BALANCE OPTIMIZATION SUBSET SELECTION: A FRAMEWORK FOR CAUSAL INFERENCE WITH OBSERVATIONAL DATA

BY

JASON JAMES SAUPPE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Sheldon H. Jacobson, Chair
Professor Allen Holder, Rose-Hulman Institute of Technology
Professor Chandra S. Chekuri
Assistant Professor P. Brighten Godfrey

Abstract

Observational data are prevalent in many fields of research, and it is desirable to use this data to explore potential causal relationships. Additional assumptions and methods for post-processing the data are needed to construct unbiased estimators of causal effects because such data is non-random. This dissertation describes the Balance Optimization Subset Selection (BOSS) framework to apply causal inference to observational data.

BOSS is designed to identify the subset of observational data that is most appropriate for computing causal estimates. To do this, it compares the available treatment units to potential sets of control units on a set of confounding factors, called covariates, with the goal of identifying a control group that minimizes a measure of covariate imbalance. Which imbalance measure to use with BOSS is an important consideration that depends both on the quality of the available observational data and on the assumptions that a researcher is willing to make.

The standard assumption for observational data, known as strong ignorability, is extended in several ways to be directly applicable to BOSS. Under these additional assumptions, specific levels of covariate balance are both necessary and sufficient for the treatment effect estimate to be unbiased. There is a trade-off in that weaker assumptions require a higher level of covariate balance in order to guarantee estimator unbiasedness. These additional assumptions bridge the gap between existing parametric and non-parametric methods.

Each imbalance measure for BOSS leads to an associated optimization problem. The computational complexity of these problems is discussed, and efficient algorithms are developed to handle several special cases. A constant factor approximation algorithm is also presented for one imbalance measure.

Given the potential applications of BOSS, identifying optimal or near-optimal solutions for these problems is of great practical interest. Heuristics and exact algorithms are considered, and computational tests demonstrate their effectiveness at minimizing imbalance. Additional tests validate BOSS on a well-studied dataset from the literature and highlight the value of alternate optima as a way to corroborate the assumptions that are made.

For my family

Acknowledgments

First, I would like to express my gratitude to my family: my wife Allie, whose love and support were instrumental in helping me get through graduate school; my parents Ted and Jan, for their love and guidance over the course of my life and for providing a wonderful example for me to follow; my siblings Josh and Stacy, for sharing in the long journey through their own graduate programs and for helping me maintain a good sense of humor throughout; my in-laws Doug, Louise, Lizzy, and John for welcoming another computer science graduate student into their family; and my many relatives scattered throughout the Midwest and beyond for their support and encouragement, particularly over the last six years.

I would also like to thank my academic mentors: my advisor Dr. Sheldon H. Jacobson, for providing me with advice and guidance during my graduate studies, the time and freedom I needed to explore ideas at my own pace, and several interesting problems on which to work, as well as for his flexibility in my work location which allowed me to spend the second half of my graduate career co-located with my wife; my undergraduate thesis advisor Dr. David Rader, for introducing me to operations research and helping me begin my academic career; Dr. Edward Sewell, for his support and advice on this project and several others during my time in graduate school; and my committee members Dr. Allen Holder, Dr. Chandra Chekuri, and Dr. Brighten Godfrey, for their time and input on this research, which greatly improved its quality and has also produced many directions for future work.

Finally, I would like to thank several groups of people for their support and camaraderie over the last six years: my academic siblings Alex, Arash, Banafsheh, David, Doug, Golsheed, Laura, and J. D.; my academic siblings-in-law Chien-Ming, Dan and Dani, Faisal, Irene and Ian, Margaret, Sean, Steve, Tomislav, and Zhi; my long-time friends from home Alex and Laura, Mike and Kay, Nick and Ali, and Rob; and my college friends Berry, Chris and Katie, Drew, Evan and Katie, Kevin, and Matthew.

Nothing would be done at all, if a man waited till he could do it so well that no one could find fault with it.

— Blessed John Henry Cardinal Newman

Deo omnis gloria.

This research has been supported in part by the Air Force Office of Scientific Research (FA9550-10-1-0387), the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (32 CFR 168a), the National Science Foundation (SES-0849223, SES-0849170), and the National Science Foundation Graduate Research Fellowship Program (DGE-1144245). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, the United States Air Force, the Department of Defense, or the United States Government.

Table of Contents

List of Notations

U	Universe or population of units under study (assumed to be infinite)
$u \in U$	A single unit within the universe
$S \subset U$	A finite set of units from U
y_u^1, y_u^0	Treatment and control responses for a (non-random) unit $u \in U$
τ_u	$\equiv y_u^1 - y_u^0$, the treatment effect for a (non-random) unit $u \in U$
Y^1, Y^0	Random variables representing the treatment response and control response, respectively, of a randomly sampled unit in U
τ	$\equiv \mathbf{E}\left[Y^1 - Y^0\right]$, the average treatment effect (ATE) across all units in U
$\widetilde{\tau}$	An estimator for τ
z_u	Indicator for the treatment status of unit $u \in U$; $z_u = 1(0)$ indicates that u is treated (untreated)
Z	Random variable representing the treatment status of a randomly sampled unit in U
U^1, U^0	$\equiv \{u \in U : z_u = 1\}$ and $\{u \in U : z_u = 0\}$, respectively; the treatment and control subpopulations
τ^1	$\equiv \mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right]$, the average treatment effect for the treated (ATT)
$\widetilde{\tau}^1$	An estimator for τ^1
$\mathcal{P}$	$\equiv \{1, 2, \ldots, p\}$, the set of (labels or indices for) the observed covariates
$\mathbf{x}_u \in \mathbb{R}^p$	Vector of covariate values for a (non-random) unit $u \in U$
$\mathbf{X}$	Random vector representing the covariate values of a randomly sampled unit in U
$\mathbf{X}_u$	Random covariate vector with u serving as an index ($u \in S \subset U$)
x_{ui}, X_i, X_{ui}	The component of the vector corresponding to covariate $i \in \mathcal{P}$
$\mathcal{X}$	$\equiv \{\mathbf{x}_u : u \in U\}$, the set of all covariate vectors in U; equivalently, the support of $\mathbf{X}$
$T \subset U^1$	A finite set of treatment units from U^1
$C \subset U^0$	A finite set of control units from U^0
$C' \subseteq C$	A subset of control units from C

$(\mathbf{X}_t, Y_t^1, Y_t^0)$	Random variables representing the covariate vector and treatment and control response of a treatment unit t sampled from U^1
$(\mathbf{X}_c, Y_c^1, Y_c^0)$	Random variables representing the covariate vector and treatment and control response of a control unit c sampled from U^0
$\bar{\mathbf{X}}_T$	$\equiv \sum_{t\in T} \mathbf{X}_t/\lvert T\rvert$, a random variable representing the vector of average covariate values for units in T (similar definitions for $\bar{\mathbf{X}}_C$, $\bar{\mathbf{X}}_{C'}$)
$\bar{X}_{Ti}$	Component of $\bar{\mathbf{X}}_T$ corresponding to covariate i (similar definitions for $\bar{X}_{Ci}$, $\bar{X}_{C'i}$)
$\bar{Y}_T^1$, $\bar{Y}_T^0$	$\equiv \sum_{t\in T} Y_t^1/\lvert T\rvert$ and $\sum_{t\in T} Y_t^0/\lvert T\rvert$, respectively; random variables representing the average treatment response and average control response of units in T (similar definitions for $\bar{Y}_C^1$, $\bar{Y}_C^0$, $\bar{Y}_{C'}^1$, $\bar{Y}_{C'}^0$)
$\tau_{T,C}$	Sample average treatment effect (SATE), computed across sampled units in T and C
τ_T^1	Sample average treatment effect for the treated (SATT), computed across sampled units in T
$\widetilde{\tau}_T^1$	An estimator for τ_T^1
$\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$	A distance function for matching
$M \subseteq T \times C$	A set of matched treatment-control pairs
$\mathcal{I} : \mathbb{N}^{\mathcal{X}} \times \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^+$	An imbalance measure defined on pairs of multisets of covariate vectors
$D \subseteq \mathcal{P}$	A covariate cluster
$\mathbf{D} \subseteq 2^{\mathcal{P}}$	A set of covariate clusters
$\mathbf{P}^D$	The projection matrix associated with covariate cluster D
x	A (non-random) scalar
$\mathbf{x}$	A (non-random) vector
$\mathcal{X}_i(S)$	$\equiv \{x_{ui} : u \in S\}$, the finite set of values for covariate $i \in \mathcal{P}$ possessed by the units in $S \subset U$
$\mathcal{X}_D(S)$	$\equiv \left\{\mathbf{P}^D\mathbf{x} : \mathbf{x} \in \mathbb{R}^p,\ x_i \in \mathcal{X}_i(S)\ \forall\ i \in D\right\}$, the projection of all combinations of covariate values for units in $S \subset U$ onto the space of covariates in D
$\widehat{F}_i(S, x)$	$\equiv \lvert\{u \in S : x_{ui} \leq x\}\rvert / \lvert S\rvert$, the empirical distribution function (empirical CDF) for covariate $i \in \mathcal{P}$
$\widehat{F}_D(S, \mathbf{x})$	$\equiv \left\lvert\{u \in S : \mathbf{P}^D\mathbf{x}_u \leq \mathbf{P}^D\mathbf{x}\}\right\rvert / \lvert S\rvert$, the joint empirical distribution function for covariate cluster $D \subseteq \mathcal{P}$
n_i	Number of bins for covariate $i \in \mathcal{P}$
b_{ij}	The jth bin boundary for covariate $i \in \mathcal{P}$
N_i	$\equiv \{1, 2, \ldots, n_i\}$, the set of indices for histogram bins for covariate $i \in \mathcal{P}$
B_{ij}	Set of units in the jth histogram bin for covariate i; $j \in N_i$, $i \in \mathcal{P}$
$\eta_{ij}(S)$	$\equiv \lvert S \cap B_{ij}\rvert / \lvert S\rvert$, the proportion of units in S occupying bin j for covariate i

N_D Set of tuples of indices for histogram bins for covariate cluster $D = \{i_1, i_2, \ldots, i_k\}$; $N_D \equiv N_{i_1} \times N_{i_2} \times \ldots \times N_{i_k}$

B_{Dj} The set of units in the jth histogram bin for covariate cluster D; $j \in N_D$, $D \subseteq \mathcal{P}$

$\eta_{Dj}(S)$ $\equiv |S \cap B_{Dj}|/|S|$, the proportion of units in S occupying bin j for cluster D

$h^0 : \mathcal{X} \to \mathbb{R}$ Control response function that maps vectors of covariate values to control responses

ε_u^0 $\equiv y_u^0 - \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}_u\right]$, the control response error for a (non-random) unit $u \in U$

$\mathcal{E}^0$ Random variable representing the control response error of a randomly sampled unit in U

$\mathcal{E}_t^0$, $\mathcal{E}_c^0$ Random variables representing the control response errors for randomly sampled units t and c

$\mathcal{B}(T, C')$ $\equiv \sum_{t \in T} h^0(\mathbf{X}_t)/|T| - \sum_{c \in C'} h^0(\mathbf{X}_c)/|C'|$, the control response function bias

Chapter 1

Introduction

Researchers in many fields are interested in establishing causal relationships to answer questions such as whether a new drug is effective in treating cancer or if a job training workshop improves the long-term earnings of participants. Questions of this form involve *actions*, which are procedures that can potentially be applied or withheld to the units in a study (e.g., patients can be given a new drug or a placebo; individuals may or may not participate in a training workshop), and *outcomes*, which are measurable responses that are exhibited by each of the units (e.g., quality of life, annual income). Actions are commonly referred to as *treatments*; the application of the action to a unit is indicated by stating that the unit is treated, receives treatment, or is exposed to treatment.

The process of determining if an action influences an outcome is known as *causal inference.* Causal inference is often applied by comparing the outcomes of units that were exposed to treatment with the outcomes for units that were not exposed. In some situations, a researcher can determine which units receive treatment and which do not. For example, in a *randomized controlled trial* (RCT), random assignments are used to select which units will receive treatment. Randomization is powerful if applying causal inference because it ensures that the measured outcomes are independent of confounding factors, and as such, randomized trials are viewed as the gold standard.

In many situations randomized trials are impractical for a number of reasons, such as cost or ethics. For example, it is unethical to intentionally expose units to a treatment that may potentially be harmful, such as smoking. However, it may be possible to observe units that were exposed for other reasons (e.g., by personal choice, by accident). By comparing exposed units with unexposed ones, a researcher may be able to provide evidence to support a causal relationship between treatment and outcome. Studies of this kind are called *observational* because they use data from observations rather than data generated from randomized experiments. With the rapid expansion of data collection in recent years, sources of observational data are burgeoning, resulting in numerous opportunities to apply causal inference.

Observational studies are prevalent in many fields. In medicine and health care, they have been used to understand the impacts of an extra chromosome on aggression (Witkin et al., 1976), compare generic drugs against brand name counterparts (Rubin, 1991), analyze how birth weight is affected by smoking during pregnancy (da Veiga and Wilder, 2008), determine the impact of hospital stay times on health outcomes for preterm infants (Zubizarreta et al., 2013b), compare differences in health outcomes after surgery with either general or regional anesthesia (Zubizarreta et al., 2012), and assess the impact of age and obesity on risk levels for kidney injury following surgery (Kelz et al., 2013; Reinke et al., 2012). Within the social sciences, observational studies have been used to explore the connections between the likelihood of revolution in a country in the presence or absence of foreign threats (Sekhon, 2004), assess the impact of "get-out-the-vote" calls on voter turnout (Imai, 2005), refute allegations of voter fraud for new election technologies (Herron and Wand, 2007), and establish a relationship between earthquake severity and post-traumatic stress symptoms (Zubizarreta et al., 2013a). There is increasing interest in using observational studies within comparative effectiveness research in order to determine the most appropriate treatment options for different groups of patients (Concato et al., 2010; Marko and Weil, 2010).

The lack of randomization in observational studies means that treatment units might differ from the untreated units in some way (e.g., treatment units may be older, on average, than control units). If these differences influence the response to treatment, then it becomes difficult to determine whether the observed outcomes are a result of the treatment or are caused by confounding factors, called *covariates*. Distortion in the estimated treatment effect due to differences in the underlying treatment and control populations is called *selection bias*. Controlling or limiting this source of bias is one of the primary challenges in observational studies (Cochran and Rubin, 1973).

This dissertation proposes the Balance Optimization Subset Selection (BOSS) framework to identify causal relationships in observational data. BOSS addresses the above difficulties by searching for the subset of control units that is most similar to the set of treatment units on the observed covariates. Subset similarity is typically assessed through an imbalance measure that compares the values of the covariates between two groups, or sets of units, and identifies differences between them. Two groups are similar if there is no imbalance across their covariates. There are numerous ways to measure covariate balance, and the BOSS framework is flexible with regards to the choice of imbalance measure. This allows researchers and practitioners to select the imbalance measure that is most appropriate for the data at hand.

BOSS provides a number of benefits compared to existing methods. First, covariate imbalance is a major source of bias in observational studies. By seeking to minimize covariate imbalance directly, BOSS ensures that any measured difference in outcomes between the treatment units and the control units is attributable solely to the treatment itself (assuming that no unobserved covariates are present). BOSS can also provide a guarantee on whether or not it is possible to remove all covariate imbalance. If the best set of control units features significant imbalance with respect to the treatment group, then this serves as an indicator that the researcher should focus his or her efforts on acquiring more data instead of continuing to search in vain for a solution in the current dataset.

Another benefit of BOSS is that it can produce multiple estimates of the treatment effect by identifying alternate optima. A comparison of these estimates might strengthen the evidence of a causal relationship or call such a relationship into question. For example, if two sets of control units are both equally well-balanced with respect to the set of treatment units but produce significantly different estimates of the treatment effect, then a researcher should question and scrutinize his or her assumptions. On the other hand, if the estimates are similar, then the researcher gains confidence in asserting a causal relationship.

The remainder of this dissertation is organized as follows. Chapter 2 provides some background on causal inference in observational studies and common methods, such as matching and regression, that have been proposed to address selection bias. An overview of the BOSS framework along with a discussion of several potential imbalance measures is also included. Chapter 3 develops the necessary statistical assumptions for BOSS. The theory builds on well-known assumptions from the causal inference literature in order to highlight the requirements that must be met to use BOSS. Chapter 4 looks at BOSS from a complexity perspective and explores the difficulty of solving the associated optimization problems for various imbalance measures. In many cases, BOSS is shown to be **NP-Hard**, though there are a few exceptions. Several approximation results are noted. Chapter 5 presents several methods for solving BOSS. These include heuristics and exact algorithms based on integer programming models. A set of computational tests demonstrate the effectiveness of BOSS with several different imbalance measures on both simulated and real datasets. A comparison to matching methods is included. Chapter 6 provides concluding remarks and highlights several directions for future research.

Chapter 2

Background

This chapter reviews the Rubin Causal Model as well as the standard assumptions for observational data. An overview of relevant methods for estimation, including regression and matching, is provided, along with a review of the Balance Optimization Subset Selection framework.

2.1 Rubin Causal Model

The process of applying causal inference can be formalized using the Rubin Causal Model (Rubin, 1974; Holland, 1986). The model considers an infinite population or universe U of units under study and is concerned with how the presence or absence of an action or treatment influences an outcome. Each $u \in U$ has two potential *responses*, or values for the outcome of interest: a *treatment response* y_u^1 that is exhibited if u receives the treatment, and a *control response* y_u^0 that is exhibited if u does not receive the treatment. For each $u \in U$, the effect of the treatment relative to the control response is $\tau_u \equiv y_u^1 - y_u^0$. The Rubin Causal Model assumes that the potential responses of one unit are unaffected by the treatment status (treated or untreated) of other units. This is known as the *Stable Unit Treatment Value Assumption* (SUTVA) (Rubin, 1978; Sekhon, 2009).

The fundamental problem of causal inference is that it is impossible to observe both y_u^1 and y_u^0 for any single unit $u \in U$. If u is exposed to the treatment, then y_u^1 is exhibited, otherwise y_u^0 is exhibited. In either case, there is no way to observe both responses for any given unit (under identical conditions). Hence, it is impossible to determine τ_u for any $u \in U$ (Holland, 1986).

The statistical solution to the fundamental problem of causal inference is to shift attention from the "impossible-to-observe" unit-level treatment effects to the "possible-to-estimate" *average treatment effect* (ATE) for the population (Holland, 1986). This is accomplished by introducing the random variables Y^1 and Y^0, which are the treatment and control response, respectively, of a unit selected at random (uniformly)

Some of the material in this chapter has been adapted from Sauppe et al. (2014), *INFORMS Journal on Computing* **26**(3), with the permission of the copyright holder.

from U. Then the average treatment effect is defined as

$$\tau \equiv \mathbf{E}\left[Y^1 - Y^0\right] = \mathbf{E}\left[Y^1\right] - \mathbf{E}\left[Y^0\right]. \tag{2.1}$$

The average treatment effect, τ, is a parameter of the population U, and as such can be estimated. While it is impossible to estimate τ directly from unit-level treatment effects τ_u for individual units because these values are unknown, (2.1) reveals that it is possible to estimate τ by estimating both the *average treatment response*, $\mathbf{E}\left[Y^1\right]$, and the *average control response*, $\mathbf{E}\left[Y^0\right]$, of the population. Most importantly, these two estimates need not come from the same units; instead, they can be constructed using units for which the appropriate potential response is observed.

