization parameter r_0 ranging from 0.05 to 0.45 in increments of 0.05. For each data set, we randomly select a quarter of the observations as the training set and leave the rest as a test set. We then fit tournament classifier models (one for each level of r_0) and lasso to the training set, and calculate the test and training error for each model. Because the error rates vary depending on which quarter we randomly select, we repeat this procedure 20 times for each data set and average the resulting error rates.

We present the results in Figure 5.1. For each data set, we give the test set accuracy for lasso, along with the minimum, median, and maximum value from the various tournament classifier regularization parameters. In most cases, lasso and tournament classifier perform comparably.

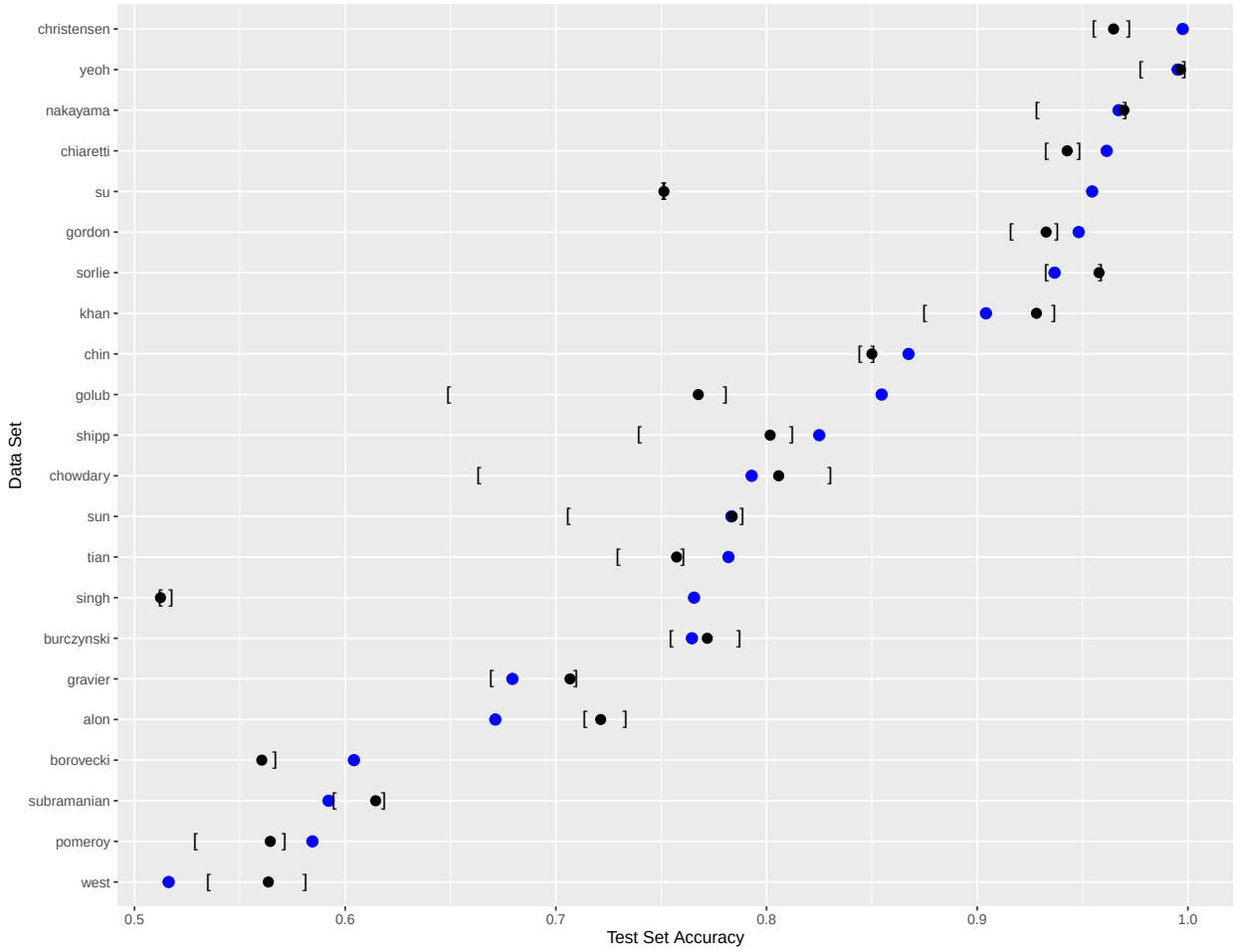

Figure 5.1: Comparison of lasso and tournament classifier for the microarray data sets. Blue dots represents the test accuracy for lasso. Black dots represent the median test accuracy for tournament classifier, and the brackets represent the minimum and maximum.

We also present more detailed results for a selection of the data sets. In each of the plots, the dotted line shows the test accuracy for lasso, and each point represents the test accuracy for the tournament classifier for a given value of r_0. As noted above, lasso and the tournament classifier perform similarly in most cases. On the left side of Figure 5.2, we see that tournament classifier performs better than lasso for higher r_0 (more regularization), and this performance declines as we use less regularization. In other cases, lower r_0 may result in worse performance. On the right side of Figure 5.2, performance is slightly worse for low values of r_0.

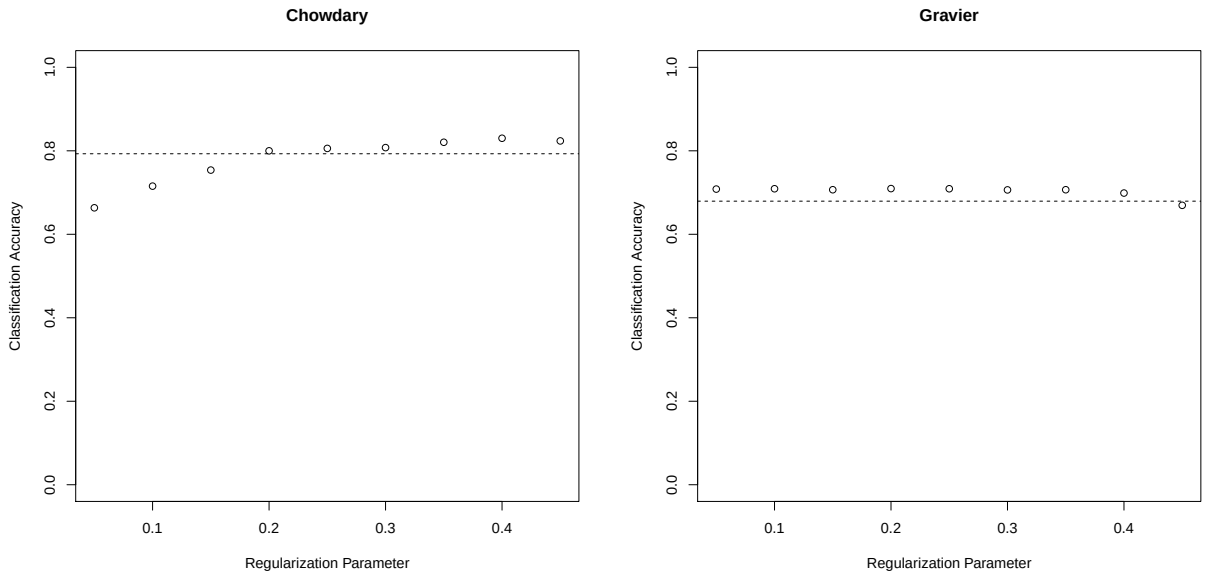

Figure 5.2: Comparison of lasso and tournament classifier for Chowdary (2006) and Gravier (2010). Dotted line shows the test set accuracy for lasso. Circles show the test set accuracy for tournament classifier.

There are also cases where the performance of the two methods differ considerably. We present two examples in Figure 5.3.

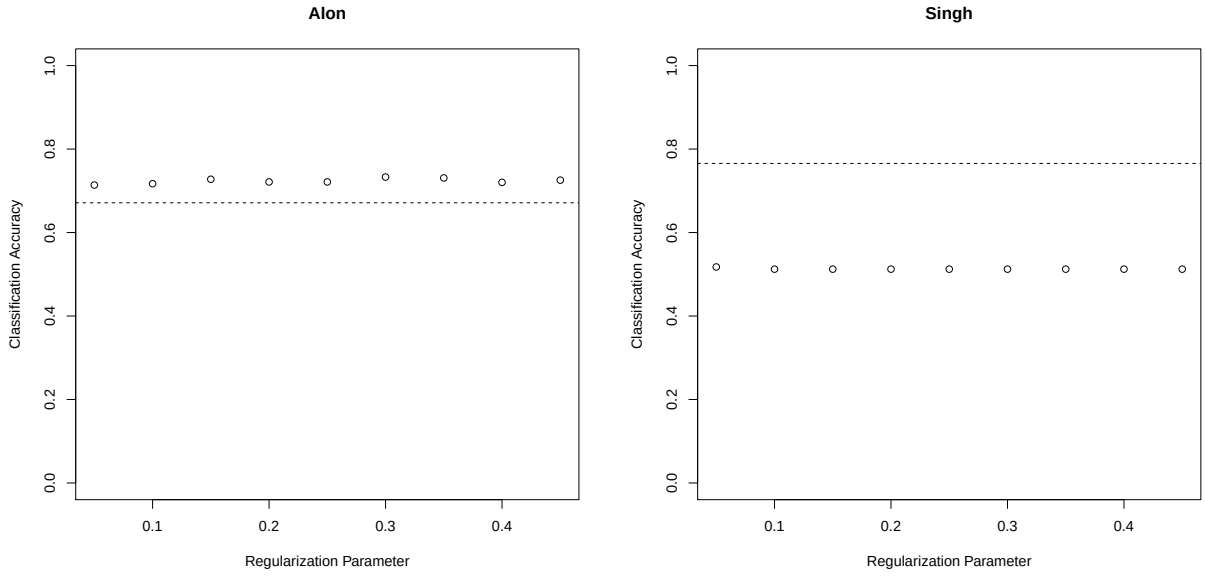

Figure 5.3: Comparison of lasso and tournament classifier for Alon (1999) and Singh (2002). Dotted line shows the test set accuracy for lasso. Circles show the test set accuracy for tournament classifier.

As noted above, the two methods generally perform comparably across the data sets. To assess why the performance diverges in some cases (*e.g.*, Singh (2002)), we perform a variety of analyses. This includes looking at the number of nonzero coefficients selected by lasso, scree plots, and examining the coefficients of the tournament classifier. However, we observe no discernible patterns to explain the differences in performance. We present the scree plots and tournament classifier coefficients in Appendix P.

5.5 Discussion

In this chapter, we introduce a novel framework for high dimensional classification. We present a specific approach that generally performed comparably to existing methods when applied to the microarray data sets in Section 5.4. However, the general method itself is modular, and we can choose to substitute various parts of the method to better suit the specific application. In this chapter, we present a specific approach that yields a sparse classifier by shrinking coefficients towards zero. On the other hand, as noted by Pan and Gagnon-Bartsch (2020), some data sets may contain dense signals resulting from the presence of latent biological factors. Our method does not explicitly attempt to recover dense latent signals, and modifying the comparison method could improve performance in situations where these latent signals are prominent. We propose a tournament that involves pairing the variables, and combining these pairs to yield a linear classifier. Some additional modifications would include creating non-linear classifiers and using a different group size within the tournament. Another natural modification would include extending the method for the case with multiple classes.

CHAPTER VI

Discussion

This dissertation has presented several methods for addressing the challenges of analyzing randomized experiments. For example, sample sizes may be limited due to practical constraints, which in turn could limit the precision of the treatment effect estimate. There may also be a large number of covariates to choose from. In cases where the number of covariates exceeds the sample size, traditional covariate adjustment methods can perform poorly or fail entirely. This issue is exacerbated in cases where the statistical analyses must be specified in advance. Choosing to adjust using the wrong covariates or making an overly aggressive adjustment could harm performance relative to the unadjusted estimator. Our methods address these issues, often by integrating modern machine learning techniques into the traditional analysis performed under the Neyman-Rubin model. In Chapters II and III, we present flexible methods for making covariate adjustments in Bernoulli and pair randomized experiments. These methods allow for automatic variable selection, eliminating the need for guesswork when choosing the covariates ahead of time. They are also design-based and generally outperform the unadjusted estimator, even if the covariates are not predictive of outcome.

An important feature of these methods are that they reconcile seemingly different approaches. For example, while these estimators are design-based, they allow for the use of models to improve performance, and we do not need to assume that these models are cor-

rectly specified. Model-assisted estimators have been used in survey sampling dating back to at least Cassel et al. (1976), and more recently for design-based covariate adjustments in randomized experiments. This includes the estimator of Aronow and Middleton (2013), of which the LOOP estimator is a special case. We also combine the strengths of randomized experiments and observational studies in Chapter IV. Randomized experiments are free from confounding bias and the randomization allows uncertainty to be easily quantified. However, as noted above, sample sizes (and therefore precision) may be limited. On the other hand, observational data sets are often large, potentially allowing for improved precision, but are not free from confounding variables. We present the ReLOOP method to take advantage of these complementary strengths. Like LOOP, ReLOOP is design-based and will generally outperform the unadjusted estimator. We also take advantage of an external data set to improve precision in the randomized experiment without allowing confounding bias to leak into our analysis.

There are also a number of logical extensions and future work for the methods presented in this dissertation. In Chapter III, we build on the LOOP estimator, creating a comparable method for paired experiments that deals with an issue unique to paired experiments, the pair inclusion trade-off. A similar issue can occur in blocked experiments, where it can be unclear the extent to which we should account for the block structure while making adjustments. While we do discuss a modification to the LOOP procedure for blocked experiments in Chapter II, this modification does not address this issue. Future work for a block randomized trial would include a covariate adjustment method that deals with this issue while also allowing for blocks of varying size. Other logical extensions include methods for other study designs (*e.g.*, cluster randomization) and for studies with multiple treatments.

While ReLOOP works within the LOOP framework to make covariate adjustments using auxiliary data, further work may also be required to extend ReLOOP to other study designs. For example, suppose we wish to incorporate external predictions in block or pair randomized experiments. One potential challenge in such work would be using a remnant which does

not share the blocked structure of the randomized experiment. There are also other areas for further exploration for ReLOOP. In Chapter IV, we generally assume that the external predictions Z^r have already been constructed rather than discussing methods for fitting the external prediction model f_{ext}. For example, we can exploit similarities and differences between the target RCT and the remnant in order to improve the performance of the external prediction model. Sales et al. (2018a) suggest an approach in rebar where the remnant is split into a group of participants that is similar to the RCT participants and a group that is not. We can then use these two groups to gauge the sensitivity of the external prediction model to extrapolation. Other exploration include tailoring f_{ext} for the specific application, using the observational data set to improve subgroup analyses, and dealing with external data sets where one or both of the covariates and outcome are not shared with the covariates and outcome within the randomized experiment. For the last issue, borrowing from the transfer learning literature may be helpful.

APPENDICES

APPENDIX A

Equivalence between LOOP and the Simple Difference Estimator

Consider the case where we impute the potential outcomes without making use of covariates. We estimate t_i as the mean of the observed outcomes in the treatment group and c_i as the mean of the observed outcomes in the control group (excluding observation i in both cases):

$$\hat{t}_i = \frac{\sum_{k \in \mathcal{T} \setminus i} Y_k}{n - T_i} \tag{A.1}$$

$$\hat{c}_i = \frac{\sum_{k \in \mathcal{C} \setminus i} Y_k}{(N - n) - (1 - T_i)}. \tag{A.2}$$

If the assignment probabilities are all equal, *i.e.*, if $p_i = p$ for all i and for some fixed p, then

the LOOP estimator is exactly equivalent to the simple difference estimator:

$$
\begin{aligned}
\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{m}_i\right)U_i \\
&= \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{1}{p}\left(Y_i - \hat{m}_i\right)T_i + \sum_{i=1}^{N}\frac{1}{1-p}\left(\hat{m}_i - Y_i\right)\left(1-T_i\right)\right\} \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\frac{1}{p}\left\{Y_i - \left(\frac{\sum_{k\in\mathcal{T}\setminus i}(1-p)Y_k}{n-T_i} + \frac{\sum_{k\in\mathcal{C}\setminus i}pY_k}{(N-n)-(1-T_i)}\right)\right\}T_i + \right. \\
&\qquad \left. \sum_{i=1}^{N}\frac{1}{1-p}\left\{\left(\frac{\sum_{k\in\mathcal{T}\setminus i}(1-p)Y_k}{n-T_i} + \frac{\sum_{k\in\mathcal{C}\setminus i}pY_k}{(N-n)-(1-T_i)}\right) - Y_i\right\}\left(1-T_i\right)\right] \\
&= \frac{1}{N}\left\{\sum_{i\in\mathcal{T}}\left(\frac{Y_i}{p} - \frac{1-p}{p}\frac{\sum_{k\in\mathcal{T}\setminus i}Y_k}{n-1} - \frac{\sum_{k\in\mathcal{C}}Y_k}{N-n}\right) + \right. \\
&\qquad \left. \sum_{i\in\mathcal{C}}\left(\frac{\sum_{k\in\mathcal{T}}Y_k}{n} + \frac{p}{1-p}\frac{\sum_{k\in\mathcal{C}\setminus i}Y_k}{(N-n)-1} - \frac{Y_i}{1-p}\right)\right\} \\
&= \frac{1}{N}\left\{\sum_{i\in\mathcal{T}}\frac{Y_i}{p} - \sum_{i\in\mathcal{C}}\frac{Y_i}{1-p} - \frac{1-p}{p}\frac{(n-1)\sum_{k\in\mathcal{T}}Y_k}{n-1} - \frac{n\sum_{k\in\mathcal{C}}Y_k}{N-n} + \right. \\
&\qquad \left. \frac{(N-n)\sum_{k\in\mathcal{T}}Y_k}{n} + \frac{p}{1-p}\frac{((N-n)-1)\sum_{k\in\mathcal{C}}Y_k}{(N-n)-1}\right\} \\
&= \frac{1}{N}\left\{\sum_{i\in\mathcal{T}}\frac{Y_i-(1-p)Y_i}{p} - \sum_{i\in\mathcal{C}}\frac{Y_i-pY_i}{1-p} - \frac{n\sum_{k\in\mathcal{C}}Y_k}{N-n} + \frac{(N-n)\sum_{k\in\mathcal{T}}Y_k}{n}\right\} \\
&= \frac{1}{N}\left\{\sum_{i\in\mathcal{T}}Y_i - \sum_{i\in\mathcal{C}}Y_i - \frac{n\sum_{k\in\mathcal{C}}Y_k}{N-n} + \frac{(N-n)\sum_{k\in\mathcal{T}}Y_k}{n}\right\} \\
&= \frac{1}{N}\left\{\frac{((N-n)+n)\sum_{k\in\mathcal{T}}Y_k}{n} - \frac{(n+(N-n))\sum_{k\in\mathcal{C}}Y_k}{N-n}\right\} \\
&= \frac{\sum_{k\in\mathcal{T}}Y_k}{n} - \frac{\sum_{k\in\mathcal{C}}Y_k}{N-n} \\
&= \hat{\tau}_{sd}.
\end{aligned}
$$

Technical note: One minor difference between the simple difference estimator and the LOOP estimator in this case is that the simple difference estimator is undefined whenever n is equal to 0 or N, whereas the LOOP estimator is undefined whenever n is equal to 0, 1, $N-1$, or N.

APPENDIX B

Variance of the LOOP Estimator

B.1 Variance and Covariance of $\hat{\tau}_i$

In this section, we calculate $\text{Var}(\hat{\tau}_i)$ and $\text{Cov}(\hat{\tau}_i,\hat{\tau}_j)$. We begin with the variance of a single $\hat{\tau}_i$:

$$\begin{aligned}
\text{Var}(\hat{\tau}_i) &= \text{Var}\left[\mathbb{E}(\hat{\tau}_i|\hat{m}_i)\right] + \mathbb{E}[\text{Var}(\hat{\tau}_i|\hat{m}_i)] \\
&= \text{Var}(\tau_i) + \mathbb{E}\left[\text{Var}\left(\frac{1}{p_i}(Y_i - \hat{m}_i)T_i + \frac{1}{1-p_i}(\hat{m}_i - Y_i)(1-T_i)|\hat{m}_i\right)\right] \\
&= 0 + \mathbb{E}\left[\text{Var}\left(\frac{1}{p_i}(t_i - \hat{m}_i)T_i + \frac{1}{1-p_i}(\hat{m}_i - c_i)(1-T_i)|\hat{m}_i\right)\right] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}\left[\text{Var}((1-p_i)(t_i - \hat{m}_i)T_i + p_i(\hat{m}_i - c_i)(1-T_i)|\hat{m}_i)\right] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}[\text{Var}[((1-p_i)t_i + p_ic_i - \hat{m}_i)T_i + p_i(\hat{m}_i - c_i)|\hat{m}_i]] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}[\text{Var}[(m_i - \hat{m}_i)T_i + p_i(\hat{m}_i - c_i)|\hat{m}_i]] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}[(m_i - \hat{m}_i)^2\text{Var}(T_i|\hat{m}_i)] \\
&= \frac{1}{p_i(1-p_i)}\mathbb{E}[(m_i - \hat{m}_i)^2] \\
&= \frac{1}{p_i(1-p_i)}\text{MSE}(\hat{m}_i). \qquad \text{(B.1)}
\end{aligned}$$

We now analyze the covariance term, γ_{ij}.

$$\begin{aligned}\gamma_{ij} &= \mathrm{Cov}[(Y_i - \hat{m}_i)U_i, (Y_j - \hat{m}_j)U_j] \\ &= \mathrm{Cov}(Y_iU_i, Y_jU_j) - \mathrm{Cov}(Y_iU_i, \hat{m}_jU_j) \\ &\quad - \mathrm{Cov}(\hat{m}_iU_i, Y_jU_j) + \mathrm{Cov}(\hat{m}_iU_i, \hat{m}_jU_j). \end{aligned} \tag{B.2}$$

The first term is zero, as Y_iU_i and Y_jU_j are independent. The second and third terms are also zero; for example, in the case of the second term,

$$\begin{aligned}\mathrm{Cov}(Y_iU_i, \hat{m}_jU_j) &= \mathbb{E}(Y_iU_i\hat{m}_jU_j) - \mathbb{E}(Y_iU_i)\mathbb{E}(\hat{m}_jU_j) \\ &= \mathbb{E}(Y_iU_i\hat{m}_j)\mathbb{E}(U_j) - \mathbb{E}(Y_iU_i)\mathbb{E}(\hat{m}_j)\mathbb{E}(U_j) \\ &= 0\end{aligned}$$

and a similar argument applies to the third term. Thus,

$$\begin{aligned}\gamma_{ij} &= \mathrm{Cov}(\hat{m}_iU_i, \hat{m}_jU_j) \\ &= \rho_{ij}\sqrt{\mathrm{Var}(\hat{m}_iU_i)\mathrm{Var}(\hat{m}_jU_j)} \\ &= \rho_{ij}\sqrt{\frac{\mathrm{Var}(\hat{m}_i)\mathrm{Var}(\hat{m}_j)}{p_ip_j(1-p_i)(1-p_j)}}\end{aligned} \tag{B.3}$$

where

$$\rho_{ij} = \mathrm{Corr}(\hat{m}_iU_i, \hat{m}_jU_j).$$

B.2 Bound for the First Term in the Variance of $\hat{\tau}$

In this section, we provide a bound for the first term in (2.12). Once again, we assume that $p_i = p$ for all i, and we wish to bound the following quantity:

$$\frac{\overline{\text{MSE}}}{Np(1-p)} = \frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p(1-p)}\text{MSE}(\hat{m}_i)\right]. \tag{B.4}$$

We can bound the MSE of $\hat{m}_i$ in terms of the MSEs of $\hat{t}_i$ and $\hat{c}_i$:

$$\begin{aligned}
\text{MSE}(\hat{m}_i) &= [\mathbb{E}(\hat{m}_i - m_i)]^2 + \text{Var}(\hat{m}_i) \\
&= \left[\mathbb{E}[(1-p)\hat{t}_i + p\hat{c}_i - (1-p)t_i - pc_i]\right]^2 + \text{Var}[(1-p)\hat{t}_i + p\hat{c}_i] \\
&= [\mathbb{E}[(1-p)(\hat{t}_i - t_i)] + \mathbb{E}[p(\hat{c}_i - c_i)]]^2 + (1-p)^2\text{Var}(\hat{t}_i) + p^2\text{Var}(\hat{c}_i) \\
&\quad + 2p(1-p)\text{Cov}(\hat{t}_i, \hat{c}_i) \\
&= [(1-p)\text{Bias}(\hat{t}_i) + p\text{Bias}(\hat{c}_i)]^2 + (1-p)^2\text{Var}(\hat{t}_i) + p^2\text{Var}(\hat{c}_i) \\
&\quad + 2p(1-p)\text{Cov}(\hat{t}_i, \hat{c}_i) \\
&= (1-p)^2\text{Bias}^2(\hat{t}_i) + p^2\text{Bias}^2(\hat{c}_i) + 2p(1-p)\text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i) \\
&\quad + (1-p)^2\text{Var}(\hat{t}_i) + p^2\text{Var}(\hat{c}_i) + 2p(1-p)\text{Cov}(\hat{t}_i, \hat{c}_i)] \\
&= (1-p)^2\text{MSE}(\hat{t}_i) + p^2\text{MSE}(\hat{c}_i) + 2p(1-p)\left[\text{Cov}(\hat{t}_i, \hat{c}_i) + \text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i)\right] \\
&\leq (1-p)^2\text{MSE}(\hat{t}_i) + p^2\text{MSE}(\hat{c}_i) + 2p(1-p)\sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)}.
\end{aligned} \tag{B.5}$$

To show inequality (B.5), we prove that:

$$\text{Cov}(\hat{t}_i, \hat{c}_i) + \text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i) \leq \sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)}.$$

The proof is trivial, but is included here for the sake of completeness.

Let $\mathrm{Cov}(\hat{t}_i, \hat{c}_i) = C$, $\mathrm{Bias}(\hat{t}_i) = B_t$, $\mathrm{Bias}(\hat{c}_i) = B_c$, $\mathrm{Var}(\hat{t}_i) = V_t$, $\mathrm{Var}(\hat{c}_i) = V_c$:

$$
\begin{aligned}
C + B_t B_c &\leq \sqrt{\mathrm{MSE}(\hat{t}_i)\mathrm{MSE}(\hat{c}_i)} \\
(C + B_t B_c)^2 &\leq (B_t^2 + V_t)(B_c^2 + V_c) \\
C^2 + 2CB_t B_c + B_t^2 B_c^2 &\leq V_t V_c + V_t B_c^2 + V_c B_t^2 + B_t^2 B_c^2 \\
C^2 + 2CB_t B_c &\leq V_t V_c + V_t B_c^2 + V_c B_t^2.
\end{aligned}
$$

$\mathrm{Cov}(\hat{t}_i, \hat{c}_i)$ is less than or equal to $\sqrt{\mathrm{Var}(\hat{t}_i)\mathrm{Var}(\hat{c}_i)}$ so it is sufficient to show:

$$
\begin{aligned}
V_t V_c + 2\sqrt{V_t V_c} B_t B_c &\leq V_t V_c + V_t B_c^2 + V_c B_t^2 \\
2\sqrt{V_t V_c} B_t B_c &\leq V_t B_c^2 + V_c B_t^2 \\
0 &\leq V_t B_c^2 - 2\sqrt{V_t V_c} B_t B_c + V_c B_t^2 \\
0 &\leq (\sqrt{V_t} B_c - \sqrt{V_c} B_t)^2.
\end{aligned}
$$

Finally, we plug (B.5) into (B.4) to obtain the following bound:

$$
\begin{aligned}
\frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p(1-p)}\mathrm{MSE}(\hat{m}_i)\right] &\leq \frac{1}{N}\left[\frac{1-p}{p}\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{t}_i) + \frac{p}{1-p}\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{c}_i) + \right. \\
&\qquad \left. 2\frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathrm{MSE}(\hat{t}_i)\mathrm{MSE}(\hat{c}_i)}\right] \\
&\leq \frac{1}{N}\left[\frac{1-p}{p}\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{t}_i) + \frac{p}{1-p}\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{c}_i) + \right. \\
&\qquad \left. 2\sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{t}_i)\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{c}_i)}\right] \\
&= \frac{1}{N}\left[\frac{1-p}{p}M_t + \frac{p}{1-p}M_c + 2\sqrt{M_t M_c}\right].
\end{aligned}
$$

B.3 $\hat{M}_t$ and $\hat{M}_c$ are Unbiased

Recall that

$$M_t = \frac{1}{N}\sum_{i=1}^{N} \text{MSE}(\hat{t}_i)$$
$$M_c = \frac{1}{N}\sum_{i=1}^{N} \text{MSE}(\hat{c}_i)$$

and that

$$\hat{M}_t = \frac{1}{Np}\sum_{i\in\mathcal{T}} (\hat{t}_i - t_i)^2$$
$$\hat{M}_c = \frac{1}{N(1-p)}\sum_{i\in\mathcal{C}} (\hat{c}_i - c_i)^2.$$

We will show that $\hat{M}_t$ and $\hat{M}_c$ are unbiased.

$$\begin{aligned}
\mathbb{E}(\hat{M}_t) &= \mathbb{E}\left[\frac{1}{Np}\sum_{i\in\mathcal{T}} (\hat{t}_i - t_i)^2\right] \\
&= \mathbb{E}\left[\frac{1}{Np}\sum_{i=1}^{N} T_i(\hat{t}_i - t_i)^2\right] \\
&= \frac{1}{Np}\sum_{i=1}^{N} \mathbb{E}\left[T_i(\hat{t}_i - t_i)^2\right] \\
&= \frac{1}{Np}\sum_{i=1}^{N} \mathbb{E}\left(T_i\right)\mathbb{E}\left[(\hat{t}_i - t_i)^2\right] \\
&= \frac{1}{Np}\sum_{i=1}^{N} p\mathbb{E}\left[(\hat{t}_i - t_i)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \text{MSE}\left(\hat{t}_i\right)
\end{aligned}$$

The argument for $\hat{M}_c$ is analogous.

B.4 Estimating γ_{ij}

In this section, we provide an estimate for γ_{ij}. First,

$$\begin{aligned}\gamma_{ij} &= \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j) = \mathrm{Cov}\left[[(1-p_i)\hat{t}_i + p_i\hat{c}_i]U_i, [(1-p_j)\hat{t}_j + p_j\hat{c}_j]U_j\right] \\ &= (1-p_i)(1-p_j)\mathrm{Cov}(\hat{t}_i U_i, \hat{t}_j U_j) + (1-p_i)p_j\mathrm{Cov}(\hat{t}_i U_i, \hat{c}_j U_j) \\ &\quad + p_i(1-p_j)\mathrm{Cov}(\hat{c}_i U_i, \hat{t}_j U_j) + p_i p_j \mathrm{Cov}(\hat{c}_i U_i, \hat{c}_j U_j). \end{aligned} \tag{B.6}$$

Now, we let $\hat{t}_i^{+j}$ denote the estimate of t_i including the j-th observation, where all the treatment assignments of the other $N-2$ observations are kept as is. Similarly, we let $\hat{t}_i^{-j}$ denote the estimate of t_i excluding the j-th observation. Then we have

$$\begin{aligned}\mathrm{Cov}(\hat{t}_i U_i, \hat{t}_j U_j | U_{k\notin\{i,j\}}) &= \hat{t}_i^{+j}\hat{t}_j^{+i} - \hat{t}_i^{-j}\hat{t}_j^{+i} - \hat{t}_i^{+j}\hat{t}_j^{-i} + \hat{t}_i^{-j}\hat{t}_j^{-i} \\ &= (\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}) \\ \mathrm{Cov}(\hat{t}_i U_i, \hat{c}_j U_j | U_{k\notin\{i,j\}}) &= \hat{t}_i^{+j}\hat{c}_j^{-i} - \hat{t}_i^{-j}\hat{c}_j^{-i} - \hat{t}_i^{+j}\hat{c}_j^{+i} + \hat{t}_i^{-j}\hat{c}_j^{+i} \\ &= \left(\hat{t}_i^{+j} - \hat{t}_i^{-j}\right)\left(\hat{c}_j^{-i} - \hat{c}_j^{+i}\right) \\ \mathrm{Cov}(\hat{c}_i U_i, \hat{t}_j U_j | U_{k\notin\{i,j\}}) &= \hat{c}_i^{-j}\hat{t}_j^{+i} - \hat{c}_i^{+j}\hat{t}_j^{+i} - \hat{c}_i^{-j}\hat{t}_j^{-i} + \hat{c}_i^{+j}\hat{t}_j^{-i} \\ &= \left(\hat{c}_i^{-j} - \hat{c}_i^{+j}\right)\left(\hat{t}_j^{+i} - \hat{t}_j^{-i}\right) \\ \mathrm{Cov}(\hat{c}_i U_i, \hat{c}_j U_j | U_{k\notin\{i,j\}}) &= \hat{c}_i^{-j}\hat{c}_j^{-i} - \hat{c}_i^{+j}\hat{c}_j^{-i} - \hat{c}_i^{-j}\hat{c}_j^{+i} + \hat{c}_i^{+j}\hat{c}_j^{+i} \\ &= \left(\hat{c}_i^{-j} - \hat{c}_i^{+j}\right)\left(\hat{c}_j^{-i} - \hat{c}_j^{+i}\right). \end{aligned} \tag{B.7}$$

Note that $\hat{t}_i^{+j}$ is calculable when $T_j = 1$, but not when $T_j = 0$, as t_j is not observable when $T_j = 0$. Similarly, $\hat{c}_i^{+j}$ is calculable when $T_j = 0$, but not when $T_j = 1$. Thus, we use

following estimate of the covariance (where all the terms are estimable):

$$\hat{\gamma}_{ij} = \begin{cases} \frac{(1-p_i)(1-p_j)}{p_i p_j}(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}), & T_i = T_j = 1 \\ (\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}), & T_i = 0, T_j = 1 \\ (\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}), & T_i = 1, T_j = 0 \\ \frac{p_i p_j}{(1-p_i)(1-p_j)}(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}), & T_i = T_j = 0 \end{cases} \tag{B.8}$$

which is an unbiased estimate of the covariance:

$$\begin{aligned}
&\mathbb{E}[\hat{\gamma}_{ij}|U_{k\notin\{i,j\}}] \\
&= p_i p_j \frac{(1-p_i)(1-p_j)}{p_i p_j}(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}) + (1-p_i)p_j(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}) \\
&\quad + p_i(1-p_j)(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}) + (1-p_i)(1-p_j)\frac{p_i p_j}{(1-p_i)(1-p_j)}(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}) \\
&= (1-p_i)(1-p_j)\mathrm{Cov}(\hat{t}_i U_i, \hat{t}_j U_j|U_{k\notin\{i,j\}}) + (1-p_i)p_j\mathrm{Cov}(\hat{t}_i U_i, \hat{c}_j U_j|U_{k\notin\{i,j\}}) \\
&\quad + p_i(1-p_j)\mathrm{Cov}(\hat{c}_i U_i, \hat{t}_j U_j|U_{k\notin\{i,j\}}) + p_i p_j\mathrm{Cov}(\hat{c}_i U_i, \hat{c}_j U_j|U_{k\notin\{i,j\}}) \\
&= \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j|U_{k\notin\{i,j\}}).
\end{aligned}$$

We take the expectation across all randomizations to show $\widehat{\mathrm{Cov}}(\hat{m}_i U_i, \hat{m}_j U_j)$ is unbiased.

$$\begin{aligned}
\mathbb{E}[\mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j|U_{k\notin\{i,j\}})] &= \mathbb{E}[\mathbb{E}(\hat{m}_i U_i \hat{m}_j U_j|U_{k\notin\{i,j\}}) - \mathbb{E}(\hat{m}_i U_i|U_{k\notin\{i,j\}})\mathbb{E}(\hat{m}_j U_j|U_{k\notin\{i,j\}})] \\
&= \mathbb{E}[\mathbb{E}(\hat{m}_i U_i \hat{m}_j U_j|U_{k\notin\{i,j\}})] \\
&= \mathbb{E}(\hat{m}_i U_i \hat{m}_j U_j)] \\
&= \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j)
\end{aligned}$$

Averaging across all i, j pairs yields an unbiased estimate of $\bar{\gamma}$.