As an example, a first attempt at estimating τ might sample a treated unit and an untreated unit and compare the observed responses of the two units. Let z_u indicate the treatment status for each $u \in U$, with $z_u = 1$ indicating that the unit is treated and $z_u = 0$ indicating that the unit is untreated. Additionally, let the random variable Z be the treatment status of a unit selected at random (uniformly) from U. The *treatment* and *control subpopulations* are $U^1 \equiv \{u \in U : z_u = 1\}$ and $U^0 \equiv \{u \in U : z_u = 0\}$, respectively, with $U = U^1 \cup U^0$. Units in U^1 are referred to as *treatment units* and denoted by t, while units in U^0 are referred to as *control units* and denoted by c.

An estimator for τ can be calculated using the following procedure. Select a unit $t \in U^1$ at random (uniformly) and let the random variable Y_t^1 be the sampled unit's treatment response. Then select a unit $c \in U^0$ at random (uniformly) and let the random variable Y_c^0 be the sampled unit's control response. Finally, construct the estimator for τ as $\widetilde{\tau} \equiv Y_t^1 - Y_c^0$.

The properties of $\widetilde{\tau}$ provide an indication as to its suitability for estimating τ. One such property is estimator bias, computed as $\mathbf{E}\left[\widetilde{\tau}\right] - \tau$, where the expectation is taken over all possible pairs of treatment and control units. The bias of $\widetilde{\tau}$ is computed by first observing that the selection procedures for t and c ensure that

$$\mathbf{Pr}\left(Y_t^1 \leq y\right) = \mathbf{Pr}\left(Y^1 \leq y \mid Z = 1\right) \quad \forall\, y \in \mathbb{R}, \tag{2.2a}$$

$$\mathbf{Pr}\left(Y_c^0 \leq y\right) = \mathbf{Pr}\left(Y^0 \leq y \mid Z = 0\right) \quad \forall\, y \in \mathbb{R}. \tag{2.2b}$$

In words, (2.2) states that Y_t^1 has the same distribution as Y^1 in the treatment subpopulation and Y_c^0 has the same distribution as Y^0 in the control subpopulation. Using (2.1) and (2.2), the bias of $\widetilde{\tau}$ is

$$\begin{aligned}
\mathbf{E}\left[\widetilde{\tau}\right] - \tau &= \mathbf{E}\left[Y_t^1 - Y_c^0\right] - \tau \\
&= \left(\mathbf{E}\left[Y_t^1\right] - \mathbf{E}\left[Y_c^0\right]\right) - \left(\mathbf{E}\left[Y^1 - Y^0\right]\right) \\
&= \left(\mathbf{E}\left[Y^1 \mid Z = 1\right] - \mathbf{E}\left[Y^0 \mid Z = 0\right]\right) - \left(\mathbf{E}\left[Y^1\right] - \mathbf{E}\left[Y^0\right]\right).
\end{aligned}$$

Thus, $\widetilde{\tau}$ is not guaranteed to be unbiased.

The difficulty with the above procedure is that Y_t^1 provides an estimate of the average treatment response for the treatment subpopulation, $\mathbf{E}\left[Y^1 \mid Z = 1\right]$, not an estimate of the average treatment response for the entire population, $\mathbf{E}\left[Y^1\right]$. In the absence of any additional assumptions, these two quantities need not be the same. (A similar argument applies for $\mathbf{E}\left[Y^0 \mid Z = 0\right]$ and $\mathbf{E}\left[Y^0\right]$.) In order to address this difficulty, some additional assumptions are necessary.

2.1.1 Experimental Setting

One way to resolve the potential discrepancies between $\mathbf{E}\left[Y^1 \mid Z = 1\right]$ and $\mathbf{E}\left[Y^1\right]$ and $\mathbf{E}\left[Y^0 \mid Z = 0\right]$ and $\mathbf{E}\left[Y^0\right]$ is to make an additional assumption about how the subpopulations U^1 and U^0 were formed. This can be done using the concept of the *assignment mechanism*, which is the process by which units from U are selected for exposure to treatment.

One particular assignment mechanism is *random assignment*, where each unit in the population has the same probability of receiving treatment. Random assignment ensures that U^1 and U^0 form a random partition of U. Because U is infinite, it follows that the distributions of treatment and control responses in these subpopulations are identical to the distributions of treatment and control responses in the overall population U. Specifically,

$$\mathbf{Pr}\left(Y^1 \leq y \mid Z = z\right) = \mathbf{Pr}\left(Y^1 \leq y\right) \quad \forall\, y \in \mathbb{R},\ z \in \{0,1\}, \tag{2.3a}$$

$$\mathbf{Pr}\left(Y^0 \leq y \mid Z = z\right) = \mathbf{Pr}\left(Y^0 \leq y\right) \quad \forall\, y \in \mathbb{R},\ z \in \{0,1\}. \tag{2.3b}$$

From (2.3), it follows that

$$\begin{aligned}
\mathbf{E}\left[Y^1 \mid Z = 1\right] = \mathbf{E}\left[Y^1\right] = \mathbf{E}\left[Y^1 \mid Z = 0\right], \\
\mathbf{E}\left[Y^0 \mid Z = 1\right] = \mathbf{E}\left[Y^0\right] = \mathbf{E}\left[Y^1 \mid Z = 0\right],
\end{aligned}$$

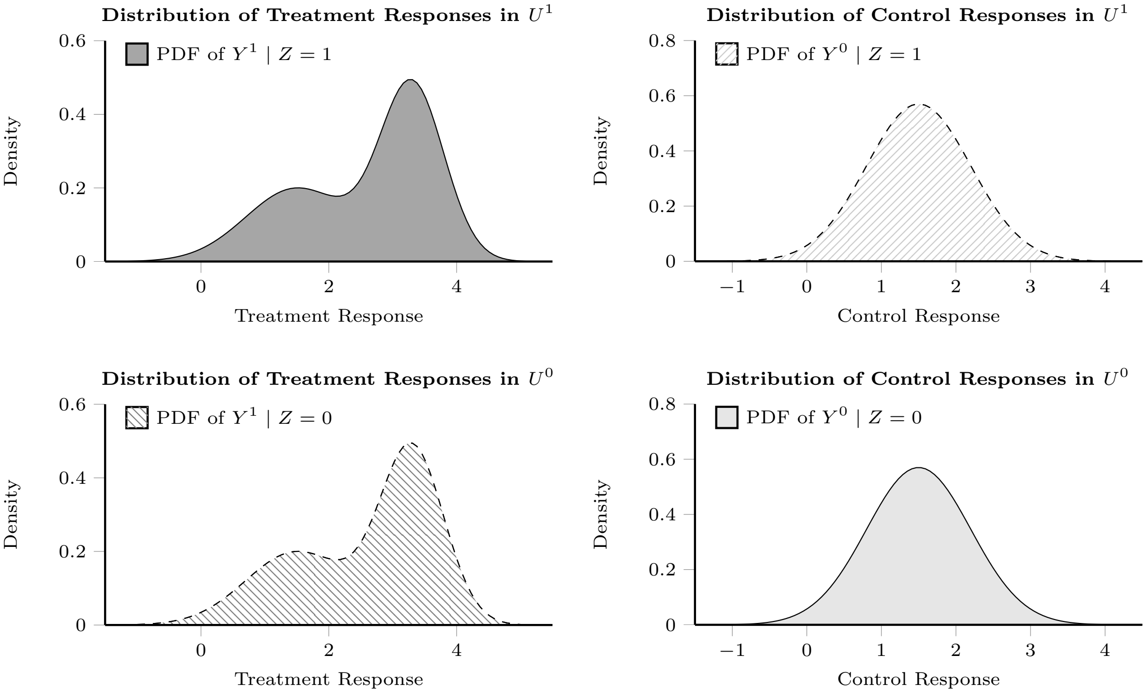


Figure 2.1: Possible distributions of treatment and control responses in the treatment and control subpopulations under the independence assumption.

and thus, the estimator $\widetilde{\tau}$ is unbiased. The properties in (2.3) are often stated in the form of the *Independence* assumption.

Assumption 2.1 (Independence)**.** The random variables Y^1 and Y^0 are independent of the treatment status variable Z: $\left(Y^1, Y^0\right) \perp Z$.

Figure 2.1 shows possible distributions for the treatment and control responses in both the treatment and control subpopulations under Assumption 2.1. The distributions for the treatment subpopulation are in the top row, while the distributions for the control subpopulation are in the bottom row. The solid plots in the figure represent distributions whose values can be sampled while the hatched plots represent distributions whose values cannot be sampled because they are unobservable.

The consequences of the independence assumption motivate the use of random assignment in *experimental studies*, where treatment assignment is not pre-determined. In this situation, a researcher can sample units from the entire population U and then decide which potential response to observe for each sampled unit. By using a random assignment mechanism, the researcher ensures that the resulting estimate of the treatment effect computed from the units' observed responses is unbiased. Experimental studies that use random assignment are referred to as *randomized controlled trials*. Assumption 2.1 allows a researcher to gain the benefits of randomization when treatment assignment is beyond his or her control.

2.1.2 Observational Setting

In many situations the assignment mechanism is unknown and almost certainly non-random, rendering Assumption 2.1 invalid. The data in these situations are classified as *observational* because they are observed as is and not generated through a known mechanism. To estimate τ from observational data, additional assumptions are needed. One assumption is an adaptation of Assumption 2.1. Instead of assuming that treatment assignment is random, the researcher assumes that treatment assignment is random *conditional* on a set of observed factors, called *covariates.* The covariates may influence the treatment and control responses of a unit or its likelihood of being exposed to treatment.

As an example, consider the outcomes from a new surgical procedure. It would be unreasonable to assume that the patients who received the procedure were selected completely at random. However, patients who received the new procedure were likely selected on the basis of existing covariates such as weight, age, gender, and prior medical history. So for two units with identical covariates, a plausible assumption is that each unit had the same probability of being selected for treatment, regardless of whether or not either unit actually received the treatment.

To formalize this assumption, let $\mathcal{P} \equiv \{1, 2, \ldots, p\}$ be the set of labels for the covariates that describe the units in U. For each $u \in U$, let $\mathbf{x}_u$ be a vector of covariate values, with x_{ui} representing unit u's value for covariate $i \in \mathcal{P}$. Let the random vector $\mathbf{X} \in \mathbb{R}^p$ be the covariate values of a unit selected at random (uniformly) from U, and let $\mathcal{X}$ denote the support of $\mathbf{X}$. Additionally, let X_i be the component of $\mathbf{X}$ corresponding to covariate $i \in \mathcal{P}$. The conditional independence assumption is known as the *Strong Ignorability* assumption (Rosenbaum and Rubin, 1983b).

Assumption 2.2 (Strong Ignorability)**.** The random variables Y^1, Y^0, $\mathbf{X}$, and Z satisfy the following:

(a) $\left(Y^1, Y^0\right) \perp\!\!\!\perp Z \mid \mathbf{X}$,

(b) $0 < \mathbf{Pr}\left(Z = 1 \mid \mathbf{X} = \mathbf{x}\right) < 1 \quad \forall\, \mathbf{x} \in \mathcal{X}$.

Assumption 2.2(a) states that Y^1 and Y^0 are conditionally independent of Z given $\mathbf{X}$, where $\perp\!\!\!\perp$ signifies conditional independence (Dawid, 1979). In other words, given $\mathbf{X}$, knowledge of Z provides no information about either Y^1 or Y^0, and vice versa. So for any fixed value of $\mathbf{x} \in \mathcal{X}$, the distribution of the treatment (control) responses for units in $\{t \in U^1 : \mathbf{x}_t = \mathbf{x}\}$ is identical to the distribution of treatment (control) responses for units in $\{c \in U^0 : \mathbf{x}_c = \mathbf{x}\}$. Assumption 2.2(b) states that there must be at least one control unit and at least one treatment unit at each possible value of $\mathbf{x} \in \mathcal{X}$. An immediate consequence of Assumption 2.2 is that

$$\mathbf{E}\left[Y^1 \mid \mathbf{X}=\mathbf{x},\ Z=1\right] = \mathbf{E}\left[Y^1 \mid \mathbf{X}=\mathbf{x}\right] = \mathbf{E}\left[Y^1 \mid \mathbf{X}=\mathbf{x},\ Z=0\right] \quad \forall\, \mathbf{x} \in \mathcal{X}, \tag{2.4a}$$

$$\mathbf{E}\left[Y^0 \mid \mathbf{X}=\mathbf{x},\ Z=1\right] = \mathbf{E}\left[Y^0 \mid \mathbf{X}=\mathbf{x}\right] = \mathbf{E}\left[Y^0 \mid \mathbf{X}=\mathbf{x},\ Z=0\right] \quad \forall\, \mathbf{x} \in \mathcal{X}. \tag{2.4b}$$

Figure 2.2 provides an example of what the the distributions of the random variables might look like under a non-random assignment mechanism. The plots in the first row show the distributions of $\mathbf{X}$, Y^1, and Y^0 within the entire population U, while the second and third rows show these distributions in U^1 and U^0, respectively. While the treatment and control subpopulations may have different unconditional distributions, Assumption 2.2 ensures that the distributions of Y^1 and Y^0 conditional on $\mathbf{X}$ are identical between U^1 and U^0; in particular, these conditional distributions should possess the symmetry shown in Figure 2.1.

If Assumption 2.2 is valid, then an unbiased estimator for τ can be calculated using the following procedure. Select a vector of covariate values $\mathbf{X}$ at random (uniformly) from the entire population U. Then select a treatment unit t at random (uniformly) from $\left\{t \in U^1 : \mathbf{x}_t = \mathbf{X}\right\}$ and let the random variable Y_t^1 be the sampled unit's treatment response. Next, select a control unit c at random (uniformly) from $\left\{c \in U^0 : \mathbf{x}_c = \mathbf{X}\right\}$ and let the random variable Y_c^0 be the sampled unit's control response. Finally, construct the estimator $\widetilde{\tau} \equiv Y_t^1 - Y_c^0$.

To see that $\widetilde{\tau}$ is unbiased, observe that the above selection procedure for t and c ensures that

$$\mathbf{Pr}\left(Y_t^1 \le y \mid \mathbf{X}=\mathbf{x}\right) = \mathbf{Pr}\left(Y^1 \le y \mid \mathbf{X}=\mathbf{x}, Z=1\right) \quad \forall\, y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}, \tag{2.5a}$$

$$\mathbf{Pr}\left(Y_c^0 \le y \mid \mathbf{X}=\mathbf{x}\right) = \mathbf{Pr}\left(Y^0 \le y \mid \mathbf{X}=\mathbf{x}, Z=0\right) \quad \forall\, y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}. \tag{2.5b}$$

Let $g(\mathbf{x}) \equiv \mathbf{E}\left[Y_t^1 - Y_c^0 \mid \mathbf{X}=\mathbf{x}\right]$. Then (2.4) and (2.5) ensure that

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{E}\left[Y_t^1 - Y_c^0 \mid \mathbf{X}=\mathbf{x}\right] \\ &= \mathbf{E}\left[Y_t^1 \mid \mathbf{X}=\mathbf{x}\right] - \mathbf{E}\left[Y_c^0 \mid \mathbf{X}=\mathbf{x}\right] \\ &= \mathbf{E}\left[Y^1 \mid \mathbf{X}=\mathbf{x},\ Z=1\right] - \mathbf{E}\left[Y^1 \mid \mathbf{X}=\mathbf{x},\ Z=0\right] \\ &= \mathbf{E}\left[Y^1 \mid \mathbf{X}=\mathbf{x}\right] - \mathbf{E}\left[Y^0 \mid \mathbf{X}=\mathbf{x}\right] \\ &= \mathbf{E}\left[Y^1 - Y^0 \mid \mathbf{X}=\mathbf{x}\right]. \end{aligned} \tag{2.6}$$

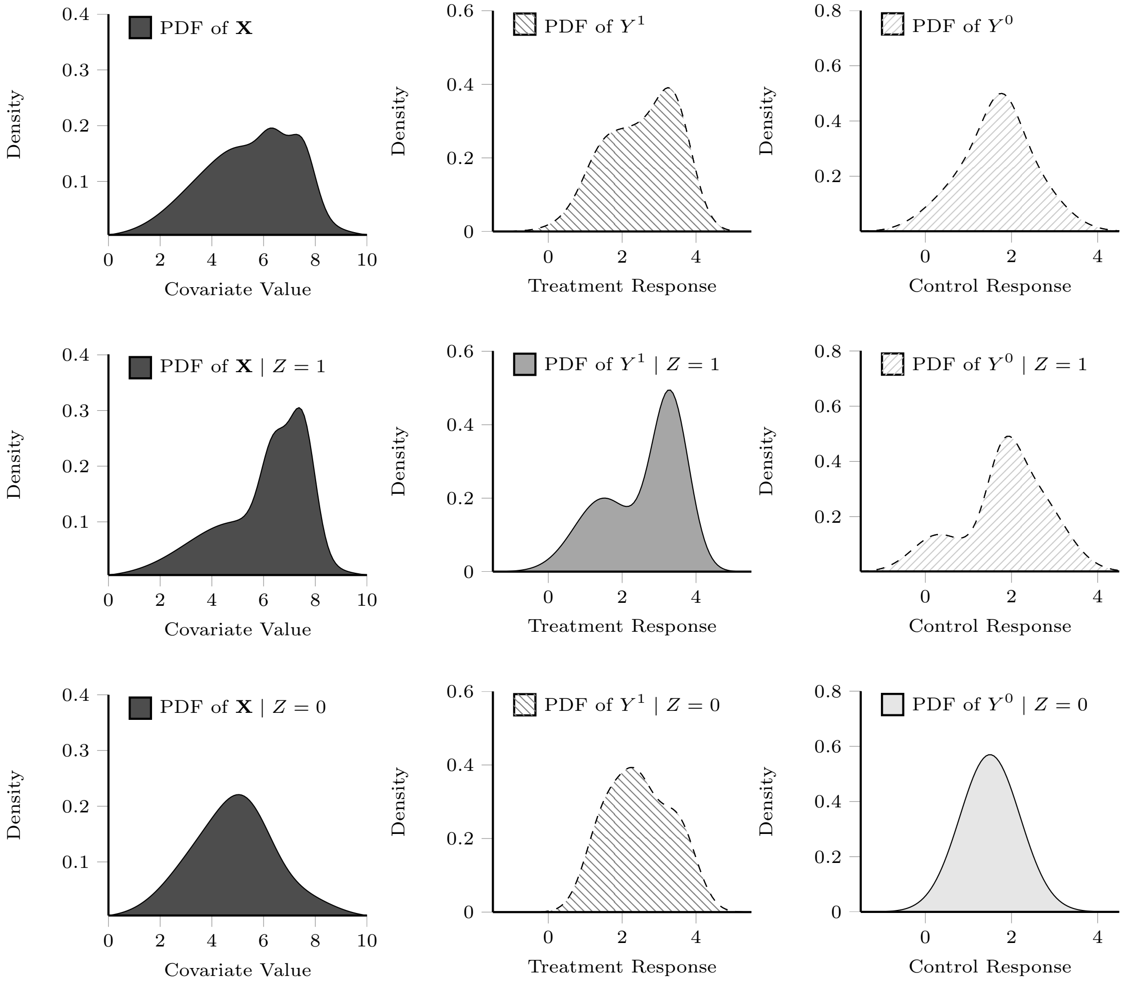


Figure 2.2: Possible distributions of covariate values and treatment and control responses for units in U, U^1, and U^0 with a non-random assignment mechanism.