APPENDIX C

Negligibility of $\bar{\gamma}$

In this section, we consider the behavior of $\bar{\gamma}$ as the sample size N grows large. In our model, the potential outcomes and the covariates are fixed parameters; they are not drawn from some probability distribution. Thus, when we speak of a growing sample size, we must imagine a growing set of parameters. Without any regularity conditions on these parameters, very little can be said, and thus some regularity conditions are necessary. However, we will not propose any specific set of regularity conditions *per se*, but rather we will assume that under some unspecified conditions that are appropriate for the imputation method under consideration (*e.g.*, OLS, random forests, etc.), the following two assumptions hold:

Assumption 5. *There exists a constant q such that, for every i, there exists a constant α_i where for all N and for all $i \leq N$*

$$\mathrm{Var}(\hat{m}_i) \leq \frac{\alpha_i}{N^q} \tag{C.1}$$

(note that when $i > N$ observation i is not yet in the model).

Assumption 6. *There exist constants C and r such that for all N*

$$\max_{i \leq N} \{\alpha_i\} \leq CN^r. \tag{C.2}$$

For example, if we impute the potential outcomes using OLS, then under suitable regularity conditions, assumption (C.1) might hold with $q = 1$ (see Freedman (2008)). Moreover, as long as the variation among the α_i is not too extreme, then assumption (C.2) might hold for some reasonably small value of r. For example, if the α_i follow a power law of the form

$$\text{fraction}\{\alpha_i \geq x\} < Kx^{-\lambda}$$

that holds for all N, then assumption (C.2) would be satisfied with $C = K^{1/\lambda}$ and $r = 1/\lambda$. Alternatively, if the tail of the distribution of the α_i decays exponentially, then any $r > 0$ would suffice, and if the α_i are bounded (which might be the case if both the response variable and the covariates are themselves bounded), then $r = 0$. In addition to the two assumptions above, we also assume in this section that $p_i = p$ for all i.

Now, combining assumptions 1 and 2 results in

$$\text{Var}(\hat{m}_i) \leq \frac{C}{N^{q-r}}$$

for all i, N such that $i \leq N$. Together with (2.10) from the main text, this implies that

$$\gamma_{ij} \leq \frac{C\rho_{ij}}{p(1-p)N^{q-r}}$$

which further implies that

$$\bar{\gamma} \leq \frac{C\bar{\rho}}{p(1-p)N^{q-r}}. \tag{C.3}$$

Thus, we find that $\bar{\gamma}$ will go to 0 at a rate faster than $1/N$, allowing us to ignore the $(N-1)\bar{\gamma}$ term in (2.12), just as long as $\bar{\rho}$ goes to 0 at a rate faster than $1/N^{1-q+r}$. In particular, if $q = 1$, then all that is necessary is that $\bar{\rho}$ goes to zero faster than $1/N^r$; if in addition $r = 0$, then all that is required is that $\bar{\rho}$ goes to zero.

In Appendix C.1 we show that if $\hat{m}_i$ is a polynomial function of degree D (or smaller) for all i, then $\bar{\rho} \leq D/(N-1)$. Combining this fact with the arguments given above in this

section, we see that for polynomial $\hat{m}_i$, $\bar{\gamma}$ will go to zero at a rate faster than $1/N$ simply as long as

$$\max_{i \leq N} \text{Var}(\hat{m}_i) \to 0.$$

C.1 Average correlation of $\hat{m}_i U_i$ and $\hat{m}_j U_j$ for polynomial $\hat{m}_i$

First, we define

$$\tilde{m}_i = \hat{m}_i - \mathbb{E}(\hat{m}_i)$$

and note that

$$\text{Corr}(\hat{m}_i, \hat{m}_j) = \text{Corr}(\tilde{m}_i, \tilde{m}_j) = \rho_{ij}.$$

Now suppose that $\tilde{m}_i$ is a polynomial function of degree D (or smaller) for all i. That is, for all i,

$$\tilde{m}_i = \sum_{d=1}^{D} \sum_{k_1, k_2, ..., k_d} a_{i.k_1 k_2 ... k_d} U_{k_1} U_{k_2} ... U_{k_d}$$

where the second sum is over all subsets $\{k_1, k_2, ..., k_d\} \subset \{1, 2, ..., N\} \setminus \{i\}$. A few comments: (a) no constant (intercept) term is needed in the expansion because $\tilde{m}_i$ has expectation 0, as do all the $U_{k_1} U_{k_2} ... U_{k_d}$ terms; (b) no higher powers of the U_k variables are needed (*e.g.*, U_k^3), since

$$U_k^2 = \frac{1}{p(1-p)} + \frac{1-2p}{p(1-p)} U_k$$

and thus by induction, any higher power of U_k can be reparameterized in terms of U_k itself; and (c) in our notation for the coefficients $a_{i.k_1 k_2 ... k_d}$, the ordering of the indices after the period does not matter. In other words, there is no distinction between $a_{2.358}$, $a_{2.583}$, $a_{2.835}$, etc. This fact will become important below when we count the number of times a specific coefficient appears in a sum.

Note that

$$\mathrm{Var}(\tilde{m}_i) = \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_d} a^2_{i.k_1k_2...k_d} \left[\frac{1}{p(1-p)}\right]^d$$

and define

$$b_{i.k_1k_2...k_d} = \frac{a_{i.k_1k_2...k_d}}{\sqrt{p^d(1-p)^d \mathrm{Var}(\tilde{m}_i)}}.$$

so that

$$\sum_{d=1}^{D} \sum_{k_1,k_2,...,k_d} b^2_{i.k_1k_2...k_d} = 1$$

and

$$\sum_{i=1}^{N} \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_d} b^2_{i.k_1k_2...k_d} = N \tag{C.4}$$

which is a fact we will make use of below.

Next observe that

$$\begin{aligned}
\gamma_{ij} &= \mathrm{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j) \\
&= \mathbb{E}\left[\left(\sum_{d=1}^{D} \sum_{k_1,k_2,...,k_d} a_{i.k_1k_2...k_d} U_{k_1} U_{k_2} ... U_{k_d} U_i\right)\right. \\
&\qquad \left. \times \left(\sum_{d=1}^{D} \sum_{k_1,k_2,...,k_d} a_{j.k_1k_2...k_d} U_{k_1} U_{k_2} ... U_{k_d} U_j\right)\right] \\
&= \mathbb{E}\left[\sum_{d=1}^{D}\sum_{e=1}^{D} \sum_{k_1,k_2,...,k_d} \sum_{l_1,l_2,...,l_e} a_{i.k_1k_2...k_d} U_{k_1} U_{k_2} ... U_{k_d} U_i a_{j.l_1l_2...l_e} U_{l_1} U_{l_2} ... U_{l_e} U_j\right] \\
&= \sum_{d=1}^{D}\sum_{e=1}^{D} \sum_{k_1,k_2,...,k_d} \sum_{l_1,l_2,...,l_e} a_{i.k_1k_2...k_d} a_{j.l_1l_2...l_e} \mathbb{E}\left(U_{k_1} U_{k_2} ... U_{k_d} U_i U_{l_1} U_{l_2} ... U_{l_e} U_j\right)
\end{aligned} \tag{C.5}$$

where again $\{k_1, k_2, ..., k_d\} \subset \{1, 2, ..., N\} \setminus \{i\}$ and $\{l_1, l_2, ..., l_e\} \subset \{1, 2, ..., N\} \setminus \{j\}$. But

$$\mathbb{E}\left(U_{k_1} U_{k_2} ... U_{k_d} U_i U_{l_1} U_{l_2} ... U_{l_e} U_j\right) = \begin{cases} \frac{1}{p^{d+1}(1-p)^{d+1}} & \{k_1, k_2, ..., k_d, i\} = \{l_1, l_2, ..., l_e, j\} \\ 0 & \text{otherwise} \end{cases}$$

and thus we may simplify (C.5) as

$$\gamma_{ij} = \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_{d-1}} a_{i.k_1k_2...k_{d-1}j} a_{j.k_1k_2...k_{d-1}i} \frac{1}{p^{d+1}(1-p)^{d+1}}$$

where now the second sum is over all subsets $\{k_1, k_2, ..., k_{d-1}\} \subset \{1, 2, ..., N\} \setminus \{i, j\}$.

Next observe that

$$\begin{aligned}
\rho_{ij} &= \frac{\gamma_{ij}}{\sqrt{\text{Var}(\tilde{m}_i)\text{Var}(\tilde{m}_j)}\,[p(1-p)]^{-1}} \\
&= \frac{\sum_{d=1}^{D} \sum_{k_1,k_2,...,k_{d-1}} a_{i.k_1k_2...k_{d-1}j} a_{j.k_1k_2...k_{d-1}i}\,[p(1-p)]^{-d}}{\sqrt{\text{Var}(\tilde{m}_i)\text{Var}(\tilde{m}_j)}} \\
&= \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_{d-1}} b_{i.k_1k_2...k_{d-1}j} b_{j.k_1k_2...k_{d-1}i}
\end{aligned}$$

and therefore

$$\sum_{i \neq j} \rho_{ij} = \sum_{i \neq j} \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_{d-1}} b_{i.k_1k_2...k_{d-1}j} b_{j.k_1k_2...k_{d-1}i}. \tag{C.6}$$

Consider now the following sum of squared coefficients

$$\sum_{i \neq j} \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_{d-1}} b^2_{i.k_1k_2...k_{d-1}j}$$

and observe that no single coefficient shows up in the sum any more than D times. For example, the coefficient $b_{1.234}$ will show up 3 times: once when $i = 1$, $j = 2$, $d = 3$, and $\{k_1, k_2\} = \{3, 4\}$; once when $i = 1$, $j = 3$, $d = 3$, and $\{k_1, k_2\} = \{2, 4\}$; and once when $i = 1$, $j = 4$, $d = 3$, and $\{k_1, k_2\} = \{2, 3\}$. Thus

$$\sum_{i \neq j} \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_{d-1}} b^2_{i.k_1k_2...k_{d-1}j} \leq D \sum_{i=1}^{N} \sum_{d=1}^{D} \sum_{k_1,k_2,...,k_d} b^2_{i.k_1k_2...k_d}$$

where the third sum on the left hand side is over all subsets $\{k_1, k_2, ..., k_{d-1}\} \subset \{1, 2, ..., N\} \setminus \{i, j\}$ and the third sum on the right hand side is over all subsets $\{k_1, k_2, ..., k_d\} \subset \{1, 2, ..., N\} \setminus \{i\}$. Applying (C.4), we therefore find that

$$\sum_{i \neq j} \sum_{d=1}^{D} \sum_{k_1, k_2, ..., k_{d-1}} b^2_{i.k_1 k_2 ... k_{d-1} j} \leq DN \tag{C.7}$$

and also similarly

$$\sum_{i \neq j} \sum_{d=1}^{D} \sum_{k_1, k_2, ..., k_{d-1}} b^2_{j.k_1 k_2 ... k_{d-1} i} \leq DN. \tag{C.8}$$

Given (C.7) and (C.8), we may now apply the Cauchy-Schwarz inequality to (C.6) and conclude

$$\sum_{i \neq j} \rho_{ij} \leq DN$$

or

$$\begin{aligned} \bar{\rho} &= \frac{1}{N(N-1)} \sum_{i \neq j} \rho_{ij} \\ &\leq \frac{D}{N-1} \end{aligned} \tag{C.9}$$

APPENDIX D

The Relationship between $\widetilde{\mathrm{Var}}(\hat{\tau})$ and the Sample Variance

We first show that

$$\tilde{M}_t = \frac{n}{n-1} s_t^2 = \frac{n}{n-1} \frac{1}{n-1} \sum_{i \in \mathcal{T}} (t_i - \bar{t})^2. \tag{D.1}$$

Without loss of generality, assume that $\mathcal{T} = \{1, ..., n\}$:

$$\begin{aligned} \tilde{M}_t &= \frac{1}{n} \sum_{i=1}^{n} (\hat{t}_i - t_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (\hat{t}_i^2 - 2\hat{t}_i t_i + t_i^2) \end{aligned} \tag{D.2}$$

We deal with the first two terms:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{t}_i^2 - 2\sum_{i=1}^{n} \hat{t}_i t_i &= \sum_{i=1}^{n} \frac{(\sum_{k\neq i} t_k)^2}{(n-1)^2} - 2\sum_{i=1}^{n} \frac{\sum_{k\neq i} t_k}{n-1} t_i \\
&= \sum_{i=1}^{n} \frac{\sum_{j,k\neq i} t_j t_k}{(n-1)^2} - 2\sum_{j\neq k} \frac{t_j t_k}{n-1} \\
&= \frac{1}{(n-1)^2}\left[(n-1)\sum_{i=1}^{n} t_i^2 + (n-2)\sum_{j\neq k} t_j t_k\right] \\
&\qquad - 2(n-1)\sum_{j\neq k} \frac{t_j t_k}{(n-1)^2} \\
&= \frac{\sum_{i=1}^{n} t_i^2}{n-1} + (n-2)\sum_{j\neq k} \frac{t_j t_k}{(n-1)^2} - 2(n-1)\sum_{j\neq k} \frac{t_j t_k}{(n-1)^2} \\
&= \frac{\sum_{i=1}^{n} t_i^2}{n-1} - n\sum_{j\neq k} \frac{t_j t_k}{(n-1)^2} \qquad \text{(D.3)}
\end{aligned}$$

Plugging (D.3) into (D.2), we can express $\tilde{M}_t$ as follows:

$$\begin{aligned}
\tilde{M}_t &= \frac{1}{n}\left[\sum_{i=1}^{n} \frac{t_i^2}{n-1} - n\sum_{j\neq k} \frac{t_j t_k}{(n-1)^2} + \frac{n-1}{n-1}\sum_{i=1}^{n} t_i^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} t_i^2 - \sum_{j\neq k} \frac{t_j t_k}{n-1}\right] \\
&= \frac{1}{n-1}\left[\frac{n-1}{n-1}\sum_{i=1}^{n} t_i^2 - \sum_{j\neq k} \frac{t_j t_k}{n-1}\right] \\
&= \frac{1}{n-1}\left[\frac{n}{n-1}\sum_{i=1}^{n} t_i^2 - \frac{\sum_{i=1}^{n} t_i^2 + \sum_{j\neq k} t_j t_k}{n-1}\frac{n}{n}\right] \\
&= \frac{1}{n-1}\frac{n}{n-1}\left[\sum_{i=1}^{n} t_i^2 - \frac{(\sum_{i=1}^{n} t_i)^2}{n}\right] \\
&= \frac{n}{n-1}\frac{1}{n-1}\left[\sum_{i=1}^{n} t_i^2 - n\bar{t}^2\right] \\
&= \frac{n}{n-1} s_t^2. \qquad \text{(D.4)}
\end{aligned}$$

An analogous calculation can be used to demonstrate that

$$\tilde{M}_c = \frac{N-n}{N-n-1} s_c^2. \tag{D.5}$$

Next, we plug (D.4) and (D.5) into (2.19):

$$\begin{aligned}
\widetilde{\mathrm{Var}}(\tilde{\tau}) &= \frac{1}{N}\left[\frac{1-p}{p}\tilde{M}_t + \frac{p}{1-p}\hat{M}_c + 2\sqrt{\tilde{M}_t \tilde{M}_c}\right] \\
&\leq \frac{1}{N}\left[\frac{1-p}{p}\tilde{M}_t + \frac{p}{1-p}\tilde{M}_c + \tilde{M}_t + \tilde{M}_c\right] \\
&= \frac{1}{N}\left[\frac{1}{p}\tilde{M}_t + \frac{1}{1-p}\tilde{M}_c\right] \\
&= \frac{1}{N}\left[\frac{n}{(n-1)p} s_t^2 + \frac{N-n}{(N-n-1)(1-p)} s_c^2\right] \\
&= \frac{n}{Np}\frac{s_t^2}{n-1} + \frac{N-n}{N(1-p)}\frac{s_c^2}{N-n-1} \\
&\approx \frac{s_t^2}{n-1} + \frac{s_c^2}{N-n-1}. \qquad \text{(D.6)}
\end{aligned}$$

where (D.6) follows from the fact that $\frac{n}{Np}$ and $\frac{N-n}{N(1-p)}$ are both approximately equal to 1.

APPENDIX E

The Random Drop Procedure

E.1 Illustrative Example

Consider an experiment with five participants, in which two participants are to be randomly assigned to treatment and the remaining three to control, and suppose we wish to estimate m_1 using the random drop procedure. If $T_1 = 1$, we randomly pick a control observation and omit it when calculating $\hat{m}_1$. Similarly, if $T_1 = 0$, we randomly drop a treatment observation.

On the left side of Table E.1, we show the 10 possible (and equally likely) treatment assignment vectors. The right side of the table shows the possible treatment assignment vectors after applying the random drop procedure; a backslash represents the dropped observation. For example, when the treatment assignment is 5) CTTCC, we could randomly drop either of the two treatment observations, resulting in either C\TCC or CT\CC.

We can use the above example to illustrate how $\hat{m}_1$ is independent of T_1. Regardless of whether T_1 is 0 or 1, we calculate $\hat{m}_1$ using a single treatment observation and two control observations; moreover, the value of T_1 does not tell us anything about which two of the four possible units will be in control, or which one of the four will be in treatment.