The expected value of $\widetilde{\tau}$ is computed by conditioning on $\mathbf{X}$ and using (2.6):

$$\begin{aligned}\mathbf{E}\left[\widetilde{\tau}\right] &= \mathbf{E}_{\mathbf{X}}\left[\mathbf{E}\left[Y_t^1 - Y_c^0 \mid \mathbf{X}\right]\right]\\ &= \mathbf{E}_{\mathbf{X}}\left[g(\mathbf{X})\right]\\ &= \mathbf{E}_{\mathbf{X}}\left[\mathbf{E}\left[Y^1 - Y^0 \mid \mathbf{X}\right]\right]\\ &= \mathbf{E}\left[Y^1 - Y^0\right]\\ &= \tau.\end{aligned}$$

Hence, $\widetilde{\tau}$ is an unbiased estimator for τ.

There are many situations in which the treatment effect for one of the subpopulations rather than the population as a whole is of interest. For example, an estimate of the effectiveness of treatment for those units that were actually treated may be sought, as the treatment subpopulation represents the set of units most likely to be exposed to the treatment in the first place. In this case, the parameter of interest is the *average treatment effect for the treated* (ATT), defined as

$$\tau^1 \equiv \mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right] = \mathbf{E}\left[Y^1 \mid Z = 1\right] - \mathbf{E}\left[Y^0 \mid Z = 1\right].$$

When estimating τ^1, Assumption 2.2 can be relaxed.

Assumption 2.3 (Strong Ignorability for ATT)**.** The random variables Y^0, $\mathbf{X}$, and Z satisfy the following:

(a) $Y^0 \perp\!\!\!\perp Z \mid \mathbf{X}$,

(b) $\mathbf{Pr}\left(Z = 1 \mid \mathbf{X} = \mathbf{x}\right) < 1 \quad \forall\, \mathbf{x} \in \mathcal{X}$.

Assumption 2.3 asserts that only the control response Y^0 is conditionally independent of Z given $\mathbf{X}$. Similarly, it is only required that there be at least one control unit at every possible value of $\mathbf{x} \in \mathcal{X}$. An immediate consequence of Assumption 2.3 is that

$$\mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x},\ Z = 1\right] = \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}\right] = \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x},\ Z = 0\right], \tag{2.7}$$

for all $\mathbf{x} \in \mathcal{X}^1$, where $\mathcal{X}^1$ is the support of $\mathbf{X}$ within the treatment subpopulation.

When Assumption 2.3 is valid, an unbiased estimator for τ^1 can be constructed using the following procedure (Rosenbaum, 1987a). Select a treatment unit $t \in U^1$ at random (uniformly), and let the random variables $\mathbf{X}_t$ and Y_t^1 be the unit's covariate vector and treatment response, respectively. Then select a control unit from $\{c \in U^0 : \mathbf{x}_c = \mathbf{X}_t\}$ at random (uniformly), and let Y_c^0 be the selected control unit's control response. Such a control unit is guaranteed to exist by Assumption 2.3(b). Finally, construct the estimator for τ^1 as $\widetilde{\tau}^1 \equiv Y_t^1 - Y_c^0$.

To see that $\widetilde{\tau}^1$ is unbiased, observe that the above selection procedure ensures that

$$\mathbf{Pr}\left(Y_t^1 \leq y \mid \mathbf{X}_t = \mathbf{x}\right) = \mathbf{Pr}\left(Y^1 \leq y \mid \mathbf{X} = \mathbf{x},\ Z = 1\right) \quad \forall\, y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}^1, \tag{2.8a}$$

$$\mathbf{Pr}\left(Y_c^0 \leq y \mid \mathbf{X}_t = \mathbf{x}\right) = \mathbf{Pr}\left(Y^0 \leq y \mid \mathbf{X} = \mathbf{x},\ Z = 0\right) \quad \forall\, y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}^1. \tag{2.8b}$$

Let $g^1(\mathbf{x}) \equiv \mathbf{E}\left[Y_t^1 - Y_c^0 \mid \mathbf{X}_t = \mathbf{x}\right]$. Then (2.7) and (2.8) ensure that

$$\begin{aligned} g^1(\mathbf{x}) &= \mathbf{E}\left[Y_t^1 - Y_c^0 \mid \mathbf{X}_t = \mathbf{x}\right] \\ &= \mathbf{E}\left[Y_t^1 \mid \mathbf{X}_t = \mathbf{x}\right] - \mathbf{E}\left[Y_c^0 \mid \mathbf{X}_t = \mathbf{x}\right] \\ &= \mathbf{E}\left[Y^1 \mid \mathbf{X} = \mathbf{x},\ Z = 1\right] - \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x},\ Z = 0\right] \\ &= \mathbf{E}\left[Y^1 \mid \mathbf{X} = \mathbf{x},\ Z = 1\right] - \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x},\ Z = 1\right] \\ &= \mathbf{E}\left[Y^1 - Y^0 \mid \mathbf{X} = \mathbf{x},\ Z = 1\right]. \end{aligned} \tag{2.9}$$

The expected value of $\widetilde{\tau}^1$ is computed by conditioning on $\mathbf{X}_t$ and using (2.9):

$$\begin{aligned} \mathbf{E}\left[\widetilde{\tau}^1\right] &= \mathbf{E}_{\mathbf{X}_t}\left[\mathbf{E}\left[Y_t^1 - Y_c^0 \mid \mathbf{X}_t\right]\right] \\ &= \mathbf{E}_{\mathbf{X}_t}\left[g\left(\mathbf{X}_t\right)\right] \\ &= \mathbf{E}_{\mathbf{X}|Z=1}\left[g\left(\mathbf{X}\right)\right] \\ &= \mathbf{E}_{\mathbf{X}|Z=1}\left[\mathbf{E}\left[Y^1 - Y^0 \mid \mathbf{X}, Z = 1\right]\right] \\ &= \mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right] \\ &= \tau^1. \end{aligned}$$

Hence, $\widetilde{\tau}^1$ is an unbiased estimator for τ^1.

2.2 Observational Data in Practice

The processes described in Section 2.1.2 for estimating τ and τ^1 both assume that units with specific covariate values can be sampled from U^1 and U^0. In practice, however, a researcher is typically unable to sample units at will and instead only has access to two subsets $T \subset U^1$ and $C \subset U^0$ (in most observational data,

$|C| \gg |T|$). Let $\left\{\left(\mathbf{X}_t, Y_t^1, Y_t^0\right)\right\}_{t\in T}$ and $\left\{\left(\mathbf{X}_c, Y_c^1, Y_c^0\right)\right\}_{c\in C}$ be the random variables corresponding to the covariate values and treatment and control responses of the units in T and C, respectively. While the values Y_t^0 and Y_c^1 are never actually observed, it is helpful to include them for discussion.

The *sample average treatment effect* (SATE) is

$$\tau_{T,C} \equiv \frac{1}{|T|+|C|}\left(\sum_{t\in T}\left(Y_t^1 - Y_t^0\right) + \sum_{c\in C}\left(Y_c^1 - Y_c^0\right)\right)$$

and the *sample average treatment effect for the treated* (SATT) is

$$\tau_T^1 \equiv \frac{1}{|T|}\sum_{t\in T}\left(Y_t^1 - Y_t^0\right).$$

Both $\tau_{T,C}$ and τ_T^1 are random variables that depend on the sampled units. To contrast these new terms, τ and τ^1 are often referred to as the *population average treatment effect* (PATE) and the *population average treatment effect for the treated* (PATT), respectively (Imai et al., 2008). The variables $\tau_{T,C}$ and τ_T^1 cannot be calculated because the values Y_t^0 and Y_c^1 are unobserved. Specifically, both $\tau_{T,C}$ and τ_T^1 require the estimation of the average control response of the treatment units, $\bar{Y}_T^0 \equiv \sum_{t\in T} Y_t^0/|T|$, and $\tau_{T,C}$ further requires the estimation of the average treatment response of the control units, $\bar{Y}_C^1 \equiv \sum_{c\in C} Y_c^1/|C|$. The following assumption supports the calculation of estimates for these quantities.

Assumption 2.4. The random variables $\left\{\left(\mathbf{X}_t, Y_t^1, Y_t^0\right)\right\}_{t\in T}$ and $\left\{\left(\mathbf{X}_c, Y_c^1, Y_c^0\right)\right\}_{c\in C}$ are mutually (jointly) independent. Conditional on the covariates, each unit is sampled at random (uniformly) from the appropriate subpopulation:

$$\begin{aligned}
\mathbf{Pr}\left(Y_t^1 \le y \mid \mathbf{X}_t = \mathbf{x}\right) &= \mathbf{Pr}\left(Y^1 \le y \mid \mathbf{X} = \mathbf{x},\ Z = 1\right) \quad \forall\, t \in T,\ y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}^1,\\
\mathbf{Pr}\left(Y_t^0 \le y \mid \mathbf{X}_t = \mathbf{x}\right) &= \mathbf{Pr}\left(Y^0 \le y \mid \mathbf{X} = \mathbf{x},\ Z = 1\right) \quad \forall\, t \in T,\ y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}^1,\\
\mathbf{Pr}\left(Y_c^1 \le y \mid \mathbf{X}_c = \mathbf{x}\right) &= \mathbf{Pr}\left(Y^1 \le y \mid \mathbf{X} = \mathbf{x},\ Z = 0\right) \quad \forall\, c \in C,\ y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X},\\
\mathbf{Pr}\left(Y_c^0 \le y \mid \mathbf{X}_c = \mathbf{x}\right) &= \mathbf{Pr}\left(Y^0 \le y \mid \mathbf{X} = \mathbf{x},\ Z = 0\right) \quad \forall\, c \in C,\ y \in \mathbb{R},\ \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

The Strong Ignorability assumptions 2.2 and 2.3 together with Assumption 2.4 make it possible to link the observed responses from one set of units with the unobserved responses in the other set of units through the units' covariate values. These links can then be used to construct estimates of $\tau_{T,C}$ and τ_T^1. Methods for doing this are discussed in Section 2.3. An additional assumption allows these methods to be applied to

τ and τ^1, as well.

Assumption 2.5. The sets T and C are simple random samples (i.e., units are sampled uniformly and without replacement) from U^1 and U^0, respectively.

Assumption 2.5 subsumes Assumption 2.4. An immediate consequence of Assumption 2.5 is:

$$\mathbf{E}\left[Y_t^1\right] = \mathbf{E}\left[Y^1 \mid Z = 1\right] \quad \forall\, t \in T, \tag{2.10a}$$

$$\mathbf{E}\left[Y_t^0\right] = \mathbf{E}\left[Y^0 \mid Z = 1\right] \quad \forall\, t \in T, \tag{2.10b}$$

$$\mathbf{E}\left[Y_c^1\right] = \mathbf{E}\left[Y^1 \mid Z = 0\right] \quad \forall\, c \in C, \tag{2.10c}$$

$$\mathbf{E}\left[Y_c^0\right] = \mathbf{E}\left[Y^0 \mid Z = 0\right] \quad \forall\, c \in C. \tag{2.10d}$$

The results from (2.10) ensure that

$$\begin{aligned}
\mathbf{E}\left[\tau_{T,C}\right] &= \mathbf{E}\left[\frac{1}{|T|+|C|}\left(\sum_{t\in T}\left(Y_t^1 - Y_t^0\right) + \sum_{c\in C}\left(Y_c^1 - Y_c^0\right)\right)\right] \\
&= \frac{1}{|T|+|C|}\left(\sum_{t\in T}\mathbf{E}\left[Y_t^1 - Y_t^0\right] + \sum_{c\in C}\mathbf{E}\left[Y_c^1 - Y_c^0\right]\right) \\
&= \frac{1}{|T|+|C|}\left(\sum_{t\in T}\mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right] + \sum_{c\in C}\mathbf{E}\left[Y^1 - Y^0 \mid Z = 0\right]\right) \\
&= \left(\frac{|T|}{|T|+|C|}\right)\mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right] + \left(\frac{|C|}{|T|+|C|}\right)\mathbf{E}\left[Y^1 - Y^0 \mid Z = 0\right]
\end{aligned} \tag{2.11}$$

and

$$\begin{aligned}
\mathbf{E}\left[\tau_T^1\right] &= \mathbf{E}\left[\frac{1}{|T|}\left(\sum_{t\in T}\left(Y_t^1 - Y_t^0\right)\right)\right] \\
&= \frac{1}{|T|}\sum_{t\in T}\mathbf{E}\left[Y_t^1 - Y_t^0\right] \\
&= \frac{1}{|T|}\sum_{t\in T}\mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right] \\
&= \mathbf{E}\left[Y^1 - Y^0 \mid Z = 1\right] \\
&= \tau^1.
\end{aligned} \tag{2.12}$$

Thus, under Assumption 2.5, SATT equals ATT in expectation. If it is further assumed that $\mathbf{Pr}\left(Z = 1\right) =$

$|T| / (|T| + |C|)$ and $\mathbf{Pr}\left(Z=0\right)$, then (2.11) simplifies to

$$\begin{aligned}\mathbf{E}\left[\tau_{T,C}\right] &= \mathbf{Pr}\left(Z=1\right) \cdot \mathbf{E}\left[Y^1 - Y^0 \mid Z=1\right] + \mathbf{Pr}\left(Z=0\right) \cdot \mathbf{E}\left[Y^1 - Y^0 \mid Z=0\right] \\ &= \mathbf{E}\left[Y^1 - Y^0\right] \\ &= \tau,\end{aligned}$$

so SATE equals ATE in expectation.

As mentioned previously, τ^1 is often more important than τ. The results in (2.12) reveal that τ^1 can be estimated by estimating τ_T^1 under Assumption 2.5, which requires estimating $\bar{Y}_T^0$. The next section presents several of the most common approaches that have been used for this purpose.

2.3 Existing Methods for Observational Data

Within the causal inference literature, two prominent methods for estimating τ_T^1 are *regression* and *matching* (Rubin, 1973b, 1979). Regression attempts to construct a model of the relationship between the covariates and the control responses using the units in C, and then uses this model to estimate Y_t^0 from $\mathbf{X}_t$ for each $t \in T$. The average of these estimates is then used as an estimate of $\bar{Y}_T^0$. Matching attempts to pair each treatment unit $t \in T$ with a similar control unit $c \in C$ and then uses the average control response of the matched control units as an estimator for $\bar{Y}_T^0$.

2.3.1 Regression

Regression methods for causal inference typically entail: (1) constructing a hypothetical model for the relationship between the covariates and the control response; (2) estimating the parameters of the model from the available data; (3) using the model and estimated parameters to predict each treatment unit's control response from its covariate values.

Linear regression is the most well-known of these methods, and it hypothesizes a linear relationship between (functions of) the covariate values and the control responses. For example, a researcher may assume that each unit's control response is determined by the function

$$y_u^0 = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_u + \alpha + \varepsilon_u^0,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$ are parameters of the model and ε_u^0 is an error term for unit $u \in U$. The researcher then uses the available data, specifically $\left\{\left(\mathbf{X}_c, Y_c^0\right)\right\}_{c\in C}$, to determine likely values for $\boldsymbol{\beta}$ and α. Parameter estimates $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\alpha}$ can be used to compute *residuals* for each of the control units as $\widetilde{\varepsilon}_u^0 \equiv y_u^0 - \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_u - \widetilde{\alpha}$.

Many techniques have been developed to construct estimates of the model parameters $\boldsymbol{\beta}$ and α in linear regression. The method of *least squares* provides an estimate that minimizes the sum of squared residuals, while the *least absolute deviations* method minimizes the sum of absolute values of the residuals. Under certain additional assumptions, several useful properties can be established. For example, if the error terms ε_u^0 have conditional mean zero given the covariates and the model is correctly specified, then the least squares estimate of $\boldsymbol{\beta}$ and α is unbiased (Cochran and Rubin, 1973).

Linear regression models can include higher-order terms such as $\gamma_i x_{ui}^2$ and $\beta_{ij}x_{ui}x_{uj}$; only the parameters have to be linear. More complicated models between the covariates and the control responses (e.g., $y_u^0 = \beta_i x_{ui}^{\gamma} + \beta_j x_{uj} + \alpha + \varepsilon_u^0$, where γ is a parameter to be estimated) require the use of nonlinear regression techniques. Both linear and nonlinear regression are classified as *parametric methods* because they involve the estimation of a finite number of parameters for a predetermined model. In contrast, *nonparametric regression* does not require the specification of a model and instead attempts to estimate both the form of the model and its parameters from the available data.

King and Zeng (2006) note that one of the primary difficulties when constructing causal estimates from regression models is *extrapolation*, which occurs when the model is used to estimate Y_t^0 from a value $\mathbf{X}_t$ that lies outside the range of the covariate values $\{\mathbf{X}_c\}_{c\in C}$. The farther away $\mathbf{X}_t$ is from the control units' covariate values, the more dependent the estimate of Y_t^0 becomes on the model itself. In such a situation, a misspecification of the model can lead to incorrect estimates of τ_T^1.

2.3.2 Exact Matching

Exact matching is motivated by Assumption 2.3 and the estimation procedure described in Section 2.1.2 in which a treatment unit is paired with a control unit having identical covariate values (Rubin, 1973a). One of the most common forms of exact matching is *one-to-one matching*, which pairs each $t \in T$ with a $c \in C$ to form a *complete matched-pair sample* $M \equiv \{(t,c)\}_{t\in T}$ (Rosenbaum, 1989). The matching may be done either *with replacement*, in which case control units can be matched to more than one treatment unit, or *without replacement*, in which case each control unit can be used at most once. Let $C_M \equiv \{c \in C : \exists\ t \in T,\ (t,c) \in M\}$ be the set of matched control units for a matched-pair sample M.