Table E.1: Illustration of the Random Drop Procedure

#	Treatment Assignments	Potential Drops		
1)	T T C C C	T T \ C C	T T C \ C	T T C C \
2)	T C T C C	T \ T C C	T C T \ C	T C T C \
3)	T C C T C	T \ C T C	T C \ T C	T C C T \
4)	T C C C T	T \ C C T	T C \ C T	T C C \ T
5)	C T T C C	C \ T C C	C T \ C C	
6)	C T C T C	C \ C T C	C T C \ C	
7)	C T C C T	C \ C C T	C T C C \	
8)	C C T T C	C C \ T C	C C T \ C	
9)	C C T C T	C C \ C T	C C T C \	
10)	C C C T T	C C C \ T	C C C T \	

For example, consider the arrangement T\CC for the last four observations. We can see that this arrangement occurs in exactly one in twelve of the combinations where $T_1 = 1$ and one in twelve of the combinations where $T_1 = 0$. That is,

$$\mathrm{P}(\mathrm{T\backslash CC}|T_1 = 1) = \mathrm{P}(\mathrm{T\backslash CC}|T_1 = 0) = 1/12.$$

The same is true of all of the other 11 possible arrangements of the last four observations. Thus T_1 and $\hat{m}_1$ are independent.

E.2 Expectation of the Random Drop Procedure

In this section, we show that $\hat{\tau}$ remains relatively unchanged by the random drop procedure in the case where we estimate m_i without using covariates. To do this, we show that the expectation (over random drops) of the estimate of the average treatment effect obtained from the random drop procedure is exactly equal to the estimate had we not used the random drop procedure at all.

Consider the case where we estimate m_i without using covariates. That is, we impute t_i as the average of the treated units and c_i as the average of the control units (omitting observation i each time). If unit i was in the control group, then each time we estimate m_i,

we would drop a random observation in the treatment group before taking the averages of the observed outcomes. While we could repeat this procedure many times and average the resulting estimates to get our final estimate of m_i, we could instead take the expected value of the "random drop" estimate over all possible drops. In this case, the value of $\hat{m}_i$ is exactly equal to the estimate had we not dropped any observations in the first place. Without loss of generality, we assume that observation i is assigned to control. Let $\hat{m}_{i,-k}$ and $\hat{\tau}_{i,-k}$ denote the estimates where we randomly dropped the k-th observation and let $\hat{m}_{i,\cdot}$ and $\hat{\tau}_{i,\cdot}$ denote their expected values over all possible drops.

$$
\begin{aligned}
\mathbb{E}_k(\hat{m}_{i,-k}) &= \frac{1}{n}\sum_{k\in\mathcal{T}} \hat{m}_{i,-k} \\
&= \frac{1}{n}\sum_{k\in\mathcal{T}} \left[\frac{\sum_{j\in\mathcal{T}\setminus\{i,k\}} Y_j}{n-1} + \frac{\sum_{j\in\mathcal{C}\setminus\{i,k\}} Y_j}{N-n-1}\right] \\
&= \frac{1}{n}\sum_{k\in\mathcal{T}} \left[\frac{\sum_{j\in\mathcal{T}\setminus\{k\}} Y_j}{n-1}\right] + \frac{1}{n}\sum_{k\in\mathcal{T}} \left[\frac{\sum_{j\in\mathcal{C}\setminus\{i\}} Y_j}{N-n-1}\right] \\
&= \frac{1}{n}\left[\frac{(n-1)\sum_{j\in\mathcal{T}} Y_j}{n-1}\right] + \frac{1}{n}\left[\frac{n\sum_{j\in\mathcal{C}\setminus\{i\}} Y_j}{N-n-1}\right] \\
&= \frac{\sum_{j\in\mathcal{T}} Y_j}{n} + \frac{\sum_{j\in\mathcal{C}\setminus\{i\}} Y_j}{N-n-1}.
\end{aligned}
$$

This last line is equal to the value of $\hat{m}_i$ that we would have gotten had we not dropped any observations besides i. Our estimate for τ_i would also be the same as if we had not used the random drop procedure (*i.e.*, $\mathbb{E}_k(\hat{\tau}_{i,-k}) = \hat{\tau}_i$). A similar argument can be used to show that if we were to use the random drop procedure when estimating $\hat{m}_i$ using a decision tree, the expected value of $\hat{\tau}$ would still be the post-stratified estimate.

APPENDIX F

Supplementary Results for Chapter II

F.1 Simulation 3 Results

In Table F.1, we provide the results for Simulation 3. The first column contains the sample size N, the second column contains the values of $\tilde{\gamma}$, and the third column contains the standard error of $\tilde{\gamma}$.

In Section 2.8.3, we observe that the simulation estimate $N\tilde{\gamma}$ for $N\bar{\gamma}$ begins to taper off at around $N = 70$ and note that this is due to the standard error of our estimate. While $N\bar{\gamma}$ may decline as N increases, the standard error of the simulation estimate $N\tilde{\gamma}$ does not. We can see this in Table F.1: the standard error estimate declines much more slowly than the value of $\tilde{\gamma}$.

Table F.1: Simulation 3 Results

N	$\bar{\gamma}$ **Estimate**	**Standard Error**
10	1.42×10^{-2}	2.79×10^{-4}
20	2.13×10^{-3}	5.05×10^{-5}
30	3.56×10^{-4}	1.79×10^{-5}
40	5.35×10^{-5}	5.62×10^{-6}
50	1.62×10^{-5}	3.15×10^{-6}
60	3.68×10^{-6}	1.62×10^{-6}
70	2.37×10^{-6}	1.34×10^{-6}
80	2.74×10^{-6}	1.46×10^{-6}
90	1.29×10^{-6}	9.81×10^{-7}
100	-7.68×10^{-7}	9.93×10^{-7}

F.2 Results from Barrera-Osorio et al. (2011)

In Table F.2, we provide the results from Barrera-Osorio et al. (2011) including the point estimates for each method. We also include two additional methods: OLS with interactions (as proposed by Lin (2013)) and LOOP in which OLS is used as the imputation method.

Table F.2: Effect of Treatment on Missing Status and Re-enrollment Status

Treatments	**Method**	**Missing Status**		**Re-enrollment Status**	
		Est. ($\times 10^{-3}$)	**SE** ($\times 10^{-3}$)	**Est.** ($\times 10^{-3}$)	**SE** ($\times 10^{-3}$)
Basic vs. Savings	LOOP with RF	-0.1	6.0	-25.0	11.8
	Simple Difference	6.7	7.4	-28.4	11.8
	OLS	3.8	6.3	-29.4	11.6
	OLS with Interactions	3.9	6.3	-29.5	11.6
	LOOP with OLS	3.8	6.3	-29.5	11.6
	Cross Estimation	-0.9	6.0	-25.4	11.6
Basic vs. Control	LOOP with RF	-2.2	5.8	15.9	11.6
	Simple Difference	4.1	7.1	16.6	11.6
	OLS	1.3	6.1	15.8	11.5
	OLS with Interactions	1.3	6.1	15.8	11.5
	LOOP with OLS	1.4	6.1	15.9	11.4
	Cross Estimation	-1.7	5.7	16.4	11.5
Saving vs. Control	LOOP with RF	-1.6	5.7	42.5	11.3
	Simple Difference	-2.5	7.0	45.0	11.4
	OLS	-2.3	6.1	46.3	11.2
	OLS with Interactions	-2.3	6.1	46.5	11.2
	LOOP with OLS	-2.3	6.1	46.4	11.2
	Cross Estimation	-1.6	5.7	41.7	11.2

APPENDIX G

Asymptotic Normality of the P-LOOP Estimator

In this section we prove that the P-LOOP estimator is asymptotically normally distributed under the assumptions outlined in Section 3.3.2. Recall the assumptions:

1. There exists some $0 < C < \infty$ and $q > 0$ such that for all i,

$$\mathrm{Var}(\tilde{d}_i) = \mathrm{Var}(\hat{d}_i) \leq C/N^q.$$

2. Let ρ_{ij} be the correlation of $\tilde{d}_i U_i$ and $\tilde{d}_j U_j$, and $\bar{\rho} = \frac{\sum_{i \neq j} \rho_{ij}}{N(N-1)}$. We assume that

$$N^{1-q}\bar{\rho} \longrightarrow 0.$$

3. Recall that $d_{0i}^{(N)} = \mathrm{E}(\hat{d}_i^{(N)})$ for some fixed N. For each pair i, we assume that the limit of $d_{0i}^{(N)}$ exists and denote the limit as $d_{\infty i}$. We also assume

$$\frac{1}{N}\sum_{i=1}^{N}\left(d_{0i}^{(N)} - d_{\infty i}\right)^2 \longrightarrow 0.$$

4. Let $V_N = \sum_{i=1}^{N}(d_i - d_{\infty i})^2$. There exists $0 < K < \infty$ such that

$$\frac{V_N}{N} \longrightarrow K,$$

and

$$\max_{i=1,\ldots,N} \frac{(d_i - d_{\infty i})^2}{V_N} \longrightarrow 0.$$

We now proceed with the proof. First, we write $N(\hat{\tau} - \tau)/\sqrt{V_N}$ as

$$\begin{aligned}
N(\hat{\tau} - \tau)/\sqrt{V_N} &= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \left\{ \left(W_i - \hat{d}_i \right) T_i + \left(W_i + \hat{d}_i \right) (1 - T_i) \right\} - \frac{N\tau}{\sqrt{V_N}} \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \left(W_i - \hat{d}_i U_i - 0.5(a_i + b_i) \right) \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} (W_i - d_{0i} U_i - 0.5(a_i + b_i)) + \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{d}_i U_i \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} (W_i - (d_{\infty i} - d_{\infty i} + d_{0i}) U_i - 0.5(a_i + b_i)) + \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{d}_i U_i \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} X_i + \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} (d_{\infty i} - d_{0i}) U_i + \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{d}_i U_i \qquad \text{(G.1)} \\
&= \text{(A)} + \text{(B)} + \text{(C)}
\end{aligned}$$

where we define $X_i = W_i - d_{\infty i} U_i - 0.5(a_i + b_i)$. We show that the first term converges in distribution to a standard normal random variable and the other two terms converge in probability to zero.

Asymptotic Normality of (A) We show that

$$\frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} X_i \xrightarrow{d} \mathrm{N}(0, 1)$$

by the Lindeberg-Feller central limit theorem (see for example, Chapter 9.8 of Resnick (2003)). We note that

$$\begin{aligned} W_i - 0.5(a_i + b_i) &= a_i T_i + b_i(1 - T_i) - 0.5(a_i + b_i) \\ &= 0.5(a_i - b_i)T_i + 0.5(b_i - a_i)(1 - T_i) \\ &= d_i U_i \end{aligned}$$

and therefore $X_i = (d_i - d_{\infty i})U_i$. The X_i are independent random variables, as the treatment assignments U_i are all independent, and d_i and $d_{\infty i}$ are constants. The X_i have expectation and variance

$$\begin{aligned} \mathrm{E}\left((d_i - d_{\infty i})U_i\right) &= (d_i - d_{\infty i})\mathrm{E}(U_i) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}\left((d_i - d_{\infty i})U_i\right) &= \mathrm{E}\left\{\left((d_i - d_{\infty i})U_i\right)^2\right\} \\ &= (d_i - d_{\infty i})^2 \mathrm{E}\left(U_i^2\right) \\ &= (d_i - d_{\infty i})^2. \end{aligned}$$

Thus $V_N = \sum_{i=1}^N \text{Var}(X_i)$. In addition, X_i satisfies the Lindeberg condition. For any $t > 0$, we have

$$\begin{aligned}
\frac{1}{V_N}\sum_{i=1}^N \text{E}\left(X_i^2 \mathbb{1}\{|X_i| > t\sqrt{V_N}\}\right) &= \frac{1}{V_N}\sum_{i=1}^N \text{E}\left((d_i - d_{\infty i})^2 U_i^2 \mathbb{1}\{|(d_i - d_{\infty i})U_i| > t\sqrt{V_N}\}\right) \\
&= \frac{1}{V_N}\sum_{i=1}^N (d_i - d_{\infty i})^2 \text{E}\left(\mathbb{1}\{|(d_i - d_{\infty i})U_i| > t\sqrt{V_N}\}\right) \\
&= \frac{1}{V_N}\sum_{i=1}^N (d_i - d_{\infty i})^2 \text{E}\left(\mathbb{1}\{(d_i - d_{\infty i})^2 U_i^2 > t^2 V_N\}\right) \\
&= \frac{1}{V_N}\sum_{i=1}^N (d_i - d_{\infty i})^2 \mathbb{1}\left\{\frac{(d_i - d_{\infty i})^2}{V_N} > t^2\right\} \longrightarrow 0
\end{aligned}$$

by assumption 4. Thus, the first term of equation (M.1) converges to a normal distribution by the Lindeberg-Feller central limit theorem.

(B) Converges in Probability to Zero Next we show that

$$\frac{1}{\sqrt{V_N}}\sum_{i=1}^N (d_{\infty i} - d_{0i})U_i$$

converges in L^2-norm to zero.

$$\begin{aligned}
\text{E}\left\{\left(\frac{1}{\sqrt{V_N}}\sum_{i=1}^N (d_{\infty i} - d_{0i})U_i\right)^2\right\} &= \frac{1}{V_N}\text{E}\left\{\sum_{i=1}^N (d_{\infty i} - d_{0i})^2 + \sum_{i\neq j}(d_{\infty i} - d_{0i})(d_{\infty j} - d_{0j})U_i U_j\right\} \\
&= \frac{1}{V_N}\sum_{i=1}^N (d_{\infty i} - d_{0i})^2 \\
&= \frac{N}{V_N}\frac{1}{N}\sum_{i=1}^N (d_{\infty i} - d_{0i})^2
\end{aligned}$$

where $\text{E}\{(d_{\infty i} - d_{0i})(d_{\infty j} - d_{0j})U_i U_j\} = 0$ since U_i and U_j have expectation zero and are independent. The quantity $\sum_{i=1}^N (d_{\infty i} - d_{0i})^2/N$ converges to zero by assumption 3, while

V_N/N (and therefore N/V_N) converges to a constant by assumption 4. Thus

$$\mathrm{E}\left\{\left(\frac{1}{\sqrt{V_N}}\sum_{i=1}^{N}(d_{\infty i}-d_{0i})U_i\right)^2\right\}\longrightarrow 0.$$

Since (B) converges in L^2-norm to zero, it also converges in probability to zero.

(C) Converges in Probability to Zero Finally we show that the last term in equation (M.1) converges in L^2-norm to zero:

$$\begin{aligned}
\mathrm{E}\left\{\left(\frac{1}{\sqrt{V_N}}\sum_{i=1}^{N}\tilde{d}_iU_i\right)^2\right\} &= \frac{1}{V_N}\sum_{i,j}\mathrm{E}\Big(\tilde{d}_iU_i\tilde{d}_jU_j\Big)\\
&= \frac{1}{V_N}\sum_{i,j}\mathrm{Cov}\Big(\tilde{d}_iU_i,\tilde{d}_jU_j\Big)\\
&= \frac{1}{V_N}\sum_{i,j}\rho_{ij}\sqrt{\mathrm{Var}\Big(\tilde{d}_iU_i\Big)\mathrm{Var}\Big(\tilde{d}_jU_j\Big)}\\
&= \frac{1}{V_N}\sum_{i,j}\rho_{ij}\sqrt{\mathrm{Var}\Big(\tilde{d}_i\Big)\mathrm{Var}\Big(\tilde{d}_j\Big)}\\
&\leq \frac{1}{V_N}\sum_{i,j}C\rho_{ij}/N^q\\
&= \frac{C}{V_NN^q}\sum_{i,j}\rho_{ij}\\
&= \frac{C}{V_NN^q}\left\{N+\sum_{i,j}\rho_{ij}\right\}\\
&= \frac{N}{V_N}\frac{C}{N^q}+\frac{N}{V_N}\frac{C}{N^q}\left\{(N-1)\bar{\rho}\right\}.
\end{aligned}$$

By assumption 4, N/V_N converges to a constant. Due to assumption 2, $\frac{C}{N^q}\left\{(N-1)\bar{\rho}\right\}$ converges to zero. It follows that the last term of equation (M.1) converges in L^2-norm (and therefore in probability) to zero.

Combining the results for (A), (B), and (C) implies that equation (M.1) converges in distribution to a standard normal random variable.

APPENDIX H

True Variance of the P-LOOP Estimator

First, we calculate the variance of a single $\hat{\tau}_i$:

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}_i) &= \mathrm{E}\left\{\mathrm{Var}(\hat{\tau}_i \mid \hat{d}_i)\right\} + \mathrm{Var}\left\{\mathrm{E}(\hat{\tau}_i \mid \hat{d}_i)\right\} \\
&= \mathrm{E}\left\{\mathrm{Var}(\hat{\tau}_i \mid \hat{d}_i)\right\} + \mathrm{Var}(\tau_i) \\
&= \mathrm{E}\left[\mathrm{Var}\left\{(W_i - \hat{d}_i)T_i + (W_i + \hat{d}_i)(1 - T_i) \mid \hat{d}_i\right\}\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left\{(a_i - \hat{d}_i)T_i + (b_i + \hat{d}_i)(1 - T_i) \mid \hat{d}_i\right\}\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left\{(a_i - b_i - 2\hat{d}_i)T_i + b_i + \hat{d}_i \mid \hat{d}_i\right\}\right] \\
&= \mathrm{E}\left\{(2d_i - 2\hat{d}_i)^2 \mathrm{Var}\left(T_i \mid \hat{d}_i\right)\right\} \\
&= \mathrm{E}\left\{4(d_i - \hat{d}_i)^2 \times 1/4\right\} \\
&= \mathrm{E}\left\{(d_i - \hat{d}_i)^2\right\} = \mathrm{MSE}(\hat{d}_i).
\end{aligned}$$

Let $\gamma_{ij} = \mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$. Then we have the following expression for the variance of the LOOP

estimator

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i\right) \\
&= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\mathrm{Var}(\hat{\tau}_i) + \sum_{i\neq j}\mathrm{cov}(\hat{\tau}_i,\hat{\tau}_j)\right\} \\
&= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\mathrm{MSE}(\hat{d}_i) + \sum_{i\neq j}\gamma_{ij}\right\}.
\end{aligned}$$

APPENDIX I

Negligibility of the Covariance Terms

In this section, we show that

$$\frac{\sum_{i \neq j} \gamma_{ij}}{\sum_{i=1}^{N} \text{MSE}(\hat{d}_i)} \longrightarrow 0$$

under the assumptions outlined in Section 3.3.2. We show that $\frac{1}{N} \sum_{i \neq j} \gamma_{ij}$ converges to zero, while $\frac{1}{N} \sum_{i=1}^{N} \text{MSE}(\hat{d}_i)$ converges to a constant. Recall that $U_i = 2T_i - 1$. We can therefore rewrite $\hat{\tau}_i$ as $W_i - \hat{d}_i U_i$. For any $i \neq j$, we have the following expression for γ_{ij}:

$$\begin{aligned}
\gamma_{ij} &= \text{Cov}(\hat{\tau}_i, \hat{\tau}_j) \\
&= \text{Cov}(W_i - \hat{d}_i U_i, W_j - \hat{d}_j U_j) \\
&= \text{Cov}(W_i, W_j) - \text{Cov}(W_i, \hat{d}_j U_j) - \text{Cov}(\hat{d}_i U_i, W_j) + \text{Cov}(\hat{d}_i U_i, \hat{d}_j U_j).
\end{aligned}$$

The first term is zero, as W_i and W_j are independent due to the independence of T_i and T_j. The second and third terms are also zero. Note that U_j is independent of W_i due to the independence of T_i and T_j, and recall that $\hat{d}_j$ is independent of T_j (and therefore U_j). Then

we have for the second term

$$\begin{aligned}\mathrm{Cov}(W_i, \hat{d}_j U_j) &= \mathrm{E}(W_i \hat{d}_j U_j) - \mathrm{E}(W_i)\mathrm{E}(\hat{d}_j U_j) \\ &= \mathrm{E}(W_i \hat{d}_j)\mathrm{E}(U_j) - \mathrm{E}(W_i)\mathrm{E}(\hat{d}_j)\mathrm{E}(U_j) \\ &= 0\end{aligned}$$

where the last line follows because $\mathrm{E}(U_j) = 0$. Next, note that $\mathrm{Cov}(d_{0i}U_i, d_{0j}U_j) = 0$ due to the independence of U_i and U_j, and that

$$\begin{aligned}\mathrm{Var}(\tilde{d}_i U_i) &= \mathrm{E}(\tilde{d}_i^2 U_i^2) \\ &= \mathrm{E}(\tilde{d}_i^2) \\ &= \mathrm{Var}(\tilde{d}_i).\end{aligned}$$

Then we have

$$\begin{aligned}\gamma_{ij} &= \mathrm{Cov}(\hat{d}_i U_i, \hat{d}_j U_j) \\ &= \mathrm{Cov}(\tilde{d}_i U_i + d_{0i} U_i, \tilde{d}_j U_j + d_{0j} U_j) \\ &= \mathrm{Cov}(\tilde{d}_i U_i, \tilde{d}_j U_j) \\ &= \mathrm{Corr}(\tilde{d}_i U_i, \tilde{d}_j U_j)\sqrt{\mathrm{Var}(\tilde{d}_i U_i)\mathrm{Var}(\tilde{d}_j U_j)} \\ &= \rho_{ij}\sqrt{\mathrm{Var}(\tilde{d}_i)\mathrm{Var}(\tilde{d}_j)}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{N}\sum_{i \neq j}\gamma_{ij} &= \frac{1}{N}\sum_{i \neq j}\rho_{ij}\sqrt{\mathrm{Var}(\tilde{d}_i)\mathrm{Var}(\tilde{d}_j)} \\ &\leq \frac{1}{N}\sum_{i \neq j}\rho_{ij} C/N^q \\ &= \frac{C(N-1)\bar{\rho}}{N^q}.\end{aligned}$$

The quantity $C(N-1)\bar{\rho}/N^q$ converges to zero by condition 2 in Section 3.3.2.

Next, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(\hat{d}_i) &= \frac{1}{N}\sum_{i=1}^{N}\left\{\left[\mathrm{E}(\hat{d}_i - d_i)\right]^2 + \mathrm{Var}(\hat{d}_i)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{(d_{0i} - d_i)^2 + \mathrm{Var}(\hat{d}_i)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{[(d_{0i} - d_{\infty i}) + (d_{\infty i} - d_i)]^2 + \mathrm{Var}(\hat{d}_i)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{(d_{0i} - d_{\infty i})^2 + (d_{\infty i} - d_i)^2 + 2(d_{0i} - d_{\infty i})(d_{\infty i} - d_i) + \mathrm{Var}(\hat{d}_i)\right\}.
\end{aligned}$$

Under the assumptions in Section 3.3.2, $\frac{1}{N}\sum_{i=1}^{N}(d_{\infty i} - d_i)^2$ converges to a constant K while the remaining terms converge to zero.

This implies that

$$\frac{\sum_{i\neq j}\gamma_{ij}}{\sum_{i=1}^{N}\mathrm{MSE}(\hat{d}_i)} \longrightarrow 0.$$

APPENDIX J

Bound on the Mean Squared Error of $\hat{d}_i$

We bound the term $\sum_{i=1}^{N} \mathrm{MSE}(\hat{d}_i)/N^2$. We can express the mean squared error of $\hat{d}_i$ as

$$\begin{aligned}
\mathrm{MSE}(\hat{d}_i) &= \mathrm{E}\left\{(d_i - \hat{d}_i)^2\right\} \\
&= \mathrm{E}\left\{\left(\frac{1}{2}(a_i - b_i) - \frac{1}{2}(\hat{a}_i - \hat{b}_i)\right)^2\right\} \\
&= \mathrm{E}\left\{\left(\frac{1}{2}(a_i - \hat{a}_i) - \frac{1}{2}(b_i - \hat{b}_i)\right)^2\right\} \\
&= \frac{1}{4}\mathrm{E}\left\{(a_i - \hat{a}_i)^2 - 2(a_i - \hat{a}_i)(b_i - \hat{b}_i) + (b_i - \hat{b}_i)^2\right\}
\end{aligned}$$

and the mean squared error of $\hat{W}_i$ as

$$\begin{aligned}
\mathrm{MSE}(\hat{W}_i) &= \mathrm{E}\left\{(W_i - \hat{W}_i)^2\right\} \\
&= \mathrm{E}\left[\mathrm{E}\left\{(W_i - \hat{W}_i)^2 | \hat{a}_i, \hat{b}_i\right\}\right] \\
&= \frac{1}{2}\mathrm{E}\left\{(a_i - \hat{a}_i)^2 + (b_i - \hat{b}_i)^2\right\}.
\end{aligned}$$

Next, we show that $\mathrm{MSE}(\hat{d}_i) \leq \mathrm{MSE}(\hat{W}_i)$:

$$\begin{aligned}
\mathrm{MSE}(\hat{W}_i) - \mathrm{MSE}(\hat{d}_i) &= \frac{1}{2}\mathrm{E}\left\{(a_i - \hat{a}_i)^2 + (b_i - \hat{b}_i)^2\right\} - \\
&\qquad \frac{1}{4}\mathrm{E}\left\{(a_i - \hat{a}_i)^2 - 2(a_i - \hat{a}_i)(b_i - \hat{b}_i) + (b_i - \hat{b}_i)^2\right\} \\
&= \mathrm{E}\left\{\frac{1}{4}(a_i - \hat{a}_i)^2 + \frac{1}{4}(b_i - \hat{b}_i)^2 + \frac{1}{2}(a_i - \hat{a}_i)(b_i - \hat{b}_i)\right\} \\
&= \mathrm{E}\left[\left\{\frac{1}{2}(a_i - \hat{a}_i) + \frac{1}{2}(b_i - \hat{b}_i)\right\}^2\right] \\
&\geq 0.
\end{aligned}$$

We therefore have the bound

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathrm{MSE}(\hat{d}_i) \leq \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{MSE}(\hat{W}_i).$$

APPENDIX K

Equivalence of P-LOOP and the Simple Difference Estimator

We consider the special case where we set $\hat{a}_i = \hat{b}_i = \bar{W}^{(-i)}$ where $\bar{W}^{(-i)} = \sum_{j \neq i} W_j/(N-1)$. In this case we estimate the variance of P-LOOP as:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\tau}) &= \frac{1}{N^2}\sum_{i=1}^{N}(W_i - \hat{W}_i)^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{W_i - \frac{1}{N-1}\sum_{j\neq i} W_j\right\}^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{W_i - \frac{1}{N-1}\left(\sum_{j=1}^{N} W_j - W_i\right)\right\}^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{N-1}{N-1}W_i + \frac{1}{N-1}W_i - \frac{N}{N-1}\frac{1}{N}\sum_{j=1}^{N} W_j\right\}^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{N}{N-1}W_i - \frac{N}{N-1}\hat{\tau}_{sd}\right\}^2 \\
&= \frac{1}{(N-1)^2}\sum_{i=1}^{N}(W_i - \hat{\tau}_{sd})^2
\end{aligned}$$

The standard variance estimator for the simple difference estimator in paired experiments is

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} (W_i - \hat{\tau}_{sd})^2 .$$

For example, see Imai (2008). Thus in this special case, the variance estimate for P-LOOP is equal to $N/(N-1)$ times the standard variance estimate for the simple difference estimator.

APPENDIX L

Simulation Procedure for Chapter III

L.1 Simulation Procedure for Sections 3.5.1 and 3.5.2

We describe the simulation procedure used in Section 3.5. For each of the two scenarios presented in Table 3.1 and for the scenario presented in Table 3.2, we generate a single set of potential outcomes and covariates. For each method, we obtain Var_{true} (an estimate for the true variance) and $\text{E}(\text{Var}_{\text{nom}})$ (an estimate for the expected value of the nominal variance) for the generated potential outcomes and covariates assuming pair randomization. We also estimate the coverage probability at a 95% confidence level.

To estimate these quantities, we generate 10,000 treatment assignment vectors. For each treatment assignment vector, we obtain a point estimate and nominal variance using each estimator. We calculate Var_{true} by taking the variance of the 10,000 point estimates. We calculate $\text{E}(\text{Var}_{\text{nom}})$ as the average of the 10,000 nominal variance estimates.

To estimate the coverage probability for a given method, we create a confidence interval of the form (point estimate) $\pm$ 1.96 $\times$ (nominal standard error) for each treatment assignment vector. We then estimate the coverage probability for that method as the proportion of time that the constructed confidence intervals include the average treatment effect.

Finally, we also estimate the standard errors for the variance estimates, which we will call $\text{SE}(\text{Var}_{\text{true}})$ and $\text{SE}(\text{E}(\text{Var}_{\text{nom}}))$. To estimate $\text{SE}(\text{Var}_{\text{true}})$, we split the 10,000 point estimates into 200 sets of 50 estimates. We take the variance of each set of 50 point estimates and take the standard deviation of these 200 variances, which we divide by the square root of 200 to obtain an estimate for $\text{SE}(\text{Var}_{\text{true}})$. We take the standard deviation of the 10,000 nominal variance estimates divided by the square root of 10,000 to estimate $\text{SE}(\text{E}(\text{Var}_{\text{nom}}))$.

In Table L.1, we show the results from Section 3.5.1 with additional columns for the values of $\text{SE}(\text{Var}_{\text{true}})$ and $\text{SE}(\text{E}(\text{Var}_{\text{nom}}))$. We show the results from Section 3.5.2 in Table L.2. In both tables, we refer to the differences, outcomes, and interpolation imputation approaches for P-LOOP using "(D)", "(O)", and "(I)" respectively.

Table L.1: Simulation Results for Section 3.5.1

Simpson's paradox					
Method	Var_{true}	$\text{SE}(\text{Var}_{\text{true}})$	$\text{E}(\text{Var}_{\text{nom}})$	$\text{SE}(\text{E}(\text{Var}_{\text{nom}}))$	Cov Prob
Simple Difference	0.343	0.0045	0.342	0.000064	0.943
P-LOOP RF (D)	0.154	0.0022	0.167	0.000064	0.951
P-LOOP RF (O)	0.440	0.0065	0.462	0.000225	0.952
P-LOOP RF (I)	0.152	0.0022	0.170	0.000063	0.953
P-LOOP OLS (D)	0.152	0.0022	0.160	0.000043	0.950
P-LOOP OLS (O)	0.442	0.0066	0.462	0.000227	0.953
P-LOOP OLS (I)	0.152	0.0022	0.164	0.000048	0.952
Regression 1	0.151	0.0022	0.150	0.000019	0.943
Regression 2	0.153	0.0022	0.148	0.000027	0.942
Uninformative pairs					
Method	Var_{true}	$\text{SE}(\text{Var}_{\text{true}})$	$\text{E}(\text{Var}_{\text{nom}})$	$\text{SE}(\text{E}(\text{Var}_{\text{nom}}))$	Cov Prob
Simple Difference	0.361	0.0049	0.365	0.000071	0.947
P-LOOP RF (D)	0.151	0.0023	0.168	0.000063	0.952
P-LOOP RF (O)	0.146	0.0022	0.154	0.000029	0.949
P-LOOP RF (I)	0.148	0.0023	0.156	0.000038	0.948
P-LOOP OLS (D)	0.148	0.0023	0.160	0.000042	0.950
P-LOOP OLS (O)	0.146	0.0022	0.154	0.000029	0.949
P-LOOP OLS (I)	0.148	0.0023	0.156	0.000033	0.949
Regression 1	0.148	0.0023	0.149	0.000018	0.944
Regression 2	0.149	0.0023	0.148	0.000026	0.940

Table L.2: Simulation Results for Section 3.5.2

Method	Var_{true}	$\text{SE}(\text{Var}_{\text{true}})$	$\text{E}(\text{Var}_{\text{nom}})$	$\text{SE}(\text{E}(\text{Var}_{\text{nom}}))$	Cov Prob
Simple Difference	0.094	0.00131	0.373	0.000375	1
P-LOOP RF (D)	0.069	0.00102	0.160	0.000109	0.996
P-LOOP RF (O)	0.046	0.00064	0.097	0.000047	0.991
P-LOOP RF (I)	0.046	0.00062	0.097	0.000048	0.992
P-LOOP OLS (D)	0.068	0.00101	0.371	0.000423	1
P-LOOP OLS (O)	0.062	0.00091	0.364	0.000401	1
P-LOOP OLS (I)	0.065	0.00097	0.363	0.000392	1
Regression 1	0.066	0.00097	0.351	0.000363	1
Regression 2	0.066	0.00097	0.358	0.000380	1

L.2 Simulating the Remainder Terms

As noted in Appendix I, $\gamma_{ij} = \text{Cov}\Big(\hat{d}_i U_i, \hat{d}_j U_j\Big) = \text{Cov}\Big(\tilde{d}_i U_i, \tilde{d}_j U_j\Big)$. Then we have

$$\begin{aligned}
\frac{1}{N}\sum_{i\neq j}\gamma_{ij} &= \frac{1}{N}\sum_{i\neq j}\text{Cov}\Big(\tilde{d}_i U_i, \tilde{d}_j U_j\Big) \\
&\leq \frac{1}{N}\sum_{i,j}\text{Cov}\Big(\tilde{d}_i U_i, \tilde{d}_j U_j\Big) \\
&= \frac{1}{N}\sum_{i,j}\text{E}\Big(\tilde{d}_i U_i \tilde{d}_j U_j\Big) \\
&= \frac{1}{N}\text{E}\left\{\left(\sum_{i=1}^{N}\tilde{d}_i U_i\right)^2\right\}.
\end{aligned}$$

Similarly, we can show

$$\frac{1}{N}\sum_{i\neq j}\gamma_{ij} \geq -\frac{1}{N}\text{E}\left\{\left(\sum_{i=1}^{N}\tilde{d}_i U_i\right)^2\right\}.$$

We now describe the simulation procedure in more detail. For each of the three data

generation processes used in Sections 3.5.1 and 3.5.2, we generate potential outcomes and covariates for 1000 pairs. We then estimate the quantity (3.5) for each of the first $N = 50, 100, \ldots, 1000$ of these pairs. For a given N, we generate treatment assignment vectors $U^{(t)} = (U_{1,t}, \ldots, U_{N,t})$ for $t = 1, \ldots, 1000$. For each treatment assignment vector $U^{(t)}$, we calculate $\hat{d}_{i,t}$ for each pair i, resulting in 1000 estimates for d_i. Although d_{0i} is not known, we can estimate it using the quantity $\sum_{k=1}^{1000} \hat{d}_{i,k}/1000$. We then obtain simulation estimates for $\tilde{d}_i$ by centering the values of $\hat{d}_{i,t}$:

$$\tilde{d}^*_{i,t} = \hat{d}_{i,t} - \frac{1}{1000} \sum_{k=1}^{1000} \hat{d}_{i,k}.$$

Finally, we estimate (3.5) as

$$\frac{1}{N} \left\{ \frac{1}{1000} \sum_{t=1}^{1000} \left(\sum_{i=1}^{N} \tilde{d}^*_{i,t} U_{i,t} \right)^2 \right\}.$$

We repeat this procedure for each sample size $N = 50, 100, \ldots, 1000$ and for both random forest and OLS imputation.