A matched-pair sample M is *exact* if each pair $(t,c) \in M$ satisfies $\mathbf{X}_t = \mathbf{X}_c$. Under Assumptions 2.3 and 2.4, an exact matched-pair sample M satisfies the following after conditioning on the covariate values

$\mathbf{X}_t = \mathbf{x}_t$ for each of the treatment units:

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{|T|}\sum_{(t,c)\in M} Y_c^0 \mid \mathbf{X}_c = \mathbf{X}_t = \mathbf{x}_t \ \forall\ t \in T\right] &= \frac{1}{|T|}\sum_{(t,c)\in M} \mathbf{E}\left[Y_c^0 \mid \mathbf{X}_c = \mathbf{x}_t\right] \\
&= \frac{1}{|T|}\sum_{(t,c)\in M} \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}_t,\ Z = 0\right] \\
&= \frac{1}{|T|}\sum_{(t,c)\in M} \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}_t,\ Z = 1\right] \\
&= \frac{1}{|T|}\sum_{t\in T} \mathbf{E}\left[Y_t^0 \mid \mathbf{X}_t = \mathbf{x}_t\right] \\
&= \mathbf{E}\left[\frac{1}{|T|}\sum_{t\in T} Y_t^0 \mid \mathbf{X}_t = \mathbf{x}_t \ \forall\ t \in T\right] \\
&= \mathbf{E}\left[\bar{Y}_T^0 \mid \mathbf{X}_t = \mathbf{x}_t \ \forall\ t \in T\right].
\end{aligned} \tag{2.13}$$

That is, the expected value of the average control response of the matched control units equals the expected value of the average control response of the treatment units, conditional on the observed covariate values of the treatment units. Thus, the exact matched-pair sample M can be used to construct an unbiased estimate of τ_T^1. Under Assumption 2.5, this also provides an unbiased estimate of τ^1.

The major difficulty with exact matching is that it is unlikely for each treatment unit to have an exact match in C, even for a limited number of covariates. One way to address this is to use *incomplete matching*, which drops treatment units that do not have an identical control unit (Rosenbaum and Rubin, 1985). However, dropping treatment units is generally regarded as undesirable because it alters the quantity being estimated from τ_T^1 to an estimate of the treatment effect for the matched treatment units.

2.3.3 Propensity Score Matching

Propensity score matching seeks to replace the problem of exactly matching treatment and control units on a multi-dimensional covariate vector with that of exactly matching units on a scalar summary of their covariates known as the *propensity score* (Rosenbaum and Rubin, 1983b). The propensity score represents the probability of receiving treatment given the covariate values, defined as $e(\mathbf{x}) \equiv \mathbf{Pr}\left(Z = 1 \mid \mathbf{X} = \mathbf{x}\right)$. Rosenbaum and Rubin show that if Assumption 2.2 holds, then it is also the case that $\left(Y^1, Y^0\right) \perp\!\!\!\perp Z \mid e(\mathbf{X})$ and $0 < \mathbf{Pr}\left(Z = 1 \mid e(\mathbf{X}) = r\right) < 1 \ \forall\ r \in (0, 1)$. From this, it can be shown that a matched-pair sample M that is exact with respect to the propensity score (i.e., $e(\mathbf{X}_t) = e(\mathbf{X}_c)$ for all $(t, c) \in M$) yields an unbiased estimate of τ_T^1, conditioned on the covariate values of the treatment units.

Propensity score matching is one of the most popular approaches for causal inference with observational data. There are two major difficulties with the propensity score approach, however. The first is that exact matches on the propensity score may still be difficult to obtain. The second is that the true propensity score itself is almost always unknown, and thus must be estimated. This is typically done using a logistic regression model relating the covariates with the treatment status (1 or 0) for each of the units, though it is often impossible to determine if the correct model was used.

Rosenbaum and Rubin (1984) provide some guidance on how to assess the quality of the estimated propensity score model by using a result from Rosenbaum and Rubin (1983b, Theorem 1), which states that $\mathbf{X} \perp\!\!\!\perp Z \mid e(\mathbf{X})$. Intuitively, this result says that the distributions of covariate values in the subpopulations $\left\{t \in U^1 : e(\mathbf{x}_t) = r\right\}$ and $\left\{c \in U^0 : e(\mathbf{x}_c) = r\right\}$ are identical for all $r \in (0,1)$. Under Assumption 2.4, this means that any set of k treatment units $t_1, t_2, \ldots, t_k$ from T and k control units $c_1, c_2, \ldots, c_k$ from C that have the same set of propensity scores $r_1, r_2, \ldots, r_k$ satisfy

$$\begin{aligned}&\mathbf{Pr}\left(\mathbf{X}_{t_1} \leq \mathbf{x}_1,\ \ldots,\ \mathbf{X}_{t_k} \leq \mathbf{x}_k \mid e(\mathbf{X}_{t_1}) = r_1,\ \ldots,\ e(\mathbf{X}_{t_k}) = r_k\right) \\ &\qquad = \mathbf{Pr}\left(\mathbf{X}_{c_1} \leq \mathbf{x}_1,\ \ldots,\ \mathbf{X}_{c_k} \leq \mathbf{x}_k \mid e(\mathbf{X}_{c_1}) = r_1,\ \ldots,\ e(\mathbf{X}_{c_k}) = r_k\right),\end{aligned} \tag{2.14}$$

for all $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. From (2.14), it can be shown that any function f defined on the covariate values satisfies

$$\begin{aligned}&\mathbf{E}\left[f(\mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_k}) \mid e(\mathbf{X}_{t_1}) = r_1,\ \ldots,\ e(\mathbf{X}_{t_k}) = r_k\right] \\ &\qquad = \mathbf{E}\left[f(\mathbf{X}_{c_1}, \ldots, \mathbf{X}_{c_k}) \mid e(\mathbf{X}_{c_1}) = r_1,\ \ldots,\ e(\mathbf{X}_{c_k}) = r_k\right].\end{aligned} \tag{2.15}$$

An example is the mean function $f_\mu : \mathcal{X} \times \mathcal{X} \times \ldots \times \mathcal{X} \to \mathbb{R}^p$ given by $f_\mu(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k) \equiv (1/k)\sum_{i=1}^k \mathbf{x}_i$. So if a complete matched-pair sample M matches units exactly on the propensity score, then

$$\mathbf{E}\left[\frac{1}{|T|}\sum_{(t,c)\in M} (\mathbf{X}_t - \mathbf{X}_c) \mid e(\mathbf{X}_t) = e(\mathbf{X}_c) = r_t \ \forall\ (t,c) \in M\right] = \begin{pmatrix} 0 & 0 & \ldots & 0 \end{pmatrix}^{\mathrm{T}}. \tag{2.16}$$

Hence the treatment units and the matched control units are *stochastically balanced* on their covariate means. More generally, as (2.15) shows, the two sets of units are stochastically balanced with respect to any function of the covariates.

A comparison of the function values for two (multi)sets of covariate vectors is referred to as a *covariate balance measure*. If the result of the comparison is a score indicating dissimilarity, such as

$$\left\| f_\mu(\mathbf{x}_1, \ldots, \mathbf{x}_k) - f_\mu(\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{2k}) \right\|_1,$$

then the comparison is referred to as an *imbalance measure*. Formally, an imbalance measure is a function $\mathcal{I} : \mathbb{N}^{\mathcal{X}} \times \mathbb{N}^{\mathcal{X}} \to [0, \infty)$, where $\mathbb{N}$ is the set of natural numbers (including zero), and $\mathbb{N}^{\mathcal{X}}$ is the set of all multisets containing elements from $\mathcal{X}$. To simplify notation, let $\mathcal{I}(S_1, S_2) \equiv \mathcal{I}\big(\{\mathbf{x}_u\}_{u \in S_1}, \{\mathbf{x}_u\}_{u \in S_2}\big)$ for any finite sets of units $S_1 \subset U$ and $S_2 \subset U$.

Rosenbaum and Rubin (1984) use (2.15) to propose the following procedure for constructing a propensity score model: (1) construct a model for the propensity score and estimate the parameters from the sampled units T and C; (2) match units on their estimated propensity scores; (3) compute covariate imbalance measures for the matched samples; (4) if the imbalance is statistically significant, adjust the propensity score model and repeat the process. Determining appropriate imbalance measures in step (3) is an open question (Nikolaev et al., 2013).

In practice, Ho et al. (2007) advocate using propensity score matching not for its theoretical properties but for its ability to produce a matched-pair sample that features covariate balance. They refer to the process of estimating the propensity score as the *propensity score tautology*: if matching on the estimated propensity score produces covariate balance in the matched-pair sample, then the estimated propensity score is regarded as a reasonable estimate of the true propensity score. However, they argue that once covariate balance is achieved, the estimated propensity score is no longer needed, since the quality of the estimates from the matching can be shown by covariate balance alone (Ho et al., 2007, pg. 219).

2.3.4 Inexact Matching

Inexact matching is a general term for any matching method that does not require exactly matched pairs, either on the covariates themselves or on an estimated propensity score. Because such pairs are rare in practice, most matching methods are inexact. Reviews of inexact matching methods are provided by Imbens (2004) and Stuart (2010).

Inexact matching requires a distance function $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ that scores each potential pair $(t, c) \in T \times C$. A matched-pair sample M that minimizes $\sum_{(t,c) \in M} \delta(\mathbf{X}_t, \mathbf{X}_c)$ is sought. Distance functions have been constructed using the Mahalanobis metric (Rubin, 1980), subclassification on estimated propensity scores (Rosenbaum and Rubin, 1983b), and *matching calipers*, which are maximum acceptable distances for a matched pair (Althauser and Rubin, 1970). Combinations of these metrics have also been considered (Rosenbaum and Rubin, 1985).

If matching is done with replacement, identifying an optimal matched-pair sample for a given distance function δ is simply a matter of pairing each treatment unit with the closest control unit. If replacement is not allowed, however, then there may be competition between treatment units that share a closest control unit. Deciding how to pair each treatment unit with a control unit then becomes an optimization problem, specifically the *assignment problem* (Kuhn, 1955). The assignment problem is well-studied in combinatorial optimization (Edmonds, 1965; Ahuja et al., 1993) and can be solved in polynomial time.

Researchers have undertaken several investigations into the theoretical properties of inexact matching. Rubin (1976a,b) identifies matching methods that are *equal percent bias reducing*, meaning that the matched-pair sample M reduces mean imbalance in each of the covariates by the same percentage compared to the mean imbalance in the samples T and C. Abadie and Imbens (2006) investigate the properties of estimators for τ constructed from an inexact matched-pair sample M. Abadie and Imbens (2011) incorporate a bias correction factor into the estimator of Abadie and Imbens (2006) and analyzes its properties.

The theoretical results provide little guidance on which distance metric should be used in practice. Instead, the generally accepted principle is that a researcher should try several metrics and use the one that produces the matched-pair sample with the best covariate balance measures (Rosenbaum and Rubin, 1985; Diamond and Sekhon, 2013). However, there is no clear consensus regarding which covariate balance measures should be used or when the balance is sufficient.

2.3.5 Matching for Covariate Balance

Several matching methods have been proposed that incorporate covariate balance measures more directly into the selection process. Greenberg (1953) and Rubin (1973a) consider covariate balance as an objective in matching methods. In particular, Rubin refers to the process of finding a matched-pair sample M that minimizes

$$\frac{1}{|T|} \sum_{(t,c) \in M} (\mathbf{X}_t - \mathbf{X}_c)$$

as *mean matching* and uses heuristics to solve the problem. However, mean matching has received little attention in subsequent literature, most likely due to the fact that the relationship between the control responses and the covariates needs to be linear in order for estimates from mean matching to be unbiased (Cochran and Rubin, 1973).

Rosenbaum (1989) and Rosenbaum et al. (2007) formulate a matching objective based on both a distance metric and a covariate balance measure, referred to as *balanced matching* and *fine balance* in the respective publications. Fine balance is defined as "exactly balancing a nominal variable, often one with many categories, without trying to match individuals on this variable" (Rosenbaum et al., 2007, pg. 75).

In this approach, one nominal covariate is selected for fine balance, and then an optimal matching (using the remaining covariates to define the distance function) is constructed subject to the constraint that each category of the balancing covariate contains an equal proportion of treatment units and matched control units. The fine balance problem can be solved efficiently by transforming it into a network flow problem, or by making appropriate modifications to the distance matrix. Zubizarreta et al. (2011) present a practical application of fine balance and discuss some of the issues associated with fine balance and matching.

Yang et al. (2012) extend fine balance to *near-fine balance*, where the requirement of exact balance on each category of the nominal covariate is relaxed by introducing upper and lower bounds on the accepted deviation in each category. Yang et al. adapt the network for fine balance to handle near-fine balance. Zubizarreta (2012) further extends fine balance and near-fine balance by seeking fine or near-fine balance for multiple covariates that are not necessarily nominal. The model can also include alternate covariate balance measures, provided that they can be modeled as linear functions. The resulting optimization problem can be formulated as a mixed integer program (MIP) and solved using appropriate techniques.

Diamond and Sekhon (2013) argue in favor of using covariate balance measures as an optimization goal in observational studies. In support of this position, they note that many current matching methods are variably effective at recovering the correct treatment effect estimate in a well-studied dataset in the literature (LaLonde, 1986), and they attribute these discrepancies to the matching methods' failure to achieve sufficient covariate balance. To remedy the lack of covariate balance, Diamond and Sekhon propose *Genetic Matching*, a genetic algorithm that iteratively adjusts the distance metric used for matching until it produces a matched-pair sample with optimal covariate balance. Diamond and Sekhon (2013) also argue that covariate imbalance should be minimized as much as possible, instead of to the point where a statistical test indicates that the imbalance is insignificant.

2.3.6 Related Work

The combination of matching and regression for constructing causal estimates has also been explored (Cochran and Rubin, 1973; Rubin, 1979; Ho et al., 2007). These methods use matching as a pre-processing step to identify control units from which the parameters of a regression model are estimated. If the treatment units and the selected control units are similar, then there is less danger of extrapolation when estimating the treatment units' control responses. Ho et al. (2007) argue that this approach is *doubly robust* because it provides an unbiased estimate if either the matching is successful or the regression model is correct. It also allows for the correction of residual covariate imbalance after matching by the regression model.

Other types of matching besides one-to-one have also been explored. In particular, *one-to-k matching* pairs each treatment unit with k control units, while *full matching* seeks the best clustering of units from T and C so that every unit has at least one paired unit in the opposite group (Rosenbaum, 1991). Hansen and Klopfer (2006) shows how to solve the full matching problem using network flow techniques. Rosenbaum (2012) presents an incomplete matching method in which unmatched treatment units incur a penalty. This creates a trade-off between including treatment units with bad matches versus dropping these units altogether. The resulting model can be solved using network optimization.

Another concern in observational studies is that Assumptions 2.2 and 2.3 may be invalid due to the presence of an *unobserved* covariate. Detecting such situations is of interest in order to ensure that the estimated treatment effect is not incorrectly attributed to the treatment itself. Rosenbaum and Rubin (1983a) and Rosenbaum (1987b,a) present methods for detecting unobserved covariates. In particular, Rosenbaum and Rubin identify the properties that an unobserved covariate would need to have in order to explain the estimated treatment effect in terms of differences in the unobserved covariate instead of the treatment itself.

Iacus et al. (2012) address the difficulty of finding exact matches by coarsening the covariate values used for matching. Their procedure, *Coarsened Exact Matching* (CEM), makes it more likely that two units will have identical (coarsened) values for all covariates. Treatment units without exact matches for the coarsened values are dropped from further consideration.

Hainmueller (2012) proposes a maximum entropy re-weighting scheme, *entropy balancing*, that adjusts weights for each of the control units in order to meet user-specified balance constraints placed on the moments of the covariate distributions. Ratkovic (2012) adapts the support vector machine classifier to identify a set of control units $C' \subseteq C$ that is indistinguishable from T with respect to the covariates.

2.4 Balance Optimization Subset Selection

The *Balance Optimization Subset Selection* (BOSS) framework (Nikolaev et al., 2013; Cho et al., 2013) is motivated by the emphasis on covariate balance measures for assessing the quality of matched-pair samples in matching methods for observational studies (Rosenbaum and Rubin, 1985; Diamond and Sekhon, 2013). Instead of pairing treatment and control units like matching methods, BOSS seeks a *control group* $C' \subseteq C$ from the *control pool* C that minimizes a given covariate imbalance measure $\mathcal{I}$ with respect to the *treatment group* T. BOSS accommodates any scalar-valued imbalance measure of the covariates, and it permits size constraints on the control group (e.g., $|C'| = |T|$) and allows for selecting control units with repetitions (i.e.,

C' can be a multiset). Following Diamond and Sekhon (2013), BOSS seeks to minimize imbalance rather than to reduce it below statistical significance.

An optimal control group $C' \subseteq C$ from BOSS is used to estimate τ_T^1 as

$$\widetilde{\tau}_T^1(C') \equiv \frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0. \tag{2.17}$$

Alternatively, BOSS can be used for pre-processing in the approach of Ho et al. (2007) where the parameters of a regression model are estimated from C'. Properties of (2.17) are explored in Chapter 3.

2.4.1 Imbalance Measures

BOSS can use any potential imbalance measure to compare T with C'. Section 2.3.3 considered the difference of covariate means in (2.16); as an imbalance measure, this is

$$\mathcal{I}_{\text{DOM}}\left(T, C'\right) \equiv \left\|\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_{C'}\right\|_1 = \sum_{i\in\mathcal{P}} \left| \frac{1}{|T|}\sum_{t\in T} X_{ti} - \frac{1}{|C'|}\sum_{c\in C'} X_{ci} \right|,$$

where $\bar{\mathbf{X}}_T \equiv \sum_{t\in T}\mathbf{X}_t/|T|$ and $\bar{\mathbf{X}}_{C'} \equiv \sum_{c\in C'}\mathbf{X}_c/|C'|$. A variant of $\mathcal{I}_{\text{DOM}}$ is

$$\mathcal{I}_{\text{SDOM}}(T, C') \equiv \sum_{i\in\mathcal{P}} \frac{\left|\sum_{t\in T} X_{ti}/|T| - \sum_{c\in C'} X_{ci}/|C'|\right|}{\left|\sum_{t\in T} X_{ti}/|T| - \sum_{c\in C} X_{ci}/|C|\right|},$$

which scales the imbalance for each covariate by the imbalance between T and C (assuming that the imbalance is nonzero).

Another imbalance measure is based on a comparison of the marginal *empirical distribution functions* of the covariates. For a covariate $i \in \mathcal{P}$, the empirical distribution function is $\widehat{F}_i(S,x) \equiv \left|\{u \in S : X_{ui} \leq x\}\right| / |S|$ for any finite set of units $S \subset U$ and value $x \in \mathbb{R}$. That is, $\widehat{F}_i(S,x)$ is the proportion of units in S with value at most x for covariate i. The two-sample Kolmogorov-Smirnov statistic values for the individual covariates lead to the imbalance measure

$$\mathcal{I}_{\text{KS}}\left(T, C'\right) \equiv \sum_{i\in\mathcal{P}} \max_{x\in\mathcal{X}_i(T\cup C)} \left|\widehat{F}_i(T,x) - \widehat{F}_i(C',x)\right|,$$

where $\mathcal{X}_i(T\cup C) \equiv \{X_{ui} : u \in T\cup C\}$ is the set of values for covariate i that are observed within the sampled units in T and C. A variant of $\mathcal{I}_{\text{KS}}$ which uses the maximum Kolmogorov-Smirnov two-sample test statistic value across all covariates is

$$\mathcal{I}_{\text{KS:max}}(T, C') \equiv \max_{i \in \mathcal{P}} \left\{ \max_{x \in \mathcal{X}_i(T \cup C)} \left| \widehat{F}_i(T, x) - \widehat{F}_i(C', x) \right| \right\}.$$

Another distribution-based balance measure relies on *coarsening* the covariate values through a histogram binning process (Nikolaev et al., 2013; Iacus et al., 2012). For covariate $i \in \mathcal{P}$, let n_i be the number of histogram bins and let

$$\min_{u \in T \cup C} X_{ui} \equiv b_{i0} < b_{i1} < \ldots < b_{in_i} \equiv \max_{u \in T \cup C} X_{ui} \tag{2.18}$$

be the $n_i + 1$ *bin boundaries.* Then let

$$\begin{aligned} B_{i1} &\equiv \{u \in T \cup C : b_{i0} \leq X_{ui} \leq b_{i1}\}, \\ B_{ij} &\equiv \{u \in T \cup C : b_{i,j-1} < X_{ui} \leq b_{ij}\} \quad \forall\, j = 2, \ldots, n_i, \end{aligned}$$

be the partition of $T \cup C$ determined by the bin boundaries in (2.18). For notational convenience, let $N_i \equiv \{1, 2, \ldots, n_i\}$. For any $S \subseteq T \cup C$, let $\eta_{ij}(S) \equiv |S \cap B_{ij}|/|S|$ be the proportion of units in S that occupy bin $j \in N_i$ for covariate $i \in \mathcal{P}$. Using these histogram bins, a marginal histogram imbalance measure is

$$\mathcal{I}_{\text{Diff}}(T, C') \equiv \sum_{i \in \mathcal{P}} \sum_{j \in N_i} |\eta_{ij}(T) - \eta_{ij}(C')|,$$

which compares the proportions of treatment units and control units in each histogram bin of each covariate. As the granularity of the bins becomes finer, $\mathcal{I}_{\text{Diff}}$ provides a stricter measure of marginal covariate imbalance.

Alternate ways to assess imbalance are possible. For example, $\mathcal{I}_{\text{Diff}}$ can be adapted to use a quadratic penalty for discrepancies within each bin, leading to

$$\mathcal{I}_{\text{Diff}^2}(T, C') \equiv \sum_{i \in \mathcal{P}} \sum_{j \in N_i} \left(\eta_{ij}(T) - \eta_{ij}(C')\right)^2.$$

A control group C' that satisfies $\mathcal{I}_{\text{Diff}}(T, C') = 0$ will also satisfy $\mathcal{I}_{\text{Diff}^2}(T, C') = 0$. However, when no control group has zero imbalance with respect to these imbalance measures, $\mathcal{I}_{\text{Diff}^2}$ helps BOSS find a control group without extreme discrepancies in any bin. A variant of $\mathcal{I}_{\text{Diff}^2}$ that is similar to the χ^2 test-statistic for control groups that satisfy $|C'| = |T|$ is

$$\mathcal{I}_{\chi^2}(T, C') \equiv \sum_{i \in \mathcal{P}} \sum_{j \in N_i} \frac{\left(|C' \cap B_{ij}| - |T \cap B_{ij}|\right)^2}{\max\left(|T \cap B_{ij}|, 1\right)}.$$

The preceding imbalance measures focus on the marginal distributions of the covariates. Joint distributions of covariates can be incorporated through *covariate clusters* (Nikolaev et al., 2013). A covariate cluster $D \subseteq \mathcal{P}$ is a (nonempty) subset of the covariates. As an example, the covariate cluster $D = \{i_1, i_2\}$ captures the pairwise joint distribution of covariates i_1 and i_2. Let $\mathbf{P}^D$ be the $p \times p$ projection matrix associated with covariate cluster D, with

$$\mathbf{P}^D_{ij} \equiv \begin{cases} 1 & \text{if } i = j \text{ and } i \in D \\ 0 & \text{otherwise.} \end{cases}$$

For a covariate cluster $D \equiv \{i_1, i_2, \ldots, i_k\}$ with $1 \leq i_1 < i_2 < \ldots < i_k \leq p$, let

$$N_D \equiv N_{i_1} \times N_{i_2} \times \ldots \times N_{i_k}$$

denote the set of tuples of indices for the histogram bins of the cluster. For $j \equiv (j_1, j_2, \ldots, j_k) \in N_D$, let $B_{Dj} \equiv \bigcap_{l=1}^k B_{i_l, j_l}$ be the units that belong to the jth bin of cluster D. For any $S \subseteq T \cup C$, let $\eta_{Dj} \equiv |S \cap B_{Dj}| / |S|$ be the proportion of units in S that occupy bin $j \in N_D$.

The *joint empirical distribution function* for a covariate cluster $D \subseteq \mathcal{P}$ is

$$\widehat{F}_D(S, \mathbf{x}) \equiv \frac{\left|\left\{u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x}\right\}\right|}{|S|} \tag{2.19}$$

for all finite $S \subset U$ and $\mathbf{x} \in \mathbb{R}^p$, where the vector inequality is component-wise so that $X_{ui} \leq x_i$ for all $i \in D$. This definition is not invariant under a change of sign for the covariates because generally

$$|\{u \in S : X_{ui} \leq x_i,\ X_{uj} \leq x_j\}| \neq |\{u \in S : X_{ui} \leq x_i,\ X_{uj} \geq x_j\}|.$$

As a result, $\widehat{F}_D(S, \mathbf{x})$ does not lead to an immediate generalization of the Kolmogorov-Smirnov statistic to the multivariate case (some possible extensions are presented by Peacock (1983), Fasano and Franceschini (1987), and Justel et al. (1997)). However, for the purposes of defining an imbalance measure it suffices to consider only one ordering with $\leq$ being used for all covariates, as indicated in (2.19).

For a set of covariate clusters $\mathbf{D}$, two possible multivariate extensions of $\mathcal{I}_{\text{KS}}$ and $\mathcal{I}_{\text{Diff}}$ are

$$\mathcal{I}_{\text{ecdf:}\mathbf{D}}\left(T, C'\right) \equiv \sum_{D \in \mathbf{D}} \max_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} \left|\widehat{F}_D(T, \mathbf{x}) - \widehat{F}_D(C', \mathbf{x})\right|$$

and

$$\mathcal{I}_{\text{Diff:}\mathbf{D}}\left(T, C'\right) \equiv \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \left|\eta_{Dj}(T) - \eta_{Dj}(C')\right|,$$

where

$$\mathcal{X}_D(T \cup C) \equiv \left\{\mathbf{P}^D \mathbf{x} : \mathbf{x} \in \mathbb{R}^p,\ x_i \in \mathcal{X}_i(T \cup C)\ \forall\ i \in \mathcal{P}\right\}$$

is the set of covariate vectors whose nonzero entries come from $\mathcal{X}_i(T \cup C)$ for each covariate $i \in D$. When $\mathbf{D} = \{\{i\} : i \in \mathcal{P}\}$ (i.e., there is a covariate cluster for each individual covariate), then $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ reduce to $\mathcal{I}_{\text{KS}}$ and $\mathcal{I}_{\text{Diff}}$, respectively. Similar extensions apply to $\mathcal{I}_{\text{KS:max}}$, $\mathcal{I}_{\text{Diff}^2}$, and $\mathcal{I}_{\chi^2}$.

The above imbalance measures only serve to illustrate what BOSS can use; they are by no means exhaustive. Other possibilities include using a combination of these measures (e.g., assessing mean balance on some covariates and distribution balance on others) or measuring imbalance as the maximum deviation across all covariates instead of the sum. The general form of the imbalance measure can also be used to include matching metrics. For example, for a given distance function δ, the imbalance measure $\mathcal{I}_\delta$ can be defined so that $\mathcal{I}_\delta(T, C')$ is equal to the minimum distance matching between T and C' if $|C'| = |T|$ and ∞ otherwise. For any $\mathcal{I}_\delta$, BOSS will not return the matched-pair sample itself, but Sauppe et al. (2014) show how the BOSS framework can be extended to explicitly include a matching component if desired.

2.4.2 Ideal Covariate Balance

With appropriately defined covariate clusters, $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ can be used to assess covariate imbalance on any number of marginal and joint distributions of the covariates. For example, the full joint distribution of covariates is captured by the covariate cluster $D = \mathcal{P}$. An ideal control group $C' \subseteq C$ would satisfy $\mathcal{I}_{\text{ecdf:}\mathbf{D}}(T, C') = 0$ with $\mathbf{D} = \{\mathcal{P}\}$. Such a group would exact match T in the sense that it could be used to construct an exact matched-pair sample M, possibly by allowing replacement of control units if $|T| > |C'|$. This is because $\mathcal{I}_{\text{ecdf:}\mathbf{D}}(T, C') = 0$ with $\mathbf{D} = \{\mathcal{P}\}$ implies that $\widehat{F}_{\mathcal{P}}(T, \mathbf{X}_t) = \widehat{F}_{\mathcal{P}}(C', \mathbf{X}_t)$ for all $t \in T$, and so each treatment unit has at least one control unit in C' with which it can be paired. With some modifications, the results of (2.13) can be extended to demonstrate that the BOSS estimator (2.17) computed from C' provides an unbiased estimate of τ_T^1 conditional on the covariate values of the treatment units.

In actuality, however, it is generally impossible to find a $C' \subseteq C$ that satisfies $\mathcal{I}_{\text{ecdf:}\mathbf{D}}(T, C') = 0$ for $\mathbf{D} = \{\mathcal{P}\}$. As such, a control group with imbalance on at least some of the higher-order joint distributions must generally be accepted. By appropriately defining the set of covariate clusters $\mathbf{D}$, a researcher is able to focus on identifying control groups that are balanced on the marginal and joint distributions that he or she deems relevant.

Identifying the covariate clusters to include in **D** is a challenging problem. Failing to include a covariate cluster may leave residual imbalance on the associated joint distribution of covariates, while including too many covariate clusters may make it impossible to identify a control group that is balanced with the treatment group on all of the associated covariate distributions. Related issues for consideration include how balance should be assessed and how residual imbalance should be handled. For example, it may be possible to exactly balance the marginal means or to moderately balance the marginal distributions themselves. Which choice is best is not always clear, and often depends on the assumptions that are made.

Chapter 3

Statistical Theory

This chapter discusses the relationship between covariate balance and bias in the BOSS estimator (2.17). In general, as covariate balance decreases, stronger assumptions are necessary to ensure that (2.17) provides an unbiased estimate of τ^1_T, which leads to a natural trade-off: weaker levels of covariate balance are easier to obtain but require stronger assumptions to increase confidence in the estimated treatment effect.

3.1 Reframing the Strong Ignorability Assumption

In order to facilitate discussion regarding covariate balance, Assumption 2.3 will be expressed in an alternative form. Define the *control response error*, hereafter referred to as error, for each $u \in U$ as $\varepsilon^0_u \equiv y^0_u - \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}_u\right]$, and let $\mathcal{E}^0$ be the (unobservable) random variable that represents the error for a randomly sampled unit from U. Define the *control response function* $h^0 : \mathcal{X} \to \mathbb{R}$ as

$$h^0(\mathbf{x}) \equiv \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}\right] \ \forall\, \mathbf{x} \in \mathcal{X}.$$

Assumption 2.3(a) implies that $\mathcal{E}^0 \perp\!\!\!\perp Z \mid \mathbf{X}$ because

$$\begin{aligned}\mathbf{Pr}\left(\mathcal{E}^0 \le r \mid \mathbf{X} = \mathbf{x},\ Z = z\right) &= \mathbf{Pr}\left(Y^0 - h^0(\mathbf{x}) \le r \mid \mathbf{X} = \mathbf{x},\ Z = z\right)\\ &= \mathbf{Pr}\left(Y^0 - h^0(\mathbf{x}) \le r \mid \mathbf{X} = \mathbf{x}\right)\\ &= \mathbf{Pr}\left(\mathcal{E}^0 \le r \mid \mathbf{X} = \mathbf{x}\right),\end{aligned}$$

for all $r \in \mathbb{R}$, $\mathbf{x} \in \mathcal{X}$, and $z \in \{0, 1\}$, which is an alternate way of expressing conditional independence. Additionally,

$$\mathbf{E}\left[\mathcal{E}^0 \mid \mathbf{X} = \mathbf{x}\right] = \mathbf{E}\left[Y^0 - h^0(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\right] = \mathbf{E}\left[Y^0 \mid \mathbf{X} = \mathbf{x}\right] - h^0(\mathbf{x}) = 0,$$

Some of the material in this chapter has been adapted from Nikolaev et al. (2013), *Operations Research* **61**(2), with the permission of the copyright holder.

for all $\mathbf{x} \in \mathcal{X}$, which implies that $\mathbf{E}\left[\mathcal{E}^0\right] = 0$ and that $\mathbf{X}$ and $\mathcal{E}^0$ are uncorrelated. Assumption 2.3(a) can now be restated as:

Assumption 3.1. The relationship between the control responses and the covariates is given by

$$Y^0 \equiv h^0\left(\mathbf{X}\right) + \mathcal{E}^0,$$

where h^0 is a (deterministic but unknown) function of the covariates, $\mathcal{E}^0$ represents an error term that satisfies $\mathbf{E}\left[\mathcal{E}^0 \mid \mathbf{X} = \mathbf{x}\right] = 0$ for all $\mathbf{x} \in \mathcal{X}$, and $\mathcal{E}^0 \perp\!\!\!\perp Z \mid \mathbf{X}$.

This alternative expression for the first portion of the strong ignorability assumption is useful both for understanding the intuition behind matching estimators and for extending this assumption to provide intuition for the BOSS estimator using various imbalance measures. First, using Assumption 3.1, the difference between the estimator (2.17) for a $C' \subseteq C$ and τ_T^1 is:

$$\begin{aligned}
\widetilde{\tau}_T^1(C') - \tau_T^1 &= \left(\frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0\right) - \frac{1}{|T|}\sum_{t\in T}\left(Y_t^1 - Y_t^0\right) \\
&= \frac{1}{|T|}\sum_{t\in T} Y_t^0 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0 \\
&= \frac{1}{|T|}\sum_{t\in T}\left(h^0(\mathbf{X}_t) + \mathcal{E}_t^0\right) - \frac{1}{|C'|}\sum_{c\in C'}\left(h^0(\mathbf{X}_c) + \mathcal{E}_c^0\right) \\
&= \left(\frac{1}{|T|}\sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c)\right) + \left(\frac{1}{|T|}\sum_{t\in T}\mathcal{E}_t^0 - \frac{1}{|C'|}\sum_{c\in C'}\mathcal{E}_c^0\right).
\end{aligned} \tag{3.1}$$

The second term with $\mathcal{E}_t^0$ and $\mathcal{E}_c^0$ is the averaged errors across the units in T and C'. By Assumptions 2.4 and 3.1, each of these quantities has expected value zero. Therefore, the only potential source of bias in the estimator $\widetilde{\tau}_T^1(C')$ is the first term in (3.1) with the control response function. Define the *control response function bias* as

$$\mathcal{B}(T, C') \equiv \frac{1}{|T|}\sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c).$$

For a matched-pair sample M consisting of pairs that are exactly matched on the covariates, the control response function bias $\mathcal{B}(T, C_M)$ equals zero because $h^0(\mathbf{X}_t)/|T| - h^0(\mathbf{X}_c)/|C_M| = 0$ for each pair $(t, c) \in M$, regardless of how h^0 is defined. If M contains inexact matches, $\mathcal{B}(T, C_M)$ is non-zero in general; however, Rubin (1973b) noted that for any set of inexact matches, there exists a polynomial control response function for which the bias equals zero. Despite this observation, researchers often justify causal estimates with inexact matches by claiming that the matched pairs feature "sufficient" covariate balance. However, lesser balance levels (i.e., balance on covariate distributions other than the full joint distribution of all covariates) are only sufficient to construct an unbiased estimate of the treatment effect if the control response function follows a specific form. This point is explored in detail in the subsequent sections by extending Assumption 3.1 in various ways.

3.2 Assumptions for Moment Balance

Moment balance measures are covariate balance measures that compare the moments of the covariate distributions between two sets of units. As noted in Chapter 2, a common way to assess covariate balance between T and $C' \subseteq C$ is through a comparison of the means, or first moments, of the covariates in T and C', given by $\bar{\mathbf{X}}_T \equiv \sum_{t \in T} \mathbf{X}_t / |T|$ and $\bar{\mathbf{X}}_{C'} \equiv \sum_{c \in C'} \mathbf{X}_c / |C'|$. Comparisons of higher raw moments (e.g., x_i^2) and interaction terms (e.g., $x_i x_j$) are also moment balance measures.

3.2.1 Mean Balance

Many researchers have informally observed that mean balance between T and C' is sufficient to guarantee an unbiased estimate of τ_T^1 only when there is a linear relationship between the covariates and the control responses (Cochran and Rubin, 1973; Rubin, 1973a; Rosenbaum and Rubin, 1985). The following *functional form* assumption for the control response function can be used to formally demonstrate this result.

Assumption 3.2. Assumption 3.1 holds with the additional requirement that $h^0(\mathbf{x}) \equiv \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \alpha$ for all $\mathbf{x} \in \mathcal{X}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$ (both $\boldsymbol{\beta}$ and α are fixed but unknown).