As noted in Section 3.5.3, the estimate for (3.5) shrinks more slowly when using random forest in the non-linear data generating process. We argue that this is due to the variance of the imputation method. Like many machine learning methods, random forests experience a bias-variance trade-off. One parameter of the random forest is the maximum number of nodes for each decision tree. If the number of nodes in a tree is large, this could result in overfitting and higher variance (for example, see James et al. (2013)). As the number of nodes increases, each node will have correspondingly fewer observations, and the predicted outcomes will be highly variable. This could be especially true if the amount of nodes increases with the number of observations. To illustrate, we perform the non-linear simulation again, but limit the number of nodes in each random forest to 10, reducing the variance of the imputed potential outcomes. We present the results in Figure L.1. When we limit the number of

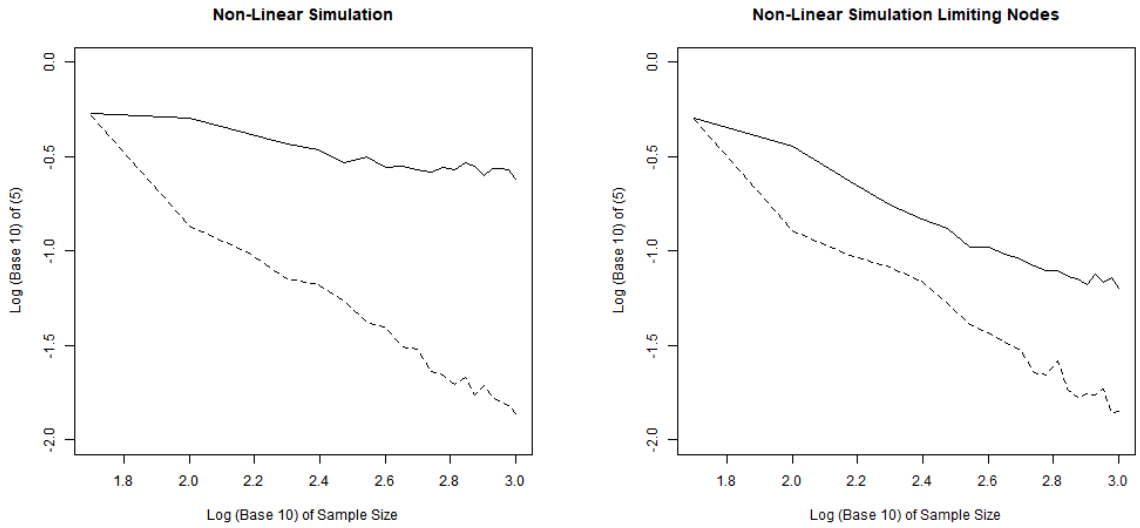

Figure L.1: We plot the estimated values of quantity (3.5) (*i.e.*, $\mathrm{E}\{(\sum_{i=1}^{N} \tilde{d}_i U_i)^2\}/N$) against the sample size N. Both values are plotted on a log base 10 scale. The left chart shows The values of (3.5) are estimated for both random forest imputation (solid line) and OLS imputation (dashed line).

nodes, the estimate of (3.5) for random forest imputation shrinks more quickly.

APPENDIX M

Asymptotic Normality of the LOOP Estimator

In this section we prove that the LOOP estimator is asymptotically normally distributed under the assumptions outlined in Section 4.4. Recall the assumptions:

1. There exists some $0 < C < \infty$ and $q > 0$ such that for all i,

$$\text{Var}(\tilde{m}_i) = \text{Var}(\hat{m}_i) \leq C/N^q.$$

2. Let ρ_{ij} be the correlation of $\tilde{m}_i U_i$ and $\tilde{m}_j U_j$, and $\bar{\rho} = \frac{\sum_{i \neq j} \rho_{ij}}{N(N-1)}$. We assume that

$$N^{1-q}\bar{\rho} \longrightarrow 0.$$

3. Recall that $m_{0i}^{(N)} = \text{E}(\hat{m}_i^{(N)})$ for some fixed N. For each observation i, we assume that the limit of $m_{0i}^{(N)}$ exists and denote the limit as $m_{\infty i}$. We also assume

$$\frac{1}{N}\sum_{i=1}^{N}\left(m_{0i}^{(N)} - m_{\infty i}\right)^2 \longrightarrow 0.$$

4. There exists $0 < K < \infty$ such that

$$\frac{\sum_{i=1}^{N}(m_i - m_{\infty i})^2}{N} \longrightarrow K,$$

and

$$\max_{i=1,\ldots,N} \frac{(m_i - m_{\infty i})^2}{\sum_{k=1}^{N}(m_k - m_{\infty k})^2} \longrightarrow 0.$$

5. There exists $0 < \epsilon < 0.5$ such that $\epsilon < p_i < 1 - \epsilon$ for all i.

We now proceed with the proof. First, let $V_N = \sum_{i=1}^{N} \frac{(m_i - m_{\infty i})^2}{p_i(1-p_i)}$. We write $N(\hat{\tau} - \tau)/\sqrt{V_N}$ as

$$\begin{aligned}
N(\hat{\tau} - \tau)/\sqrt{V_N} &= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} (\hat{\tau}_i - \tau_i) \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \{(Y_i - \hat{m}_i)U_i - \tau_i\} \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \{(Y_i - (\tilde{m}_i + m_{0i}))U_i - \tau_i\} \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \{(Y_i - m_{0i})U_i - \tau_i\} - \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{m}_i U_i \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \{(Y_i - m_{\infty i} + m_{\infty i} - m_{0i})U_i - \tau_i\} - \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{m}_i U_i \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \{(Y_i - m_{\infty i})U_i - \tau_i\} + \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} (m_{\infty i} - m_{0i})\, U_i - \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{m}_i U_i \\
&= \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} X_i + \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} (m_{\infty i} - m_{0i})\, U_i - \frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} \tilde{m}_i U_i \qquad \text{(M.1)} \\
&= \text{(A)} + \text{(B)} - \text{(C)}
\end{aligned}$$

where we define $X_i = (Y_i - m_{\infty i})U_i - \tau_i$. We show that the first term converges in distribution to a standard normal random variable and the other two terms converge in probability to zero.

Asymptotic Normality of (A) We show that

$$\frac{1}{\sqrt{V_N}} \sum_{i=1}^{N} X_i \xrightarrow{d} \mathrm{N}(0, 1)$$

by the Lindeberg-Feller central limit theorem (see for example, Chapter 9.8 of Resnick (2003)). We note that

$$\begin{aligned} Y_i U_i - \tau_i &= \frac{t_i}{p_i} T_i - \frac{c_i}{1-p_i}(1-T_i) - (t_i - c_i) \\ &= \frac{t_i - p_i(t_i - c_i)}{p_i} T_i - \frac{c_i - (1-p_i)(t_i - c_i)}{1-p_i}(1-T_i) \\ &= \frac{(1-p_i)t_i - p_i c_i}{p_i} T_i - \frac{(1-p_i)t_i - p_i c_i}{1-p_i}(1-T_i) \\ &= \frac{m_i}{p_i} T_i - \frac{m_i}{1-p_i}(1-T_i) \\ &= m_i U_i \end{aligned}$$

and therefore $X_i = (m_i - m_{\infty i})U_i$. The X_i are independent random variables, as the U_i are all independent, and m_i and $m_{\infty i}$ are constants. The X_i have expectation and variance

$$\begin{aligned} \mathrm{E}\left((m_i - m_{\infty i})U_i\right) &= (m_i - m_{\infty i})\mathrm{E}(U_i) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}\left((m_i - m_{\infty i})U_i\right) &= \mathrm{E}\left\{\left((m_i - m_{\infty i})U_i\right)^2\right\} \\ &= (m_i - m_{\infty i})^2 \mathrm{E}\left(U_i^2\right) \\ &= (d_i - d_{\infty i})^2 \frac{1}{p_i(1-p_i)}. \end{aligned}$$

Thus $V_N = \sum_{i=1}^N \text{Var}(X_i)$. In addition, X_i satisfies the Lindeberg condition. For any $t > 0$, we have

$$\begin{aligned}
\frac{1}{V_N}\sum_{i=1}^N \text{E}\left(X_i^2 \mathbb{1}\{|X_i| > t\sqrt{V_N}\}\right) &= \frac{1}{V_N}\sum_{i=1}^N \text{E}\left((m_i - m_{\infty i})^2 U_i^2 \mathbb{1}\{|(m_i - m_{\infty i})U_i| > t\sqrt{V_N}\}\right) \\
&= \frac{1}{V_N}\sum_{i=1}^N (m_i - m_{\infty i})^2 \text{E}\left(U_i^2 \mathbb{1}\{|(m_i - m_{\infty i})U_i| > t\sqrt{V_N}\}\right) \\
&= \frac{1}{V_N}\sum_{i=1}^N (m_i - m_{\infty i})^2 \text{E}\left(U_i^2 \mathbb{1}\{(m_i - m_{\infty i})^2 U_i^2 > t^2 V_N\}\right).
\end{aligned}$$

Next, note that U_i^2 can take the value $1/p_i^2$ or $1/(1-p_i)^2$. By assumption 5, it follows that $U_i^2 < 1/\epsilon^2$ and $\mathbb{1}\{(m_i - m_{\infty i})^2 U_i^2 > t^2 V_N\} \leq \mathbb{1}\{(m_i - m_{\infty i})^2/\epsilon^2 > t^2 V_N\}$. In addition, we have $V_N = \sum_{i=1}^N \frac{(m_i - m_{\infty i})^2}{p_i(1-p_i)} \geq 4\sum_{i=1}^N (m_i - m_{\infty i})^2$, as $\frac{1}{p_i(1-p_i)} \geq 4$ for any $p_i \in [0,1]$. We therefore have:

$$\begin{aligned}
&\frac{1}{V_N}\sum_{i=1}^N (m_i - m_{\infty i})^2 \text{E}\left(U_i^2 \mathbb{1}\{(m_i - m_{\infty i})^2 U_i^2 > t^2 V_N\}\right) \\
&\leq \frac{1}{V_N}\sum_{i=1}^N \frac{(m_i - m_{\infty i})^2}{\epsilon^2} \mathbb{1}\{(m_i - m_{\infty i})^2/\epsilon^2 > t^2 V_N\} \\
&= \frac{1}{V_N}\sum_{i=1}^N \frac{(m_i - m_{\infty i})^2}{\epsilon^2} \mathbb{1}\{(m_i - m_{\infty i})^2/V_N > t^2\epsilon^2\} \\
&\leq \frac{1}{V_N}\sum_{i=1}^N \frac{(m_i - m_{\infty i})^2}{\epsilon^2} \mathbb{1}\left\{\frac{(m_i - m_{\infty i})^2}{\sum_{i=1}^N (m_i - m_{\infty i})^2} > 4t^2\epsilon^2\right\} \longrightarrow 0
\end{aligned}$$

by assumption 4. It follows that $\frac{1}{V_N}\sum_{i=1}^N \text{E}\left(X_i^2 \mathbb{1}\{|X_i| > t\sqrt{V_N}\}\right) \longrightarrow 0$ for any $t > 0$, satisfying the Lindeberg condition. Thus, the first term of equation (M.1) converges to a normal distribution by the Lindeberg-Feller central limit theorem.

(B) Converges in Probability to Zero Next we show that

$$\frac{1}{\sqrt{V_N}}\sum_{i=1}^N (m_{\infty i} - m_{0i})U_i$$

converges in L^2-norm to zero.

$$\begin{aligned}
\mathrm{E}\left\{\left(\frac{1}{\sqrt{V_N}}\sum_{i=1}^{N}(m_{\infty i}-m_{0i})U_i\right)^2\right\} &= \frac{1}{V_N}\mathrm{E}\left\{\sum_{i=1}^{N}(m_{\infty i}-m_{0i})^2U_i^2+\right. \\
&\qquad \left.\sum_{i\neq j}(m_{\infty i}-m_{0i})(m_{\infty j}-m_{0j})U_iU_j\right\} \\
&= \frac{1}{V_N}\sum_{i=1}^{N}(m_{\infty i}-m_{0i})^2U_i^2 \\
&= \frac{N}{V_N}\frac{1}{N}\sum_{i=1}^{N}(m_{\infty i}-m_{0i})^2U_i^2 \\
&\leq \frac{N}{V_N\epsilon^2}\frac{1}{N}\sum_{i=1}^{N}(m_{\infty i}-m_{0i})^2 \\
&\leq \frac{N}{4\epsilon^2\sum_{i=1}^{N}(m_i-m_{\infty i})^2}\frac{1}{N}\sum_{i=1}^{N}(m_{\infty i}-m_{0i})^2
\end{aligned}$$

where $\mathrm{E}\{(m_{\infty i}-m_{0i})(m_{\infty j}-m_{0j})U_iU_j\} = 0$ since U_i and U_j have expectation zero and are independent. The quantity $\sum_{i=1}^{N}(m_{\infty i}-m_{0i})^2/N$ converges to zero by assumption 3, while $\sum_{i=1}^{N}(m_i-m_{\infty i})^2/N$ (and therefore $N/\sum_{i=1}^{N}(m_i-m_{\infty i})^2$) converges to a constant by assumption 4. Thus

$$\mathrm{E}\left\{\left(\frac{1}{\sqrt{V_N}}\sum_{i=1}^{N}(d_{\infty i}-d_{0i})U_i\right)^2\right\} \longrightarrow 0.$$

Since (B) converges in L^2-norm to zero, it also converges in probability to zero.

(C) Converges in Probability to Zero Finally we show that the last term in equation (M.1) converges in L^2-norm to zero:

$$\begin{aligned}
\mathrm{E}\left\{\left(\frac{1}{\sqrt{V_N}}\sum_{i=1}^{N}\tilde{m}_i U_i\right)^2\right\} &= \frac{1}{V_N}\sum_{i,j}\mathrm{E}(\tilde{m}_i U_i \tilde{m}_j U_j) \\
&= \frac{1}{V_N}\sum_{i,j}\mathrm{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j) \\
&= \frac{1}{V_N}\sum_{i,j}\rho_{ij}\sqrt{\mathrm{Var}(\tilde{m}_i U_i)\mathrm{Var}(\tilde{m}_j U_j)} \\
&\leq \frac{1}{V_N}\sum_{i,j}\rho_{ij}\sqrt{\frac{1}{\epsilon^2}\mathrm{Var}(\tilde{m}_i)\frac{1}{\epsilon^2}\mathrm{Var}(\tilde{m}_j)} \\
&\leq \frac{1}{V_N}\sum_{i,j}\frac{C\rho_{ij}}{N^q\epsilon^2} \\
&= \frac{C}{V_N N^q \epsilon^2}\sum_{i,j}\rho_{ij} \\
&= \frac{C}{V_N N^q \epsilon^2}\left\{N + \sum_{i\neq j}\rho_{ij}\right\} \\
&= \frac{N}{V_N\epsilon^2}\frac{C}{N^q} + \frac{N}{V_N\epsilon^2}\frac{C}{N^q}\left\{(N-1)\bar{\rho}\right\} \\
&\leq \frac{N}{4\epsilon^2\sum_{i=1}^{N}(m_i - m_{\infty i})^2}\frac{C}{N^q} + \frac{N}{4\epsilon^2\sum_{i=1}^{N}(m_i - m_{\infty i})^2}\frac{C}{N^q}\left\{(N-1)\bar{\rho}\right\}.
\end{aligned}$$

As noted above, $\frac{N}{4\epsilon^2\sum_{i=1}^{N}(m_i - m_{\infty i})^2}$ converges to a constant. Due to assumption 2, $\frac{C}{N^q}\left\{(N-1)\bar{\rho}\right\}$ converges to zero. It follows that the last term of equation (M.1) converges in L^2-norm (and therefore in probability) to zero.

Combining the results for (A), (B), and (C) implies that equation (M.1) converges in distribution to a standard normal random variable.

APPENDIX N

Asymptotic Normality when Imputing Potential Outcomes Using Simple Linear Regression

Let $\{(c_i, t_i, Z_i), i = 1, 2, \dots\}$ be an infinite sequence of experimental observations such that c_i, t_i, and Z_i are all bounded. We assume that each observation is assigned to treatment with probability p. We also assume that $\frac{1}{N}\sum_{i=1}^{N} c_i$, $\frac{1}{N}\sum_{i=1}^{N} c_i^2$, $\frac{1}{N}\sum_{i=1}^{N} t_i$, $\frac{1}{N}\sum_{i=1}^{N} t_i^2$, $\frac{1}{N}\sum_{i=1}^{N} Z_i$, and $\frac{1}{N}\sum_{i=1}^{N} Z_i^2$ all converge, and denote these limits $\bar{c}$, $\overline{c^2}$,$\bar{t}$, $\overline{t^2}$,$\bar{Z}$, and $\overline{Z^2}$. Finally, let ρ^t be the limiting correlation between the treatment outcomes and covariates, and ρ^c be the limiting correlation between the control outcomes and covariates. We assume $-1 < \rho^t < 1$ and $-1 < \rho^c < 1$. Under these conditions, we demonstrate that the LOOP estimator is asymptotically normally distributed when imputing potential outcomes using a linear regression with the single covariate Z. We do this by showing that the assumptions outlined in Appendix M are met.

For a given N, let $\mathcal{T}^{(N)}$ and $\mathcal{C}^{(N)}$ denote the indices of observations in the treated and control groups, and let $n_t^{(N)}$ and $n_c^{(N)}$ be the size of these sets. Let $\bar{Y}_{-i}^{(t,N)}$ and $\bar{Z}_{-i}^{(t,N)}$ be the averages of the Y and Z values in $\mathcal{T}^{(N)}\backslash i$, and $\bar{Y}_{-i}^{(c,N)}$, and $\bar{Z}_{-i}^{(c,N)}$ be the averages for $\mathcal{C}^{(N)}\backslash i$. These values all depend on N; however, for the remainder of this section we will generally suppress the N superscript.

For each i, we obtain $\hat{t}_i$ and $\hat{c}_i$ by regressing Y onto Z for the observations in $\mathcal{T}\backslash i$ and $\mathcal{C}\backslash i$ respectively. For example, we have

$$\hat{t}_i = \hat{\alpha}_i^t + \hat{\beta}_i^t Z_i$$

where

$$\hat{\alpha}_i^t = \bar{Y}^t - \hat{\beta}_i^t \bar{Z}^t$$

and

$$\hat{\beta}_i^t = \frac{\sum_{k\in\mathcal{T}\backslash i}(Z_k - \bar{Z}_{-i}^t)(Y_k - \bar{Y}_{-i}^t)}{\sum_{k\in\mathcal{T}\backslash i}(Z_k - \bar{Z}_{-i}^t)^2}.$$

The terms $\hat{c}_i$, $\hat{\alpha}_i^c$, and $\hat{\beta}_i^t$ are defined similarly.

Assumption (1): *There exists some* $0 < C < \infty$ *and* $q > 0$ *such that for all* i*,*

$$Var(\tilde{m}_i) = Var(\hat{m}_i) \leq C/N^q.$$

Our assumptions imply those of Freedman (2008), who shows that regression coefficients go to zero at a rate $1/N$. In addition, because the covariates and outcomes are bounded, a uniform bound on the variances for the $\tilde{m}_i$ follows. Thus assumption 1 holds with $q = 1$.

Assumption (2): *Let* ρ_{ij} *be the correlation of* $\tilde{m}_i U_i$ *and* $\tilde{m}_j U_j$*, and* $\bar{\rho} = \frac{\sum_{i\neq j}\rho_{ij}}{N(N-1)}$*. We assume that*

$$N^{1-q}\bar{\rho} \longrightarrow 0.$$

Because $q = 1$, we need to show that $\bar{\rho}$ converges to zero. We have

$$\begin{aligned}
\bar{\rho} &= \frac{\sum_{i \neq j} \rho_{ij}}{N(N-1)} \\
&= \frac{1}{N(N-1)} \sum_{i \neq j} \frac{\text{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j)}{\sqrt{\text{Var}(\tilde{m}_i U_i)\text{Var}(\tilde{m}_j U_j)}} \\
&= \frac{p(1-p)}{N(N-1)} \sum_{i \neq j} \frac{\text{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j)}{\sqrt{\text{Var}(\tilde{m}_i)\text{Var}(\tilde{m}_j)}}.
\end{aligned}$$

Because $\text{Var}(\tilde{m}_i)$ and $\text{Var}(\tilde{m}_j)$ go to zero at a rate $1/N$, we need

$$\frac{1}{N} \sum_{i \neq j} \text{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j)$$

to go to zero in order for $\bar{\rho}$ to go to zero.

We first consider the covariance for a fixed pair i and j:

$$\begin{aligned}
\text{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j) &= \text{Cov}(\hat{m}_i U_i, \hat{m}_j U_j) \\
&= \text{E}(\hat{m}_i \hat{m}_j U_i U_j) \\
&= \text{E}\Big\{ \Big[(1-p)^2 (\hat{\alpha}_i^t + \hat{\beta}_i^t Z_i)(\hat{\alpha}_j^t + \hat{\beta}_j^t Z_j) \\
&\quad + p(1-p)(\hat{\alpha}_i^t + \hat{\beta}_i^t Z_i)(\hat{\alpha}_j^c + \hat{\beta}_j^c Z_j) \\
&\quad + p(1-p)(\hat{\alpha}_i^c + \hat{\beta}_i^c Z_i)(\hat{\alpha}_j^t + \hat{\beta}_j^t Z_j) \\
&\quad + p^2 (\hat{\alpha}_i^c + \hat{\beta}_i^c Z_i)(\hat{\alpha}_j^c + \hat{\beta}_j^c Z_j) \Big] U_i U_j \Big\}.
\end{aligned}$$

We will show that the values of $\text{E}(\hat{\beta}_i^t \hat{\beta}_j^t U_i U_j)$ go to zero at a rate $1/N^2$ for all combinations of t, c, α, and β. To do this we condition on the treatment assignments for the remaining observations. Let $\hat{\alpha}_i^{t,-j}$ be the value of $\hat{\alpha}_i^t$ calculated without observation j (*i.e.*, when $T_j = 0$), and $\hat{\alpha}_i^{t,+j}$ be the value calculated with observation j ($T_j = 1$). Define $\hat{\beta}_j^{c,-i}$ and $\hat{\beta}_j^{c,+i}$ similarly. Then we have

$$\begin{aligned}\mathrm{E}(\hat{\alpha}_i^t\hat{\beta}_j^c U_i U_j | U_k, k \notin \{i,j\}) &= \hat{\alpha}_i^{t,-j}\hat{\beta}_j^{c,+i} - \hat{\alpha}_i^{t,+j}\hat{\beta}_j^{c,+i} - \hat{\alpha}_i^{t,-j}\hat{\beta}_j^{c,-i} + \hat{\alpha}_i^{t,+j}\hat{\beta}_j^{c,-i} \\ &= (\hat{\alpha}_i^{t,-j} - \hat{\alpha}_i^{t,+j})(\hat{\beta}_j^{c,+i} - \hat{\beta}_j^{c,-i}).\end{aligned}$$

To see why this quantity is on the order of $1/N^2$, consider the simple linear regression of an outcome y_i onto a predictor x_i for $i = 1, \ldots, n$, and let $\hat{\alpha}$ and $\hat{\beta}$ be the fitted slope and intercept. Suppose we calculate wish to estimate the coefficients without observation j. It can be shown the corresponding coefficients are given by

$$\hat{\alpha}_{-j} = \hat{\alpha} - \frac{S_x - n\bar{x}(x_j - \bar{x})}{(n-1)S_x - n(x_j - \bar{x})^2} \times (y_j - \hat{y}_j) \tag{N.1}$$

and

$$\hat{\beta}_{-j} = \hat{\beta} - \frac{n(x_j - \bar{x})}{(n-1)S_x - n(x_j - \bar{x})^2} \times (y_j - \hat{y}_j) \tag{N.2}$$

where $\hat{y}_j$ is the fitted value from the regression with all of the observations, $\bar{x}$ is the average of all the x_i, and $S_x = \sum_{i=1}^n (x_i - \bar{x})^2$. We can use these formulas to compare $\hat{\alpha}_i^{t,-j}$ and $\hat{\alpha}_i^{t,+j}$, and $\hat{\beta}_j^{c,+i}$ and $\hat{\beta}_j^{c,-i}$. For example,

$$\hat{\beta}_j^{c,+i} - \hat{\beta}_j^{c,-i} = \frac{n_{c'}(Z_i - \bar{Z}_{-j}^c)}{(n_{c'} - 1)\sum_{k \in \mathcal{C}\setminus j}(Z_k - \bar{Z}_{-j}^c)^2 - n_{c'}(Z_i - \bar{Z}_{-j}^c)^2} \times (Y_j - \hat{c}_j)$$

where $n_{c'}$ is the number of observations in $\mathcal{C}\setminus j$. As we will show for assumption 3, the value $\sum_{k \in \mathcal{C}\setminus j}(Z_k - \bar{Z}_{-j}^c)^2/n_c'$ converges to a constant. In addition, the values of $Z_i - \bar{Z}_{-j}^c$ are bounded. Thus, $\hat{\beta}_j^{c,+i} - \hat{\beta}_j^{c,-i}$ is on the order of $1/n_c \approx 1/N(1-p)$. A similar argument can be used to show that $\hat{\alpha}_i^{t,-j} - \hat{\alpha}_i^{t,+j}$ is also on the order of $1/N$, and therefore $\mathrm{E}(\hat{\alpha}_i^t\hat{\beta}_j^c U_i U_j | U_k, k \notin \{i,j\})$ is on the order of $1/N^2$. However, this same argument holds across randomizations, so we have $E(\hat{\alpha}_i^t\hat{\beta}_j^c U_i U_j) = \mathrm{E}(\mathrm{E}(\hat{\alpha}_i^t\hat{\beta}_j^c U_i U_j | U_k, k \notin \{i,j\}))$ is on the order of $1/N^2$. This same

argument can be repeated for the remaining combination of coefficients, which implies that $\text{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j)$ is on the order of $1/N^2$.

Finally, we can repeat the same argument for any i and j. It follows that

$$\frac{1}{N} \sum_{i \neq j} \text{Cov}(\tilde{m}_i U_i, \tilde{m}_j U_j)$$

converges to zero, as desired.

Assumption (3): *Recall that $m_{0i}^{(N)} = E(\hat{m}_i^{(N)})$ for some fixed N. For each observation i, we assume that the limit of $m_{0i}^{(N)}$ exists and denote the limit as $m_{\infty i}$. We also assume*

$$\frac{1}{N} \sum_{i=1}^{N} \left(m_{0i}^{(N)} - m_{\infty i} \right)^2 \longrightarrow 0.$$

We first show that $m_{\infty i} = \lim m_{0i}^{(N)}$ exists. For a given observation i, we have

$$\begin{aligned} \hat{\beta}_i^t &= \frac{\sum_{k \in \mathcal{T} \setminus i} (Z_k - \bar{Z}_{-i}^t)(Y_k - \bar{Y}_{-i}^t)}{\sum_{k \in \mathcal{T} \setminus i} (Z_k - \bar{Z}_{-i}^t)^2} \\ &= \frac{\frac{1}{n_t - 1} \sum_{k \in \mathcal{T} \setminus i} (Z_k - \bar{Z}_{-i}^t)(Y_k - \bar{Y}_{-i}^t)}{\frac{1}{n_t - 1} \sum_{k \in \mathcal{T} \setminus i} (Z_k - \bar{Z}_{-i}^t)^2}. \end{aligned}$$

We show that both the numerator and denominator converge. Note that $Y_k = t_k$ for all $k \in \mathcal{T}$. In addition, we can treat the observations in set $\mathcal{T}$ as a random sample drawn from the population of observations $\{(c_i, t_i, Z_i), i = 1, 2, \dots\}$. It follows that $\frac{1}{n_t} \sum_{k \in \mathcal{T}} Z_k$ converges in probability to the population mean $\bar{Z}$ by the weak law of large numbers. Because Z_i/n_t converges to zero, $\frac{1}{n_t - 1} \sum_{k \in \mathcal{T} \setminus i} Z_k$ also converges in probability to $\bar{Z}$. We can use identical arguments to show that $\sum_{k \in \mathcal{T} \setminus i} Y_k/(n_t - 1)$ converges in probability to $\bar{t}$, $\sum_{k \in \mathcal{T} \setminus i} Z_k^2/(n_t - 1)$ to $\overline{Z^2}$, and $\sum_{k \in \mathcal{T} \setminus i} Y_k^2/(n_t - 1)$ to $\overline{t^2}$.

We can write the denominator as

$$\left(\frac{1}{n_t-1}\sum_{k\in\mathcal{T}\setminus i} Z_k^2\right) - \left(\frac{1}{n_t-1}\sum_{k\in\mathcal{T}\setminus i} Z_k\right)^2 \xrightarrow{p} \overline{Z^2} - \bar{Z}^2$$

For the numerator, let $\overline{tZ}$ be the limit of $\sum_{i=1}^{N} t_i Z_i/N$. We can apply the Cauchy-Schwarz inequality to show this limit exists. Using the same argument as above, we have $\sum_{k\in\mathcal{T}\setminus i} Z_k Y_k/(n_t-1)$ converges in probability to $\overline{tZ}$. We therefore have the numerator

$$\left(\frac{1}{n_t-1}\sum_{k\in\mathcal{T}\setminus i} Z_k Y_k\right) - \left(\frac{1}{n_t-1}\sum_{k\in\mathcal{T}\setminus i} Z_k\right)\left(\frac{1}{n_t-1}\sum_{k\in\mathcal{T}\setminus i} Y_k\right) \xrightarrow{p} \overline{tZ} - \bar{Z}\bar{t}.$$

Thus for any observation i, $\hat{\beta}_i^t$ converges with probability 1, and we denote the limit β^t. It follows that $\hat{\alpha}_i^t$ converges, and we denote the limit α^t.

We can use an identical argument to show that $\hat{\beta}_i^c$ and $\hat{\alpha}_i^c$ converge. We then have

$$\begin{aligned}
\hat{m}_i &= (1-p)\hat{t}_i + p\hat{c}_i \\
&= (1-p)\left(\hat{\alpha}_i^t + \hat{\beta}_i^t Z_i\right) + p\left(\hat{\alpha}_i^c + \hat{\beta}_i^c Z_i\right) \\
&\xrightarrow{p} (1-p)\left(\alpha^t + \beta^t Z_i\right) + p\left(\alpha^c + \beta^c Z_i\right)
\end{aligned}$$

Due to assumption 1, the family of random variables $\{\hat{m}_i^{(N)}, N \in \mathbb{N}\}$ is uniformly integrable by the crystal ball condition (for example, see Chapter 6.5.1 of Resnick (2003)). We therefore have

$$\begin{aligned}
\mathrm{E}(\hat{m}_i) &\longrightarrow \mathrm{E}\left\{(1-p)\left(\alpha^t + \beta^t Z_i\right) + p\left(\alpha^c + \hat{\beta}^c Z_i\right)\right\} \\
&= (1-p)\left(\alpha^t + \beta^t Z_i\right) + p\left(\alpha^c + \beta^c Z_i\right).
\end{aligned}$$

Next, we consider

$$\frac{1}{N}\sum_{i=1}^{N}\left(m_{0i}^{(N)} - m_{\infty i}\right)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left((1-p)\left\{(\hat{\alpha}_i^t - \alpha^t) + (\hat{\beta}_i^t - \beta^t)Z_i\right\} + p\left\{(\hat{\alpha}_i^c - \alpha^c) + (\hat{\beta}_i^c - \beta^c)Z_i\right\}\right)^2 \quad \text{(N.3)}$$

After expanding the square, we can show that average of each term converges to zero. For example, let $\hat{\alpha}^t$ be the fitted intercept from regressing Y onto Z for all of the treated units, and $\hat{\beta}^c$ be the fitted slope from regressing Y onto Z for all of the control units. Let $R_i^{t,\alpha}$ be the adjustment term such that $\hat{\alpha}_{-i}^t = \hat{\alpha}^t - R_i^{t,\alpha}$, as calculated using equation (N.1). Similarly use equation (N.2) to calculate $R_i^{c,\beta}$. Fix $0 < M < \infty$ such that $M \geq |Z_i|$ for all i. Then we have

$$\frac{1}{N}\sum_{i=1}^{N} p(1-p)(\hat{\alpha}_i^t - \alpha^t)(\hat{\beta}_i^c - \beta^c)Z_i$$
$$\leq Mp(1-p)\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i^t - \alpha^t)(\hat{\beta}_i^c - \beta^c)$$
$$= Mp(1-p)\frac{1}{N}\left\{\sum_{i\in\mathcal{T}}(\hat{\alpha}^t - R_i^{t,\alpha} - \alpha^t)(\hat{\beta}^c - \beta^c) + \sum_{i\in\mathcal{C}}(\hat{\alpha}^t - \alpha^t)(\hat{\beta}^c - R_i^{c,\beta} - \beta^c)\right\}$$
$$= Mp(1-p)\frac{1}{N}\left\{\sum_{i=1}^{N}(\hat{\alpha}^t - \alpha^t)(\hat{\beta}^c - \beta^c) - \sum_{i\in\mathcal{T}} R_i^{t,\alpha} - \sum_{i\in\mathcal{C}} R_i^{c,\beta}\right\}$$
$$= Mp(1-p)(\hat{\alpha}^t - \alpha^t)(\hat{\beta}^c - \beta^c) - Mp(1-p)\frac{1}{N}\sum_{i\in\mathcal{T}} R_i^{t,\alpha} - Mp(1-p)\frac{1}{N}\sum_{i\in\mathcal{C}} R_i^{c,\beta}.$$

The first term converges to zero as $\hat{\alpha}^t \longrightarrow \alpha^t$ and $\hat{\beta}^c \longrightarrow \beta^c$. As noted earlier, the adjustment terms $R_i^{t,\alpha}$ and $R_i^{c,\beta}$ are on the order of $1/n_t$ and $1/n_c$. It follows that the second and third terms also converge to zero. Thus $\frac{1}{N}\sum_{i=1}^{N} p(1-p)(\hat{\alpha}_i^t - \alpha^t)(\hat{\beta}_i^c - \beta^c)Z_i$ converges to zero. Similar arguments can be used to show that the remaining terms of equation (N.3) go to zero as well.

Assumption (4): *There exists* $0 < K < \infty$ *such that*

$$\frac{\sum_{i=1}^{N}(m_i - m_{\infty i})^2}{N} \longrightarrow K,$$

and

$$\underset{i=1,\ldots,N}{max} \frac{(m_i - m_{\infty i})^2}{\sum_{k=1}^{N}(m_i - m_{\infty i})^2} \longrightarrow 0.$$

Let $A_i = (1-p)(t_i - (\alpha^t + \beta^t Z_i))$ and $B_i = p(c_i - (\alpha^c + \beta^c Z_i))$. We have

$$\begin{aligned}
\frac{\sum_{i=1}^{N}(m_i - m_{\infty i})^2}{N} &= \frac{\sum_{i=1}^{N}(A_i + B_i)^2}{N} \\
&\leq \frac{\sum_{i=1}^{N}(A_i^2 + B_i^2)}{N} + \frac{2}{N}\sqrt{\sum_{i=1}^{N} A_i^2 \sum_{i=1}^{N} B_i^2} \\
&= \frac{\sum_{i=1}^{N}(A_i^2 + B_i^2)}{N} + 2\sqrt{\frac{1}{N}\sum_{i=1}^{N} A_i^2 \frac{1}{N}\sum_{i=1}^{N} B_i^2}.
\end{aligned}$$

The term $\sum_{i=1}^{N} A_i^2/N$ converges:

$$\sum_{i=1}^{N} A_i^2/N = \frac{1}{N}\sum_{i=1}^{N} t_i^2 - \frac{1}{N}\sum_{i=1}^{N} 2t_i(\alpha^t + \beta^t Z_i) + \frac{1}{N}\sum_{i=1}^{N}(\alpha^t + \beta^t Z_i)^2$$

These three values converge due to the convergence of $\sum_{i=1}^{N} t_i/N$, $\sum_{i=1}^{N} Z_i/N$, $\sum_{i=1}^{N} t_i^2/N$, $\sum_{i=1}^{N} Z_i^2/N$, and $\sum_{i=1}^{N} t_i Z_i/N$. In addition, the limit of $\sum_{i=1}^{N} A_i^2/N$ is zero only if t_i and Z_i are perfectly correlated. We can use a similar argument to show that $\sum_{i=1}^{N} B_i^2/N$ converges. It follows there exists some $0 < K < \infty$ such that

$$\frac{\sum_{i=1}^{N}(m_i - m_{\infty i})^2}{N} \longrightarrow K.$$

Next, we have

$$\max_{i=1,\ldots,N} \frac{(m_i - m_{\infty i})^2}{\sum_{k=1}^{N}(m_i - m_{\infty i})^2} = \frac{\frac{1}{N}\max_{i=1,\ldots,N}(m_i - m_{\infty i})^2}{\sum_{k=1}^{N}(m_i - m_{\infty i})^2/N}.$$

The denominator converges to K. The quantity $(m_i - m_{\infty i})^2$ is bounded as t_i, c_i, and Z_i are bounded. It follows that the numerator converges to zero and so

$$\max_{i=1,\ldots,N} \frac{(m_i - m_{\infty i})^2}{\sum_{k=1}^{N}(m_k - m_{\infty k})^2} \longrightarrow 0.$$

Assumption (5): *There exists* $0 < \epsilon < 0.5$ *such that* $\epsilon < p_i < 1 - \epsilon$ *for all* i.