Theorem 3.3. *Under Assumption 3.2, a control group $C' \subseteq C$ with $\bar{\mathbf{X}}_{C'} = \bar{\mathbf{X}}_T$ (i.e., balance on first marginal moments) is both necessary and sufficient for $\mathcal{B}(T, C')$ to equal zero for all possible $\boldsymbol{\beta}$ and α.*

Proof. Sufficiency can be demonstrated for a C' with $\bar{\mathbf{X}}_{C'} = \bar{\mathbf{X}}_T$ as follows:

$$\begin{aligned}
\mathcal{B}(T, C') &= \frac{1}{|T|} \sum_{t \in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|} \sum_{c \in C'} h^0(\mathbf{X}_c) \\
&= \frac{1}{|T|} \sum_{t \in T} \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t + \alpha\right) - \frac{1}{|C'|} \sum_{c \in C'} \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_c + \alpha\right) \\
&= \boldsymbol{\beta}^{\mathrm{T}}\bar{\mathbf{X}}_T - \boldsymbol{\beta}^{\mathrm{T}}\bar{\mathbf{X}}_{C'} \\
&= \boldsymbol{\beta}^{\mathrm{T}}\left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_{C'}\right) \\
&= 0.
\end{aligned}$$

Necessity can be demonstrated by the contrapositive. Consider a C' with $\bar{\mathbf{X}}_{C'} \neq \bar{\mathbf{X}}_T$, and let $i^* \in \mathcal{P}$ be a covariate for which $\sum_{c \in C'} X_{ci^*}/|C'| \neq \sum_{t \in T} X_{ti^*}/|T|$. Then for $\boldsymbol{\beta}$ with $\beta_{i^*} = 1$ and $\beta_i = 0$ for all $i \in \mathcal{P} \backslash \{i^*\}$, the control response function bias is

$$\begin{aligned}
\mathcal{B}(T, C') &= \frac{1}{|T|} \sum_{t \in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|} \sum_{c \in C'} h^0(\mathbf{X}_c) \\
&= \frac{1}{|T|} \sum_{t \in T} \left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t + \alpha \right) - \frac{1}{|C'|} \sum_{c \in C'} \left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_c + \alpha \right) \\
&= \frac{1}{|T|} \sum_{t \in T} \beta_{i^*} X_{ti^*} + \alpha - \frac{1}{|C'|} \sum_{c \in C'} \beta_{i^*} X_{ci^*} - \alpha \\
&= \frac{1}{|T|} \sum_{t \in T} X_{ti^*} - \frac{1}{|C'|} \sum_{c \in C'} X_{ci^*},
\end{aligned}$$

which does not equal zero by choice of i^*. □

The proof of Theorem 3.3 establishes that $\mathcal{B}(T, C') = \boldsymbol{\beta}^{\mathrm{T}} \left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_{C'} \right)$. The "for all possible $\boldsymbol{\beta}$ and α" quantifier in Theorem 3.3 is needed because $\mathcal{B}(T, C')$ may equal zero when $\bar{\mathbf{X}}_{C'} \neq \bar{\mathbf{X}}_T$. For example, if $\bar{\mathbf{X}}_T \neq \bar{\mathbf{X}}_{C'}$ and $\boldsymbol{\beta} = \bar{\mathbf{X}}_T \times \bar{\mathbf{X}}_{C'}$, the vector cross product of the vectors of covariate means, then $\boldsymbol{\beta}^{\mathrm{T}} \bar{\mathbf{X}}_T = 0$ and $\boldsymbol{\beta}^{\mathrm{T}} \bar{\mathbf{X}}_{C'} = 0$, and consequently $\mathcal{B}(T, C') = 0$. Such cases are the exception rather than the norm.

The parameters $\boldsymbol{\beta}$ and α in the control response function of Assumption 3.2 can be estimated through linear regression techniques. Estimates of these parameters can then be used to construct a regression estimator for τ_T^1 as

$$\widetilde{\tau}_T^1 \equiv \frac{1}{|T|} \sum_{t \in T} \left(Y_t^1 - \widetilde{h}^0(\mathbf{X}_t) \right) = \frac{1}{|T|} \sum_{t \in T} \left(Y_t^1 - \left(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{X}_t + \widetilde{\alpha} \right) \right) = \bar{Y}_T^1 - \widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \bar{\mathbf{X}}_T - \widetilde{\alpha}.$$

However, as Theorem 3.3 shows, there is no need to estimate $\boldsymbol{\beta}$ and α when a control group C' satisfying $\mathcal{I}_{\mathrm{DOM}}(T, C') = 0$ is available.

3.2.2 Bias from Mean Imbalance

If Assumption 3.2 is valid, then BOSS can be used with $\mathcal{I}_{\mathrm{DOM}}$ to identify a $C' \subseteq C$ with $\bar{\mathbf{X}}_{C'} = \bar{\mathbf{X}}_T$, assuming that such a group exists, in order to construct an unbiased estimate of the treatment effect. When no control group meets this mean balance requirement, however, the residual imbalance between C' and T will introduce some bias in the estimate from (2.17) (in general). The following lemma provides a relationship between these two quantities.

Lemma 3.4. *Under Assumption 3.2,* $|\mathcal{B}(T, C')| \le \|\boldsymbol{\beta}\|_\infty \cdot \mathcal{I}_{DOM}(T, C')$.

Proof.

$$\begin{aligned}
|\mathcal{B}(T, C')| &= \left|\boldsymbol{\beta}^{\mathrm{T}}\left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_{C'}\right)\right| \\
&= \left|\sum_{i \in \mathcal{P}} \beta_i \left(\bar{X}_{Ti} - \bar{X}_{C'i}\right)\right| \\
&\le \sum_{i \in \mathcal{P}} |\beta_i| \cdot \left|\bar{X}_{Ti} - \bar{X}_{C'i}\right| \\
&\le \|\boldsymbol{\beta}\|_\infty \cdot \sum_{i \in \mathcal{P}} \left|\bar{X}_{Ti} - \bar{X}_{C'i}\right| \\
&= \|\boldsymbol{\beta}\|_\infty \cdot \mathcal{I}_{\mathrm{DOM}}(T, C')
\end{aligned}$$

□

An immediate consequence of Lemma 3.4 is that $|\mathcal{B}(T, C')| = |\beta_1| \cdot \mathcal{I}_{\mathrm{DOM}}(T, C')$ if there is only one covariate. Thus, by minimizing $\mathcal{I}_{\mathrm{DOM}}(T, C')$, BOSS attempts to minimize the residual bias. The bound provided by Lemma 3.4 is tight; the worst case occurs when all mean imbalance between T and C' is present in covariates $i \in \mathcal{P}$ with $\beta_i = \|\boldsymbol{\beta}\|_\infty$. Without knowing $\boldsymbol{\beta}$, $\mathcal{I}_{\mathrm{DOM}}$ may end up reducing imbalance in one covariate at the expense of increasing imbalance in another that has a larger impact on the value of the control response function. This issue led to the development of *equal percent bias reducing* (EPBR) methods (Rubin, 1976a,b) and later to *monotonic imbalance bounding* (MIB) methods (Iacus et al., 2011). The imbalance measure

$$\mathcal{I}_{\mathrm{DOM:max}}\left(T, C'\right) \equiv \left\|\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_{C'}\right\|_\infty = \max_{i \in \mathcal{P}} \left|\frac{1}{|T|}\sum_{t \in T} X_{ti} - \frac{1}{|C'|}\sum_{c \in C'} X_{ci}\right|$$

alleviates this problem by minimizing the maximum difference of means across all covariates. It may be necessary to normalize the covariate values in order for $\mathcal{I}_{\mathrm{DOM:max}}$ to provide a fair comparison between the different covariate means.

Lemma 3.5. *Under Assumption 3.2,* $|\mathcal{B}(T, C')| \le \|\boldsymbol{\beta}\|_1 \cdot \mathcal{I}_{DOM:\max}(T, C')$.

Proof.

$$\begin{aligned}
|\mathcal{B}(T,C')| &= \left|\sum_{i\in\mathcal{P}} \beta_i \left(\bar{X}_{Ti} - \bar{X}_{C'i}\right)\right| \\
&\le \sum_{i\in\mathcal{P}} |\beta_i| \cdot \left|\bar{X}_{Ti} - \bar{X}_{C'i}\right| \\
&\le \sum_{i\in\mathcal{P}} |\beta_i| \cdot \mathcal{I}_{\text{DOM:max}}(T,C') \\
&\le \|\boldsymbol{\beta}\|_1 \cdot \mathcal{I}_{\text{DOM:max}}(T,C'). \qquad \square
\end{aligned}$$

This bound is also tight, with the worst case occurring if the imbalance equals $\mathcal{I}_{\text{DOM:max}}(T,C')$ for each covariate. Estimates of $\boldsymbol{\beta}$ and α can be used to evaluate the bounds provided by Lemmas 3.4 and 3.5.

3.2.3 Balancing Additional Moments

While mean balance is necessary and sufficient for an unbiased estimate under Assumption 3.2, this is no longer the case when higher-degree terms are included in the control response function. However, as in linear regression, these higher-degree terms can also be incorporated into moment balance measures. To see how this can be accomplished, it will be useful to focus on the contribution of each term in the control response function $h^0(\cdot)$ to the bias $\mathcal{B}(T,C')$ separately rather than in aggregate, as this avoids the possibility of bias from terms canceling each other out (as in the case when $\boldsymbol{\beta} = \bar{\mathbf{X}}_T \times \bar{\mathbf{X}}_{C'}$).

For example, if the term $\gamma_i\left(x_i\right)^a$ appears in the control response function for some $i \in \mathcal{P}$ with $\gamma_i \in \mathbb{R}$ (fixed but unknown) and a known $a \in \mathbb{R}$, then its contribution to $\mathcal{B}(T,C')$ is

$$\frac{1}{|T|}\sum_{t\in T} \gamma_i \left(X_{ti}\right)^a - \frac{1}{|C'|}\sum_{c\in C'} \gamma_i \left(X_{ci}\right)^a = \gamma_i \left(\frac{1}{|T|}\sum_{t\in T} \left(X_{ti}\right)^a - \frac{1}{|C'|}\sum_{c\in C'} \left(X_{ci}\right)^a\right).$$

This contribution is zero if and only if $\sum_{t\in T}\left(X_{ti}\right)^a/|T| = \sum_{c\in C'}\left(X_{ci}\right)^a/|C'|$. Similarly, if the term $\gamma_{ij}\left(x_i\right)^a\left(x_j\right)^b$ appears in the control response function, then its contribution to $\mathcal{B}(T,C')$ is

$$\gamma_{ij}\left(\frac{1}{|T|}\sum_{t\in T}\left(X_{ti}\right)^a\left(X_{tj}\right)^b - \frac{1}{|C'|}\sum_{c\in C'}\left(X_{ci}\right)^a\left(X_{cj}\right)^b\right),$$

which is zero if and only if the two summations are equal. In both cases it is necessary to know the values of the exponents on the terms in order to assess their contributions to $\mathcal{B}(T,C')$. However, this is no different than if these terms were to be included in a linear regression model.

The above examples and proofs can be used to show that if the control response function is of the form

$$h^0(\mathbf{x}) \equiv \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \sum_{i\in\mathcal{P}} \gamma_i \left(x_i\right)^2 + \sum_{(i,j)\in\binom{\mathcal{P}}{2}} \gamma_{ij} x_i x_j + \alpha$$

for all $\mathbf{x} \in \mathcal{X}$, where $\binom{\mathcal{P}}{2}$ is the set of all unordered pairs of covariates, then $\mathcal{B}(T, C') = 0$ for all possible parameter values if and only if

$$\begin{aligned} \bar{X}_{Ti} &= \bar{X}_{C'i} && \forall\ i \in \mathcal{P}, \\ \frac{1}{|T|}\sum_{t\in T}\left(X_{ti}\right)^2 &= \frac{1}{|C'|}\sum_{c\in C'}\left(X_{ci}\right)^2 && \forall\ i \in \mathcal{P}, \\ \frac{1}{|T|}\sum_{t\in T} X_{ti}X_{tj} &= \frac{1}{|C'|}\sum_{c\in C'} X_{ci}X_{cj} && \forall\ (i,j) \in \binom{\mathcal{P}}{2}. \end{aligned} \tag{3.2}$$

A similar observation was made by Rosenbaum and Rubin (1985). The balance requirements in (3.2) were also considered by Zubizarreta (2012), though the relationship to bias was not discussed. The imbalance measure

$$\begin{aligned} \mathcal{I}_{\mathrm{DOM:2}}\left(T, C'\right) \equiv \mathcal{I}_{\mathrm{DOM}}\left(T, C'\right) &+ \sum_{i\in\mathcal{P}} \left| \frac{1}{|T|}\sum_{t\in T}\left(X_{ti}\right)^2 - \frac{1}{|C'|}\sum_{c\in C'}\left(X_{ci}\right)^2 \right| \\ &+ \sum_{(i_1,i_2)\in\binom{\mathcal{P}}{2}} \left| \frac{1}{|T|}\sum_{t\in T} X_{ti_1}X_{ti_2} - \frac{1}{|C'|}\sum_{c\in C'} X_{ci_1}X_{ci_2} \right|, \end{aligned}$$

provides one way to assess whether or not these conditions are met.

3.3 Assumptions for Distribution Balance

Whereas moment balance measures require assumptions about the specific terms in the control response function to provide guarantees on $\mathcal{B}(T, C')$, a stronger form of covariate balance such as *distribution balance* can be used to ensure that there is no contribution to $\mathcal{B}(T, C')$ from a large set of potential terms. Some assumptions regarding the functional form of the control response function are still necessary, however.

3.3.1 Marginal Distribution Balance

Marginal distribution balance measures are covariate balance measures that compare the marginal distributions of each covariate between two sets of units. Examples include $\mathcal{I}_{\mathrm{KS}}$ and $\mathcal{I}_{\mathrm{Diff}}$. These balance measures are appropriate if the control response function is separable.

Assumption 3.6. Assumption 3.1 holds with the additional requirement that $h^0(\mathbf{x}) \equiv \sum_{i\in\mathcal{P}} h_i^0(x_i)$ for all $\mathbf{x} \in \mathcal{X}$, where $h_i^0 : \mathbb{R} \to \mathbb{R}$ is a (bounded but unknown) function on the support of covariate i for each $i \in \mathcal{P}$.

Theorem 3.7. *Under Assumption 3.6, a control group $C' \subseteq C$ with $\widehat{F}_i(C', x) = \widehat{F}_i(T, x)$ for all $i \in \mathcal{P}$ and $x \in \mathcal{X}_i(T \cup C)$ (i.e., identical marginal empirical distributions) is both necessary and sufficient for $\mathcal{B}(T, C')$ to equal zero for all bounded functions h_i^0 for $i \in \mathcal{P}$.*

Proof. Sufficiency can be demonstrated for a C' that satisfies the required conditions by first reorganizing the terms in $\mathcal{B}(T, C')$:

$$\begin{aligned}\mathcal{B}(T, C') &= \frac{1}{|T|}\sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c) \\ &= \frac{1}{|T|}\sum_{t\in T}\sum_{i\in\mathcal{P}} h_i^0(X_{ti}) - \frac{1}{|C'|}\sum_{c\in C'}\sum_{i\in\mathcal{P}} h_i^0(X_{ci}) \\ &= \sum_{i\in\mathcal{P}}\left(\sum_{t\in T}\frac{h_i^0(X_{ti})}{|T|} - \sum_{c\in C'}\frac{h_i^0(X_{ci})}{|C'|}\right).\end{aligned} \tag{3.3}$$

For any covariate $i \in \mathcal{P}$, let $x_1 < x_2 < \ldots < x_k$ be the covariate values in $\mathcal{X}_i(T \cup C)$, and let $x_0 = x_1 - 1$ so that $\widehat{F}_i(T, x_0) = \widehat{F}_i(C, x_0) = 0$. Then

$$\begin{aligned}\sum_{t\in T}\frac{h_i^0(X_{ti})}{|T|} &= \sum_{j=1}^{k}\left(\sum_{t\in T:\ x_{j-1}<X_{ti}\le x_j}\frac{h_i^0(X_{ti})}{|T|}\right) \\ &= \sum_{j=1}^{k}\left(\sum_{t\in T:\ x_{j-1}<X_{ti}\le x_j}\frac{h_i^0(x_j)}{|T|}\right) \\ &= \sum_{j=1}^{k} h_i^0(x_j)\left(\widehat{F}_i(T, x_j) - \widehat{F}_i(T, x_{j-1})\right).\end{aligned}$$

An identical argument can be applied to show that

$$\sum_{c\in C'}\frac{h_i^0(X_{ci})}{|C'|} = \sum_{j=1}^{k} h_i^0(x_j)\left(\widehat{F}_i(C', x_j) - \widehat{F}_i(C', x_{j-1})\right).$$

From the fact that $\widehat{F}_i(T, x) = \widehat{F}_i(C', x)$ for all $x \in \mathcal{X}_i(T \cup C)$, it follows that

$$\begin{aligned}\sum_{t\in T}\frac{h_i^0(X_{ti})}{|T|} - \sum_{c\in C'}\frac{h_i^0(X_{ci})}{|C'|} &= \sum_{j=1}^{k} h_i^0(x_j)\left(\widehat{F}_i(T, x_j) - \widehat{F}_i(T, x_{j-1})\right) \\ &\quad - \sum_{j=1}^{k} h_i^0(x_j)\left(\widehat{F}_i(C', x_j) - \widehat{F}_i(C', x_{j-1})\right) \\ &= 0.\end{aligned}$$

Because this equality holds for all $i \in \mathcal{P}$, (3.3) simplifies to

$$\mathcal{B}(T, C') = \sum_{i \in \mathcal{P}} \left(\sum_{t \in T} \frac{h_i^0(X_{ti})}{|T|} - \sum_{c \in C'} \frac{h_i^0(X_{ci})}{|C'|} \right) = 0.$$

Necessity can be demonstrated by contrapositive. Consider a C' with $\widehat{F}_{i^*}(C', x^*) \neq \widehat{F}_{i^*}(T, x^*)$ for some $i^* \in \mathcal{P}$ and $x^* \in \mathcal{X}_{i^*}$. For the control response functions given by

$$h_i^0(x) = \begin{cases} 1 & \text{if } i = i^* \text{ and } x \leq x^* \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in \mathcal{P}$ and $x \in \mathcal{X}_i$, the control response function bias simplifies to

$$\begin{aligned} \mathcal{B}(T, C') &= \frac{1}{|T|} \sum_{t \in T} \sum_{i \in \mathcal{P}} h_i^0(X_{ti}) - \frac{1}{|C'|} \sum_{c \in C'} \sum_{i \in \mathcal{P}} h_i^0(X_{ci}) \\ &= \frac{1}{|T|} \sum_{t \in T} h_{i^*}^0(X_{ti^*}) - \frac{1}{|C'|} \sum_{c \in C'} h_{i^*}^0(X_{ci^*}) \\ &= \frac{1}{|T|} \sum_{t \in T\colon\ X_{ti^*} \leq x^*} 1 - \frac{1}{|C'|} \sum_{c \in C'\colon\ X_{ci^*} \leq x^*} 1 \\ &= \widehat{F}_{i^*}(T, x^*) - \widehat{F}_{i^*}(C', x^*), \end{aligned}$$

which does not equal zero by choice of i^* and x^*. □

As with Theorem 3.3, the "for all" quantifier is important, because there are functions h_i^0 for which $\mathcal{B}(T, C') = 0$ for a $C' \subseteq C$ without $\widehat{F}_i(C', x) = \widehat{F}_i(T, x)$ at every $x \in \mathcal{X}_i(T \cup C)$ and $i \in \mathcal{P}$. For example, if $h_i^0(x) = a_i$ for all $x \in \mathbb{R}$ and $i \in \mathcal{P}$, where each a_i is a constant, then $\mathcal{B}(T, C') = 0$ for all $C' \subseteq C$. Such functions are the exception, and in general the distribution imbalance measured by $\widehat{F}_i(\cdot)$ coincides with $\mathcal{B}(T, C')$ being nonzero under Assumption 3.6.

The role of Theorem 3.7 is to motivate methods for balancing the marginal distributions of the covariates in C' and T. By focusing on marginal balance on each covariate separately instead of joint balance on all covariates together, which leads to exact matches if achieved, marginal balancing methods avoid the difficulties due to sparsity caused by the exponential growth in volume as the number of covariates increases. One such marginal balancing method is BOSS with $\mathcal{I}_{\text{KS}}$, which identifies a $C' \subseteq C$ with $\widehat{F}_i(C', x) = \widehat{F}_i(T, x)$ for all $i \in \mathcal{P}$ and $x \in \mathcal{X}_i(T \cup C)$, if such a control group exists.

3.3.2 Joint Distribution Balance

Distribution balance can also be assessed on higher-order distributions (e.g., pairwise joints) by using covariate clusters and $\widehat{F}_D(\cdot)$. The following theorem establishes the relationship between joint distribution balance and the control response function bias.

Theorem 3.8. *For a covariate cluster $D \equiv \{i_1, i_2, \ldots, i_k\} \subseteq \mathcal{P}$, if $\widehat{F}_D(T, \mathbf{x}) = \widehat{F}_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$, then any function $h_D^0(x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ that appears in the control response function contributes nothing to $\mathcal{B}(T, C')$.*

To prove Theorem 3.8, several additional definitions that depend on the samples T and C are needed. For each covariate $i \in \mathcal{P}$, let $\rho_i : \mathbb{R}^p \to \mathbb{R}$ denote the *predecessor* function for covariate i, defined as

$$\rho_i(\mathbf{x}) \equiv \begin{cases} x_i - 1 & \text{if } x_i = \min\{X_{ui} : u \in T \cup C\} \\ \max\{X_{ui} : u \in T \cup C,\ X_{ui} < x_i\} & \text{otherwise.} \end{cases}$$

Also, let $\nu_i : \mathbb{R}^p \to \mathbb{R}$ denote the *predecessor gap*, defined as $\nu_i(\mathbf{x}) \equiv x_i - \rho_i(\mathbf{x})$, and let $\mathbf{e}_i \in \mathbb{R}^p$ be the unit vector with value 1 for coordinate i and 0 otherwise. Additionally, for any $S \subseteq T \cup C$ and $x \in \mathbb{R}$, let

$$G_i(S, x) \equiv \{u \in S : X_{ui} > x\}$$

be the set of units in S whose values for covariate i are strictly greater than x.

For any covariate cluster $D \subseteq \mathcal{P}$, set of units $S \subseteq T \cup C$, and $\mathbf{x} \in \mathbb{R}^p$, let

$$\zeta_D(S, \mathbf{x}) \equiv \left|\left\{u \in S : \mathbf{P}^D \mathbf{X}_u = \mathbf{P}^D \mathbf{x}\right\}\right| / |S|$$

be the proportion of units in S whose covariate values equal the values in $\mathbf{x}$ for the covariates in D, and let

$$L_D(S, \mathbf{x}) \equiv \left\{u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x}\right\}$$

be the set of units in S whose covariate values are no greater than the values in $\mathbf{x}$ for all covariates in D.

Lemma 3.9. *For any $S \subseteq T \cup C$, covariate cluster $D \subseteq \mathcal{P}$, and $\mathbf{x} \in \mathbb{R}^p$,*

$$\zeta_D(S, \mathbf{x}) = \frac{1}{|S|} \left| L_D(S, \mathbf{x}) \cap \left(\bigcap_{i \in D} G_i(S, \rho_i(\mathbf{x})) \right) \right| .$$

Proof.

$$\begin{aligned}
\zeta_D(S,\mathbf{x}) &= \frac{1}{|S|}\left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u = \mathbf{P}^D\mathbf{x}\right\}\right| \\
&= \frac{1}{|S|}\left|\{u \in S : \rho_i(\mathbf{x}) < X_{ui} \leq x_i \ \forall\ i \in D\}\right| \\
&= \frac{1}{|S|}\left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u \leq \mathbf{P}^D\mathbf{x},\ X_{ui} > \rho_i(\mathbf{x}) \ \forall\ i \in D\right\}\right| \\
&= \frac{1}{|S|}\left|L_D(S,\mathbf{x}) \cap \left(\bigcap_{i\in D} G_i(S,\rho_i(\mathbf{x}))\right)\right|. \qquad \square
\end{aligned}$$

Lemma 3.10. *For any $S \subseteq T \cup C$, covariate cluster $D \subseteq \mathcal{P}$, and $\mathbf{x} \in \mathbb{R}^p$,*

$$\left|L_D(S,\mathbf{x}) \cap \left(\bigcap_{i\in D'} G_i(S,\rho_i(\mathbf{x}))\right)\right| = \sum_{D''\in 2^{D'}} (-1)^{|D''|}\left|L_D\left(S,\mathbf{x} - \sum_{i\in D''}\nu_i(\mathbf{x})\mathbf{e}_i\right)\right|$$

for all $D' \subseteq D$, where $2^{D'}$ is the set of all subsets of D' including the empty set and D'.

Proof. This can be shown using induction on the size of D'.

Base Case: $|D'| = 1$

Let i be the covariate in D'. Then

$$\begin{aligned}
|L_D(S,\mathbf{x}) \cap G_i(S,\rho_i(\mathbf{x}))| &= \left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u \leq \mathbf{P}^D\mathbf{x}\right\} \cap \{u \in S : X_{ui} > \rho_i(\mathbf{x})\}\right| \\
&= \left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u \leq \mathbf{P}^D\mathbf{x},\ X_{ui} > \rho_i(\mathbf{x})\right\}\right| \\
&= \left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u \leq \mathbf{P}^D\mathbf{x}\right\}\right| - \left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u \leq \mathbf{P}^D\mathbf{x},\ X_{ui} \leq \rho_i(\mathbf{x})\right\}\right| \\
&= |L_D(S,\mathbf{x})| - \left|\left\{u \in S : \mathbf{P}^D\mathbf{X}_u \leq \mathbf{P}^D\mathbf{x} - \nu_i(\mathbf{x})\mathbf{e}_i\right\}\right| \\
&= |L_D(S,\mathbf{x})| - |L_D(S,\mathbf{x} - \nu_i(\mathbf{x})\mathbf{e}_i)| \\
&= \sum_{D''\in 2^{D'}} (-1)^{|D''|}\left|L_D\left(S,\mathbf{x} - \sum_{i\in D''}\nu_i(\mathbf{x})\mathbf{e}_i\right)\right|,
\end{aligned}$$

which establishes the base case.

Inductive Step: $|D'| > 1$

Assume that the desired result holds for all D'' with $|D''| < |D'|$. Let i^* be an arbitrary covariate in D', and let $D^* \equiv D' \setminus \{i^*\}$. Then

$$\begin{aligned}
&\left| L_D(S,\mathbf{x}) \cap \left(\bigcap_{i \in D'} G_i(S, \rho_i(\mathbf{x})) \right) \right| \\
&\qquad = \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x},\ X_{ui} > \rho_i(\mathbf{x}) \ \forall\ i \in D' \right\} \right| \\
&\qquad = \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x},\ x_{ui^*} > \rho_{i^*}(\mathbf{x}),\ X_{ui} > \rho_i(\mathbf{x}) \ \forall\ i \in D^* \right\} \right| \\
&\qquad = \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x},\ X_{ui} > \rho_i(\mathbf{x}) \ \forall\ i \in D^* \right\} \right| \\
&\qquad\quad - \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x},\ x_{ui^*} \leq \rho_{i^*}(\mathbf{x}),\ X_{ui} > \rho_i(\mathbf{x}) \ \forall\ i \in D^* \right\} \right| \\
&\qquad = \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x},\ X_{ui} > \rho_i(\mathbf{x}) \ \forall\ i \in D^* \right\} \right| \\
&\qquad\quad - \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \mathbf{x} - \nu_{i^*}(\mathbf{x})\mathbf{e}_{i^*},\ X_{ui} > \rho_i(\mathbf{x} - \nu_{i^*}(\mathbf{x})\mathbf{e}_{i^*}) \ \forall\ i \in D^* \right\} \right| \\
&\qquad = \left| L_D(S,\mathbf{x}) \cap \left(\bigcap_{i \in D^*} G_i(S, \rho_i(\mathbf{x})) \right) \right| - \left| L_D(S,\mathbf{x}^*) \cap \left(\bigcap_{i \in D^*} G_i(S, \rho_i(\mathbf{x}^*)) \right) \right|,
\end{aligned} \tag{3.4}$$

where $\mathbf{x}^* \equiv \mathbf{x} - \nu_{i^*}(\mathbf{x})\mathbf{e}_{i^*}$. The inductive hypothesis shows that

$$\begin{aligned}
\left| L_D(S,\mathbf{x}) \cap \left(\bigcap_{i \in D^*} G_i(S, \rho_i(\mathbf{x})) \right) \right| &= \sum_{D'' \in 2^{D^*}} (-1)^{|D''|} \left| L_D \left(S, \mathbf{x} - \sum_{i \in D''} \nu_i(\mathbf{x})\mathbf{e}_i \right) \right| \\
&= \sum_{D'' \in 2^{D'}:\ i^* \notin D''} (-1)^{|D''|} \left| L_D \left(S, \mathbf{x} - \sum_{i \in D''} \nu_i(\mathbf{x})\mathbf{e}_i \right) \right|
\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}
\left| L_D(S,\mathbf{x}^*) \cap \left(\bigcap_{i \in D^*} G_i(S, \rho_i(\mathbf{x}^*)) \right) \right| &= \sum_{D'' \in 2^{D^*}} (-1)^{|D''|} \left| L_D \left(S, \mathbf{x}^* - \sum_{i \in D''} \nu_i(\mathbf{x})\mathbf{e}_i \right) \right| \\
&= \sum_{D'' \in 2^{D^*}} (-1)^{|D''|} \left| L_D \left(S, \mathbf{x} - \nu_{i^*}(\mathbf{x})\mathbf{e}_{i^*} - \sum_{i \in D''} \nu_i(\mathbf{x})\mathbf{e}_i \right) \right| \\
&= (-1) \cdot \sum_{D'' \in 2^{D'}:\ i^* \in D''} (-1)^{|D''|} \left| L_D \left(S, \mathbf{x} - \sum_{i \in D''} \nu_i(\mathbf{x})\mathbf{e}_i \right) \right|.
\end{aligned} \tag{3.6}$$

Then (3.5) and (3.6) can be combined with (3.4) to show that

$$\left| L_D(S,\mathbf{x}) \cap \left(\bigcap_{i \in D'} G_i(S, \rho_i(\mathbf{x})) \right) \right| = \sum_{D'' \in 2^{D'}} (-1)^{|D''|} \left| L_D \left(S, \mathbf{x} - \sum_{i \in D''} \nu_i(\mathbf{x})\mathbf{e}_i \right) \right|,$$

which completes the proof. □

Lemma 3.11. *For any $S \subseteq T \cup C$, covariate cluster $D \subseteq \mathcal{P}$, and $\mathbf{x} \in \mathbb{R}^p$,*

$$\zeta_D(S, \mathbf{x}) = \sum_{D' \in 2^D} (-1)^{|D'|} \cdot \widehat{F}_D \left(S, \mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right).$$

Proof. Follows from Lemmas 3.9 and 3.10:

$$\begin{aligned}
\zeta_D(S, \mathbf{x}) &= \frac{1}{|S|} \left| L_D(S, \mathbf{x}) \cap \left(\bigcap_{i \in D} G_i(S, \rho_i(\mathbf{x})) \right) \right| \\
&= \frac{1}{|S|} \left(\sum_{D' \in 2^D} (-1)^{|D'|} \left| L_D \left(S, \mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right) \right| \right) \\
&= \frac{1}{|S|} \left(\sum_{D' \in 2^D} (-1)^{|D'|} \left| \left\{ u \in S : \mathbf{P}^D \mathbf{X}_u \leq \mathbf{P}^D \left(\mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right) \right\} \right| \right) \\
&= \frac{1}{|S|} \left(\sum_{D' \in 2^D} (-1)^{|D'|} \cdot |S| \cdot \widehat{F}_D \left(S, \mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right) \right) \\
&= \sum_{D' \in 2^D} (-1)^{|D'|} \cdot \widehat{F}_D \left(S, \mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right). \qquad \square
\end{aligned}$$

Theorem 3.12. *If T and C' satisfy $\widehat{F}_D(T, \mathbf{x}) = \widehat{F}_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$, then $\zeta_D(T, \mathbf{x}) = \zeta_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$.*

Proof. Let $C' \subseteq C$ be a control group that satisfies $\widehat{F}_D(T, \mathbf{x}) = \widehat{F}_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$. For an arbitrary $\mathbf{x} \in \mathcal{X}_D(T \cup C)$, Lemma 3.11 yields

$$\begin{aligned}
\zeta_D(T, \mathbf{x}) &= \sum_{D' \in 2^D} (-1)^{|D'|} \cdot \widehat{F}_D \left(T, \mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right) \\
&= \sum_{D' \in 2^D} (-1)^{|D'|} \cdot \widehat{F}_D \left(C', \mathbf{x} - \sum_{i \in D'} \nu_i(\mathbf{x}) \mathbf{e}_i \right) = \zeta_D(C', \mathbf{x}). \qquad \square
\end{aligned}$$

Proof of Theorem 3.8. For the function $h_D^0(\cdot)$, the treatment units in T contribute the following to $\mathcal{B}(T, C')$:

$$\begin{aligned}
\frac{1}{|T|} \sum_{t \in T} h_D^0(X_{ti_1}, \ldots, X_{ti_k}) &= \frac{1}{|T|} \sum_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} \left(\sum_{t \in T:\ \mathbf{P}^D \mathbf{X}_t = \mathbf{x}} h_D^0(X_{ti_1}, \ldots, X_{ti_k}) \right) \\
&= \frac{1}{|T|} \sum_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} \left(h_D^0(x_{i_1}, \ldots, x_{i_k}) \sum_{t \in T:\ \mathbf{P}^D \mathbf{x}_t = \mathbf{x}} 1 \right) \qquad (3.7) \\
&= \sum_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} h_D^0(x_{i_1}, \ldots, x_{i_k}) \cdot \zeta_D(T, \mathbf{x}).
\end{aligned}$$

Similarly, control units in C' contribute

$$\frac{1}{|C'|}\sum_{c\in C'} h_D^0(X_{ci_1},\ldots,X_{ci_k}) = \sum_{\mathbf{x}\in\mathcal{X}_D(T\cup C)} h_D^0(x_{i_1},\ldots,x_{i_k})\cdot\zeta_D(C',\mathbf{x}). \tag{3.8}$$

The total contribution of $h_D^0(\cdot)$ to $\mathcal{B}(T,C')$ is the difference between (3.7) and (3.8):

$$\begin{aligned}\frac{1}{|T|}\sum_{t\in T} h_D^0(X_{ti_1},\ldots,X_{ti_k}) - \frac{1}{|C'|}\sum_{c\in C'} h_D^0(X_{ci_1},\ldots,X_{ci_k})&\\ = \sum_{\mathbf{x}\in\mathcal{X}_D(T\cup C)} h_D^0(x_{i_1},\ldots,x_{i_k})\cdot\left(\zeta_D(T,\mathbf{x})-\zeta_D(C',\mathbf{x})\right).&\end{aligned} \tag{3.9}$$

By Theorem 3.12, $\zeta_D(T,\mathbf{x}) = \zeta_D(C',\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}_D(T\cup C)$. As a result, the summation in (3.9) equals zero, which ensures that $h_D^0(\cdot)$ makes no contribution to $\mathcal{B}(T,C')$. □

Corollary 3.13. *Distribution balance on the cluster of all covariates $D=\mathcal{P}$ (i.e., $\widehat{F}_D(T,\mathbf{x}) = \widehat{F}_D(C',\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}_D(T\cup C)$) ensures that $\mathcal{B}(T,C')=0$ under Assumption 3.1.*

Corollary 3.13 shows that distribution balance (as measured by $\widehat{F}_D$) on the full joint distribution of all covariates provides the same guarantee on $\mathcal{B}(T,C')$ that exact matching provides under the strong ignorability assumption. In fact, as noted in Chapter 2, if T and C' possess this level of distribution balance, then an exact matched-pair sample can be constructed using the control units in C'.

3.3.3 Coarsened Distribution Balance

Both Theorems 3.7 and 3.8 rely on T and C' having identical empirical cumulative distribution functions for the appropriate covariates. As Theorem 3.12 shows, this essentially requires exactly matching units on a subset of covariates. Because of this, achieving this form of covariate balance is likely to be difficult in practice, particularly if continuous covariates are involved. However, as Iacus et al. (2012) note, one way to reduce the difficulty of exact matching on continuous covariates is to *coarsen* or discretize the covariate values using histogram bins. Nikolaev et al. (2013) propose the imbalance measure $\mathcal{I}_{\chi^2}$ for this purpose; $\mathcal{I}_{\text{Diff}}$ and $\mathcal{I}_{\text{Diff}^2}$ are reasonable alternatives.

When a control group C' has no imbalance with respect to T on a coarsened distribution imbalance measure, the corresponding empirical joint distributions of the covariates are approximately equal for the two sets of units. The quality of this approximation depends on the granularity of the histogram bins used to coarsen the covariate values. This level of covariate balance is generally insufficient to remove all bias from the estimator (2.17), but under certain assumptions the residual bias can be bounded. The following assumptions and results apply to an arbitrary covariate cluster $D \equiv \{i_1, i_2, \ldots, i_k\} \subseteq \mathcal{P}$.

Assumption 3.14. The function $h^0_D(x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ is Lipschitz continuous with Lipschitz constant no greater than K_D for the Manhattan distance metric.