This assumption is met due to the constant treatment assignment probability.

APPENDIX O

Comparison Procedure for the Tournament Classifier

In this section, we provide details for the comparison method introduced in Section 5.3.

Transformation Recall that after we leave out the i-th observation, we transform each predictor for the remaning observations such that the mean of each transformed predictor will be 1 for the observations in $\mathcal{S}\backslash i$, and -1 for observations in $\mathcal{S}^c\backslash i$. In addition, for a predictor that predicts the class label Y well, we would expect the values of the transformed predictor for observations in $\mathcal{S}\backslash i$ to be closely clustered around 1, and the values for observations in $\mathcal{S}^c\backslash i$ to be closely clustered around -1. For predictors that do not predict the class label well, the transformed predictor will fluctuate more.

For a predictor X_j, we leave out the i-th observation and then we use the following procedure:

1. Let $m_{ij}^{(1)}$ be the mean of X_j for the observations in the set $\mathcal{S}\backslash i$ and $m_{ij}^{(-1)}$ be the mean of X_j for the observations in class $\mathcal{S}^c\backslash i$
2. Set $c_{ij} = 0.5 \times (m_{ij}^{(1)} + m_{ij}^{(-1)})$
3. Define $X_{kj}^{(i)} = X_{kj} - c_{ij}$ for $k \neq i$
4. Let $s_{ij} = \frac{1}{|\mathcal{S}\backslash i|} \sum_{k \in \mathcal{S}\backslash i} X_{kj}^{(i)}$

5. Set $Z_{ij} = \frac{X_{ij} - c_{ij}}{s_{ij}}$

We repeat this procedure for all i to obtain the transformed predictor $Z_j = (Z_1, \ldots, Z_n)$.

Comparing the Transformed predictors We interpolate between each pair of transformed predictors to form a single predictor. For a pair Z_1 and Z_2, we wish to obtain weights $w_{i,\{1,2\}}$ that minimize the distance between Y_i and the interpolation $Z_{i,\{1,2\}} = w_i Z_{i1} + (1 - w_i) Z_{i2}$ for $i = 1, \ldots, n$. To do this we set

$$w_{i,\{1,2\}} = \underset{x \in [0,1]}{\operatorname{argmin}} \sum_{k \neq i} \left\{ Y_k - (x Z_{k1} + (1 - x) Z_{k2}) \right\}^2 .$$

Taking the derivative with respect to x and setting equal to 0, we have

$$w_i = \frac{\sum_{k \neq i} (Y_k - Z_{k2})(Z_{k1} - Z_{k2})}{\sum_{k \neq i} (Z_{k1} - Z_{k2})^2},$$

which we then restrict to be in the interval $[0, 1]$.

APPENDIX P

Diagnostic Plots for Section 5.4.2

We display diagnostic plots related to our analysis of the microarray data sets. In Figure P.1, we show scree plots (for each data set) of the proportion of variance explained by each of the first 10 principal components. In Figure P.2, we show the values of the top 10 coefficients for the fitted tournament classifier model. We observe no discernible patterns to explain performance discrepancies between lasso and tournament classifier.

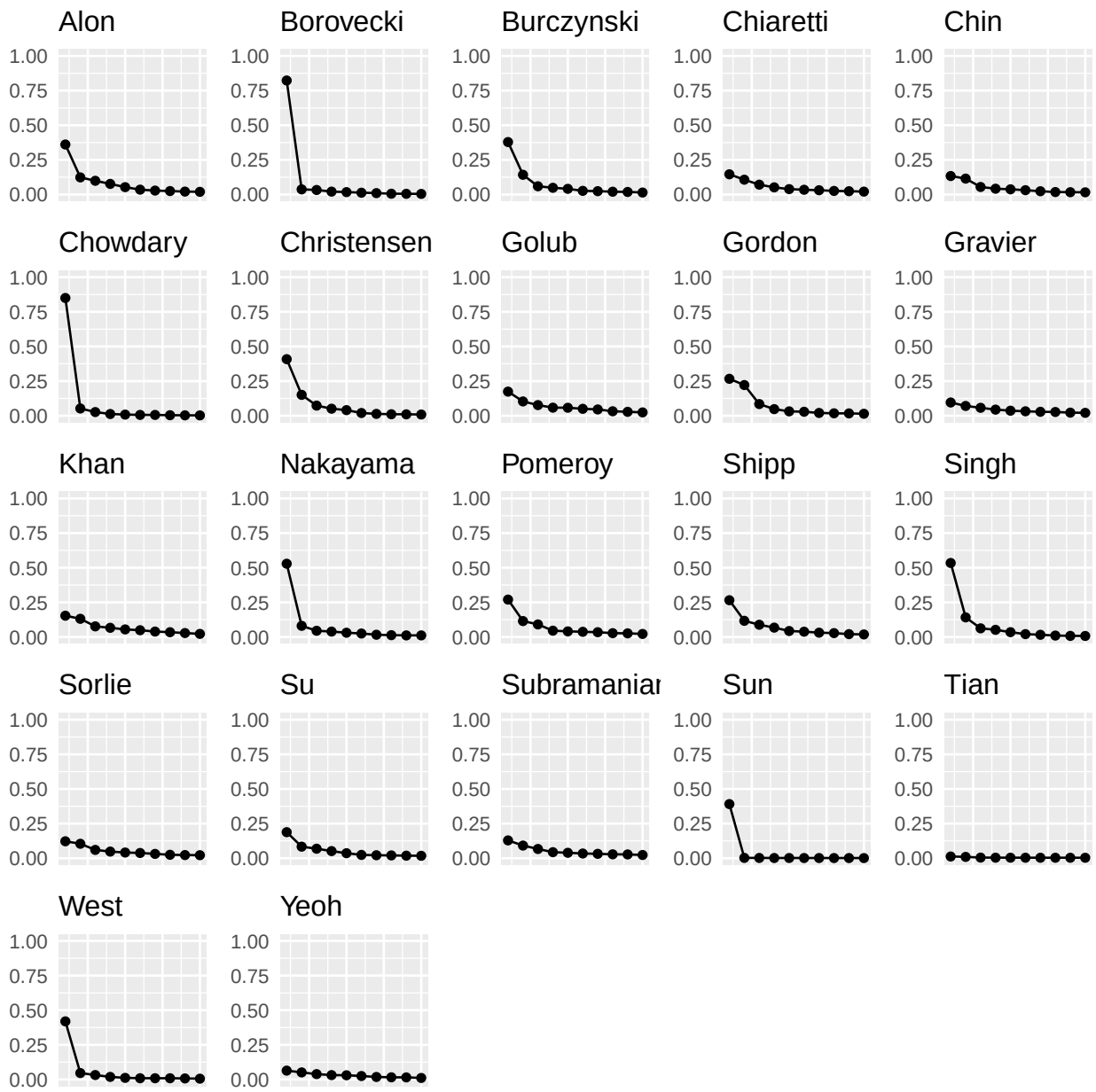

Figure P.1: Scree plots for the microarray data sets. In each plot, we plot the variance explained by each of the first 10 principal components for a given data set.

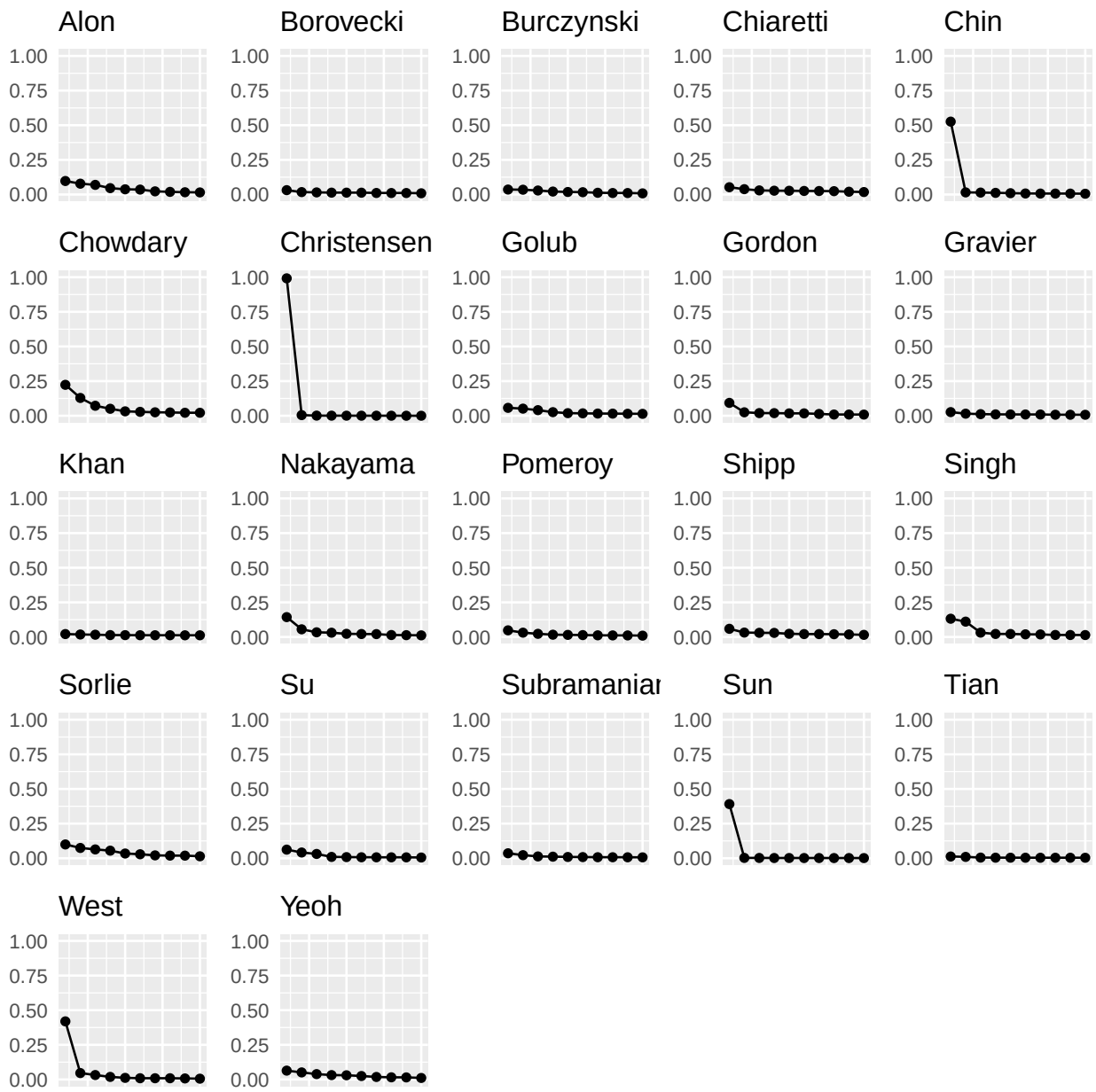

Figure P.2: Values of the top 10 tournament classifier coefficients for the microarray data sets. In each plot, we plot the top 10 coefficient values for a given data set.

BIBLIOGRAPHY

F. Abramovich and V. Grinshtein. High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079, 2018.

P. M. Aronow and J. A. Middleton. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135–154, 2013.

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

L. B. Balzer, M. L. Petersen, M. J. van der Laan, and the SEARCH Collaboration. Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Statistics in medicine*, 35(21):3717–3732, 2016a.

L. B. Balzer, M. J. van der Laan, M. L. Petersen, and the SEARCH Collaboration. Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in medicine*, 35(25):4528–4545, 2016b.

F. Barrera-Osorio, M. Bertrand, L. L. Linden, and F. Perez-Calle. Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. *American Economic Journal: Applied Economics*, 3(2):167–195, 2011.

C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, and D. F. Stroup. Improving the quality of reporting of randomized controlled trials: the consort statement. *The Journal of the American Medical Association*, 276(8):637–639, 1996.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013a.

R. Berk, E. Pitkin, L. Brown, A. Buja, E. George, and L. Zhao. Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37(3-4):170–196, 2013b.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13 (Apr):1063–1095, 2012.

P. J. Bickel, E. Levina, et al. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

A. Bloniarz, H. Liu, C. Zhang, J. S. Sekhon, and B. Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

C. M. Cassel, C. E. Särndal, and J. H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615–620, 1976.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

P. Dixon. Should blocks be fixed or random? *Conference on Applied Statistics in Agriculture*, 2016.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL `https://doi.org/10.1214/aos/1176344552`.

R. Fisher. *The Design of Experiments*. The Design of Experiments. Oliver and Boyd, 1935. URL `https://books.google.com/books?id=-EsNAQAAIAAJ`.

C. B. Fogarty. Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4):994–1000, 2018.

D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

E. Hartman, R. Grieve, R. Ramsahai, and J. S. Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 757–778, 2015.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

N. T. Heffernan and C. L. Heffernan. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

D. Holt and T. M. F. Smith. Post stratification. *Journal of the Royal Statistical Society, Series A*, pages 33–46, 1979.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

K. Imai. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in medicine*, 27(24):4857–4873, 2008.

K. Imai, G. King, and C. Nall. Rejoinder: Matched pairs and the future of cluster-randomized experiments. *Statistical Science*, 24(1):65–72, 2009.

G. W. Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic literature*, 48(2):399–423, 2010.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

N. Kallus, A. M. Puli, and U. Shalit. Removing hidden confounding by experimental grounding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/566f0ea4f6c2e947f36795c8f58ba901-Paper.pdf`.

G. G. Koch, I. A. Amara, G. W. Davis, and D. B. Gillings. A review of some statistical methods for covariance analysis of categorical data. *Biometrics*, pages 563–595, 1982.

G. G. Koch, C. M. Tangen, J.-W. Jung, and I. A. Amara. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*, 17(15-16):1863–1892, 1998.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

X. Li and P. Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):241–268, 2020.

J. Lim, R. Walley, J. Yuan, J. Liu, A. Dabral, N. Best, A. Grieve, L. Hampson, J. Wolfram, P. Woodward, et al. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. *Therapeutic innovation & regulatory science*, 52(5):546–559, 2018.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

J. Lu. Covariate adjustment in randomization-based causal inference for 2^K factorial designs. *Statistics & Probability Letters*, 119:11–20, 2016.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23 (19):2937–2960, 2004.

L. W. Miratrix, J. S. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society, Series B*, 75(2):369–396, 2013.

K. L. Moore and M. J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in Medicine*, 28(1):39–64, 2009.

D. C. Mutz, R. Pemantle, and P. Pham. The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician*, pages 1–11, 2018.

X. Nie and S. Wager. Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*, 2017.

K. S. Ostrow, D. Selent, Y. Wang, E. G. Van Inwegen, N. T. Heffernan, and J. J. Williams. The assessment of learning infrastructure (ALI) the theory, practice, and scalability of automated assessment. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 279–288, 2016.

Y. Pan and J. A. Gagnon-Bartsch. Separating and reintegrating latent variables to improve classification of genomic data. *arXiv preprint arXiv:2012.11757*, 2020.

J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144, 2014.

N. E. Pashley and L. W. Miratrix. Insights on variance estimation for blocked and matched pairs designs. *arXiv preprint arXiv:1710.10342*, 2017.

J. C. Pinheiro and D. M. Bates. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

S. J. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*, 29(3):175–188, 1976.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Science & Business Media, 1999.

S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage, 2002.

S. I. Resnick. *A probability path*. Springer, 2003.

J. M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 1999:6–10, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–866, 1994.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

E. Rosenman, A. B. Owen, M. Baiocchi, and H. Banack. Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863*, 2018.

C. Rothe. Flexible covariate adjustments in randomized experiments. *Working Paper*, 2018.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

A. Sales, A. Botelho, T. Patikorn, and N. T. Heffernan. Using big data to sharpen design-based inference in A/B tests. In *Proceedings of the Eleventh International Conference on Educational Data Mining*, 2018a.

A. C. Sales, B. B. Hansen, and B. Rowan. Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics*, 43(1):3–31, 2018b.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Rejoinder. *Journal of the American Statistical Association*, 94(448):1135–1146, 1999.

P. Z. Schochet. Statistical theory for the" rct-yes" software: Design-based causal inference for rcts. ncee 2015-4011. *National Center for Education Evaluation and Regional Assistance*, 2015.

K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine*, 152(11):726–732, 2010.

E. Scornet, G. Biau, J.-P. Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

J. Shao, Y. Wang, X. Deng, S. Wang, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics*, 39(2):1241–1265, 2011.

S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

D. S. Small, T. R. Ten Have, and P. R. Rosenbaum. Randomization inference in a group–randomized trial of treatments for depression: covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association*, 103(481):271–279, 2008.

J. Spiess. Optimal estimation when researcher and social preferences are misaligned. Technical report, 2018. Job Market Paper.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4): 465–472, 1990.

J. A. Steingrimsson, D. F. Hanley, and M. Rosenblum. Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemporary Clinical Trials*, 54:18–24, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

N. T. Trendafilov and I. T. Jolliffe. Dalass: Variable selection in discriminant analysis via the lasso. *Computational Statistics & Data Analysis*, 51(8):3718–3736, 2007.

A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, 2008.

K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45): 12673–12678, 2016.

W. Williams. Generating unbiased ratio and regression estimators. *Biometrics*, 17(2):267–274, 1961.

D. M. Witten and R. Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

H. Woltman, A. Feldstain, J. C. MacKay, and M. Rocchi. An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, 8(1):52–69, 2012.

E. Wu and J. A. Gagnon-Bartsch. The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation Review*, 42(4):458–488, 2018.

Z. Wu, C. E. Frangakis, T. A. Louis, and D. O. Scharfstein. Estimation of treatment effects in matched-pair cluster randomized trials by calibrating covariate imbalance between clusters. *Biometrics*, 70(4):1014–1022, 2014.

J. Yuan, J. Liu, R. Zhu, Y. Lu, and U. Palm. Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. *Journal of biopharmaceutical statistics*, 29(3):558–573, 2019.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.