For each $i \in D$, label the histogram bin boundaries as

$$\min_{u \in T \cup C} X_{ui} \equiv b_{i0} < b_{i1} < \ldots < b_{in_i} \equiv \max_{u \in T \cup C} X_{ui},$$

let $\bar{b}_{ij} \equiv (b_{ij} - b_{i,j-1})/2$ be the midpoint of bin $j \in N_i$, and let $m_i \equiv \max_{j \in N_i} \{b_{ij} - b_{i,j-1}\}$ be the maximum width of any bin for covariate i. Under Assumption 3.14, for any $j \in N_D$ and unit $u \in B_{Dj}$, the difference between the function $h^0_D(\cdot)$ evaluated at $\mathbf{X}_u$ and at the center of the bin is bounded by

$$\left| h^0_D(X_{u,i_1}, \ldots, X_{u_1,i_k}) - h^0_D(\bar{b}_{i_i,j_1}, \ldots, \bar{b}_{i_k,j_k}) \right| \leq K_D \sum_{l=1}^{k} \left| X_{u,i_l} - \bar{b}_{i_l,j_l} \right| \leq \frac{K_D}{2} \sum_{i \in D} m_i. \tag{3.10}$$

Theorem 3.15. *Under Assumption 3.14, if C' and T satisfy $\eta_{Dj}(T) = \eta_{Dj}(C')$ for all $j \in N_D$, then the contribution of the function $h^0_D(x_{i_1}, \ldots, x_{i_k})$ to $\mathcal{B}(T, C')$ is bounded by $K_D \sum_{i \in D} m_i$.*

Proof.

$$\begin{aligned}
&\left| \frac{1}{|T|} \sum_{t \in T} h^0_D(X_{ti_1}, \ldots, X_{ti_k}) - \frac{1}{|C'|} \sum_{c \in C'} h^0_D(X_{ci_1}, \ldots, X_{ci_k}) \right| \\
&\qquad = \left| \sum_{j \in N_D} \sum_{t \in T \cap B_{Dj}} \frac{h^0_D(X_{ti_1}, \ldots, X_{ti_k})}{|T|} - \sum_{j \in N_D} \sum_{c \in C' \cap B_{Dj}} \frac{h^0_D(X_{ci_1}, \ldots, X_{ci_k})}{|C'|} \right| \\
&\qquad = \left| \sum_{j \in N_D} \left(\sum_{t \in T \cap B_{Dj}} \frac{h^0_D(X_{ti_1}, \ldots, X_{ti_k})}{|T|} - \sum_{c \in C' \cap B_{Dj}} \frac{h^0_D(X_{ci_1}, \ldots, X_{ci_k})}{|C'|} \right) \right| \\
&\qquad \leq \sum_{j \in N_D} \left| \sum_{t \in T \cap B_{Dj}} \frac{h^0_D(X_{ti_1}, \ldots, X_{ti_k})}{|T|} - \sum_{c \in C' \cap B_{Dj}} \frac{h^0_D(X_{ci_1}, \ldots, X_{ci_k})}{|C'|} \right|.
\end{aligned} \tag{3.11}$$

For each $j \equiv (j_1, j_2, \ldots, j_k) \in N_D$, the associated term in the summation in (3.11) can be expanded to yield

$$\begin{aligned}
&\left|\sum_{t\in T\cap B_{Dj}} \frac{h_D^0(X_{ti_1},\ldots,X_{ti_k})}{|T|} - \sum_{c\in C'\cap B_{Dj}} \frac{h_D^0(X_{ci_1},\ldots,X_{ci_k})}{|C'|}\right| \\
&\quad= \left|\sum_{t\in T\cap B_{Dj}} \frac{h_D^0(X_{ti_1},\ldots,X_{ti_k})}{|T|} - \eta_{Dj}(T)h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big)\right. \\
&\qquad \left. + \eta_{Dj}(C')h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big) - \sum_{c\in C'\cap B_{Dj}} \frac{h_D^0(X_{ci_1},\ldots,X_{ci_k})}{|C'|}\right| \\
&\quad= \left|\sum_{t\in T\cap B_{Dj}} \frac{h_D^0(X_{ti_1},\ldots,X_{ti_k})}{|T|} - |T\cap B_{Dj}|\left(\frac{h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big)}{|T|}\right)\right. \\
&\qquad \left. + |C'\cap B_{Dj}|\left(\frac{h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big)}{|C'|}\right) - \sum_{c\in C'\cap B_{Dj}} \frac{h_D^0(X_{ci_1},\ldots,X_{ci_k})}{|C'|}\right| \qquad (3.12)\\
&\quad= \left|\frac{1}{|T|}\sum_{t\in T\cap B_{Dj}} \left(h_D^0(X_{ti_1},\ldots,X_{ti_k}) - h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big)\right)\right. \\
&\qquad \left. + \frac{1}{|C'|}\sum_{c\in C'\cap B_{Dj}} \left(h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big) - h_D^0(X_{ci_1},\ldots,X_{ci_k})\right)\right| \\
&\quad\le \frac{1}{|T|}\sum_{t\in T\cap B_{Dj}} \left|h_D^0(X_{ti_1},\ldots,X_{ti_k}) - h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big)\right| \\
&\qquad + \frac{1}{|C'|}\sum_{c\in C'\cap B_{Dj}} \left|h_D^0(X_{ci_1},\ldots,X_{ci_k}) - h_D^0\big(\bar{b}_{i_1,j_1},\ldots,\bar{b}_{i_k,j_k}\big)\right|.
\end{aligned}$$

The results from (3.10) and (3.12) combine with (3.11) to yield

$$\begin{aligned}
&\left|\frac{1}{|T|}\sum_{t\in T} h_D^0(X_{ti_1},\ldots,X_{ti_k}) - \frac{1}{|C'|}\sum_{c\in C'} h_D^0(X_{ci_1},\ldots,X_{ci_k})\right| \\
&\qquad\le \sum_{j\in N_D}\left(\frac{1}{|T|}\sum_{t\in T\cap B_{Dj}}\left(\frac{K_D}{2}\sum_{i\in D} m_i\right) + \frac{1}{|C'|}\sum_{c\in C'\cap B_{Dj}}\left(\frac{K_D}{2}\sum_{i\in D} m_i\right)\right) \\
&\qquad= \left(\frac{K_D}{2}\sum_{i\in D} m_i\right)\left(\sum_{j\in N_D}\sum_{t\in T\cap B_{Dj}}\frac{1}{|T|} + \sum_{j\in N_D}\sum_{c\in C'\cap B_{Dj}}\frac{1}{|C'|}\right) \\
&\qquad= K_D\sum_{i\in D} m_i. \qquad \square
\end{aligned}$$

Theorem 3.15 reveals how the width of the histogram bins affects $\mathcal{B}(T, C')$. Any set of histogram bins can be refined by subdividing each of the bins for the marginal distributions into two equal bins with half the width of the original bin. This subdivision then cuts the bound on $\mathcal{B}(T, C')$ in half. Theorem 3.15

is a modified version of Theorem 1 from Nikolaev et al. (2013), which shows that as the granularity of the histogram bins improves (i.e., the maximum width shrinks), $\mathcal{B}(T, C')$ converges to zero as long as C' continues to satisfy $\eta_{Dj}(T) = \eta_{Dj}(C')$ for all $j \in N_D$ for each set of refined histogram bins.

3.4 The Balance Hierarchy

The results for moment balance and distribution balance are summarized in Table 3.1, which lists potential terms in the control response function along with the balance requirements that are necessary and sufficient to ensure that those terms contribute nothing to $\mathcal{B}(T, C')$. As the terms become more (less) specific, the level of balance required to ensure no contribution to $\mathcal{B}(T, C')$ decreases (increases).

Table 3.1: Levels of covariate balance that are required to remove bias from various terms in the control response function.

Term	Balance Required
$\beta_i x_i,\ i \in \mathcal{P}$	$\frac{1}{\lvert T\rvert}\sum_{t\in T} X_{ti} = \frac{1}{\lvert C'\rvert}\sum_{c\in C'} X_{ci}$
$\gamma_{a_1,a_2,\ldots,a_p}\prod_{i\in\mathcal{P}} (x_i)^{a_i}$	$\frac{1}{\lvert T\rvert}\sum_{t\in T}\prod_{i\in\mathcal{P}} (X_{ti})^{a_i} = \frac{1}{\lvert C'\rvert}\sum_{c\in C'}\prod_{i\in\mathcal{P}} (X_{ci})^{a_i}$
$h_i^0(x_i),\ i \in \mathcal{P}$	$\widehat{F}_i(T,x) = \widehat{F}_i(C',x)\ \forall\ x \in \mathcal{X}_i(T \cup C)$
$h_D^0(x_{i_1}, x_{i_2}, \ldots, x_{i_k}),\ D \subseteq \mathcal{P}$	$\widehat{F}_D(T,\mathbf{x}) = \widehat{F}_D(C',\mathbf{x})\ \forall\ \mathbf{x} \in \mathcal{X}_D(T \cup C)$
$h^0(\mathbf{x})$	$\widehat{F}_{\mathcal{P}}(T,\mathbf{x}) = \widehat{F}_{\mathcal{P}}(C',\mathbf{x})\ \forall\ \mathbf{x} \in \mathcal{X}_{\mathcal{P}}(T \cup C)$

The relationships between moment balance, marginal distribution balance, and joint distribution balance are formally established in the next several results.

Lemma 3.16. *For a covariate $i \in \mathcal{P}$, if T and C' satisfy $\widehat{F}_i(T, x) = \widehat{F}_i(C', x)$ for all $x \in \mathcal{X}_i(T \cup C)$, then $\sum_{t\in T} (X_{ti})^a / |T| = \sum_{c\in C'} (X_{ci})^a / |C'|$ for any $a \in \mathbb{R}$.*

Proof. Let $x_1 < x_2 < \ldots < x_k$ be the covariate values in $\mathcal{X}_i(T \cup C)$, with $x_0 = x_1 - 1$. Then

$$\frac{1}{|T|}\sum_{t\in T} (X_{ti})^a = \sum_{j=1}^{k}\left(\sum_{\substack{t\in T:\\ x_{j-1}<X_{ti}\le x_j}} \frac{(X_{ti})^a}{|T|}\right) = \sum_{j=1}^{k} x_j^a \left(\widehat{F}_i(T, x_j) - \widehat{F}_i(T, x_{j-1})\right).$$

Similarly,

$$\frac{1}{|C'|}\sum_{c\in C'} (X_{ci})^a = \sum_{j=1}^{k} x_j^a \left(\widehat{F}_i(C', x_j) - \widehat{F}_i(C', x_{j-1})\right).$$

The desired result follows by applying the fact that $\widehat{F}_i(T, x) = \widehat{F}_i(C', x)$ for all $x \in \mathcal{X}_i(T \cup C)$. □

Lemma 3.17. *For a covariate cluster $D = \{i_1, i_2, \ldots, i_k\}$, if T and C' satisfy $\widehat{F}_D(T, \mathbf{x}) = \widehat{F}_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$, then*

$$\frac{1}{|T|} \sum_{t \in T} \prod_{j=1}^{k} \left(X_{ti_j}\right)^{a_j} = \frac{1}{|C'|} \sum_{c \in C'} \prod_{j=1}^{k} \left(X_{ci_j}\right)^{a_j}$$

for any $a_j \in \mathbb{R}$, $j \in \{1, 2, \ldots, k\}$.

Proof. Let $a_j \in \mathbb{R}$, $j \in \{1, 2, \ldots, k\}$ be arbitrary constants. Then

$$\begin{aligned} \frac{1}{|T|} \sum_{t \in T} \prod_{j=1}^{k} \left(X_{ti_j}\right)^{a_j} &= \frac{1}{|T|} \sum_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} \left(\sum_{t \in T:\ \mathbf{P}^D \mathbf{X}_t = \mathbf{x}} \left(\prod_{j=1}^{k} \left(X_{ti_j}\right)^{a_j} \right) \right) \\ &= \sum_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} \left(\prod_{j=1}^{k} \left(x_{i_j}\right)^{a_j} \right) \cdot \zeta_D(T, \mathbf{x}). \end{aligned}$$

A similar argument establishes that

$$\frac{1}{|C'|} \sum_{c \in C'} \prod_{j=1}^{k} \left(X_{ci_j}\right)^{a_j} = \sum_{\mathbf{x} \in \mathcal{X}_D(T \cup C)} \left(\prod_{j=1}^{k} \left(x_{i_j}\right)^{a_j} \right) \cdot \zeta_D(C', \mathbf{x}).$$

Theorem 3.12 guarantees that $\zeta_D(T, \mathbf{x}) = \zeta_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$, so the result follows. □

Lemma 3.18. *For a covariate cluster $D \subseteq \mathcal{P}$, if $\widehat{F}_D(T, \mathbf{x}) = \widehat{F}_D(C', \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_D(T \cup C)$, then $\widehat{F}_{D'}(T, \mathbf{x}') = \widehat{F}_{D'}(C', \mathbf{x}')$ for all $\mathbf{x}' \in \mathcal{X}_{D'}(T \cup C)$ for any sub-cluster $D' \subset D$.*

Proof. Let $D' \subset D$ be an arbitrary sub-cluster of D. For an arbitrary $\mathbf{x}' \in \mathcal{X}_{D'}(T \cup C)$, define $\mathbf{x}^*$ as

$$x_i^* = \begin{cases} x_i' & \text{if } i \in D' \\ \max \mathcal{X}_i(T \cup C) & \text{if } i \in D \setminus D' \\ 0 & \text{otherwise.} \end{cases}$$

By construction, $\mathbf{x}^* \in \mathcal{X}_D(T \cup C)$. Then

$$\begin{aligned}
\widehat{F}_{D'}(T, \mathbf{x}') &= \left|\left\{t \in T : \mathbf{P}^{D'}\mathbf{X}_t \leq \mathbf{P}^{D'}\mathbf{x}'\right\}\right| / |T| \\
&= \left|\{t \in T : X_{ti} \leq x_i' \ \forall\ i \in D'\}\right| / |T| \\
&= \left|\{t \in T : X_{ti} \leq x_i' \ \forall\ i \in D',\ X_{ti} \leq x_i^* \ \forall\ i \in D \setminus D'\}\right| / |T| \\
&= \left|\{t \in T : X_{ti} \leq x_i^* \ \forall\ i \in D\}\right| / |T| \\
&= \left|\{t \in T : \mathbf{P}^{D}\mathbf{X}_t \leq \mathbf{P}^{D}\mathbf{x}^*\}\right| / |T| \\
&= \widehat{F}_D(T, \mathbf{x}^*).
\end{aligned}$$

A similar argument shows that $\widehat{F}_{D'}(C', \mathbf{x}') = \widehat{F}_D(C', \mathbf{x}^*)$. So

$$\widehat{F}_{D'}(T, \mathbf{x}') = \widehat{F}_D(T, \mathbf{x}^*) = \widehat{F}_D(C', \mathbf{x}^*) = \widehat{F}_{D'}(C', \mathbf{x}'),$$

which completes the proof. □

Lemma 3.16 establishes that distribution balance is stronger than moment balance because distribution balance simultaneously balances all moments of the covariate. Lemma 3.17 extends this result to covariate clusters to show that distribution balance on a cluster D also balances any multivariate moment of covariates in D. Finally, Lemma 3.18 establishes how joint distribution balance for a cluster D also ensures distribution balance for joint distributions defined on subsets of D.

The results of Lemmas 3.16, 3.17, and 3.18 create a *balance hierarchy* that represents the general relationships between various forms of balance. In the balance hierarchy, stronger forms of balance subsume weaker ones. This hierarchy is a partial ordering because some forms of balance are incomparable (e.g., marginal distribution balance on covariate i is incomparable to marginal distribution balance on covariate j). Because there is a one-to-one correspondence between covariate balance levels and the terms in the control response function (as indicated by Table 3.1), the balance hierarchy implies a relationship between control response function terms. This relationship is illustrated in Figure 3.1 for three covariates (only low-degree terms are included in order to simplify the figure). Arrows indicate if one term is subsumed by another. For example, the arrow from the term $h_1^0(x_1)$ to the term $\beta_1 x_1$ indicates that the balance necessary to remove the contribution of $h_1^0(x_1)$ to $\mathcal{B}(T, C')$ is also sufficient to remove the contribution of $\beta_1 x_1$.

Figure 3.1 illustrates how regression and matching methods differ in their approach to observational data. Regression methods are typically used in a "bottom-up" fashion by including the low-degree terms

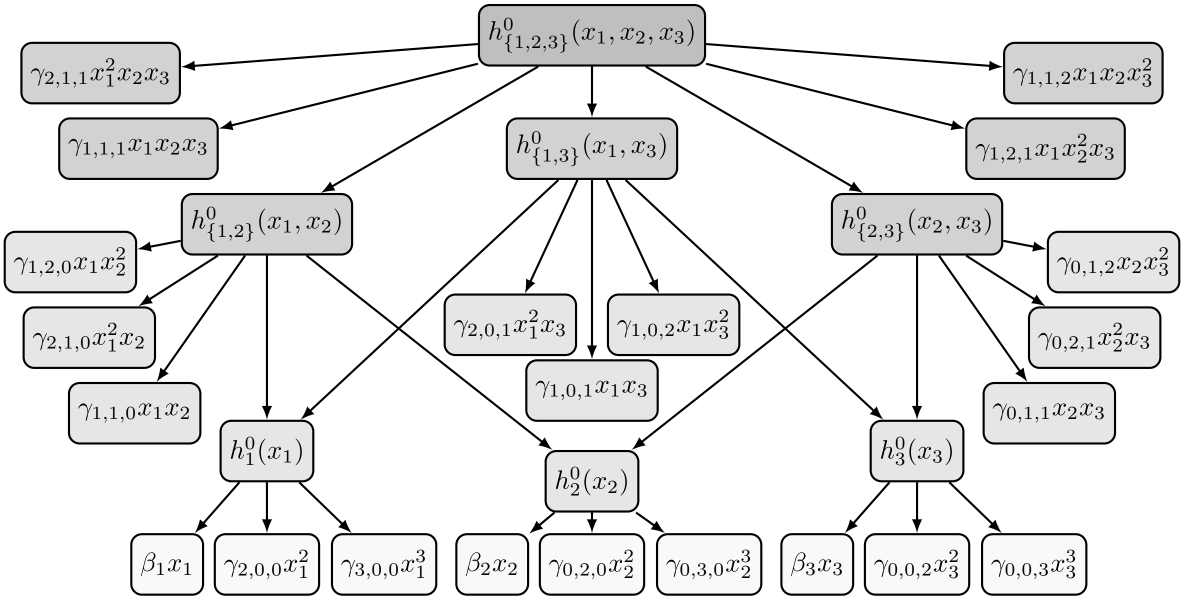


Figure 3.1: The balance hierarchy for three covariates, $\mathcal{P} = \{1, 2, 3\}$. All possible one, two, and three-degree terms and several four-degree terms are included. An arrow from one term to another indicates that the first term includes the second term as a special case. Shading indicates approximate ordering of terms, with darker terms requiring higher levels of covariate balance.

(the leaf nodes in the balance hierarchy) into a model of the control response function, estimating the model parameters given the data, and then adding higher-degree terms as needed to improve the model fit. In contrast, matching methods operate "top-down", making no assumptions about the individual terms and instead seeking exactly matched pairs in order to remove bias from all possible terms contained within the function $h^0(\mathbf{x})$ simultaneously.

BOSS bridges the gap between matching and regression. In general, BOSS works from some assumptions about the form of the control response function in order to identify which covariate balance measures are important. These assumptions can be specific (e.g., by identifying individual terms) or they can be general (e.g., by identifying which covariates are likely to have interactions). Once these assumptions are made, they can be translated directly to the covariate balance requirements in Table 3.1, assembled into an imbalance measure, and optimized using BOSS.

One claim of matching methods is that they are a non-parametric approach to causal inference in observational studies because they do not require any functional form assumptions. This is true if exact matches are obtained. However, when inexact matches are found and a claim is made about the matches possessing "sufficient covariate balance" to ensure an unbiased estimate, this is implicitly an assumption about the form of the control response function. This can be seen by looking at the above results from the reverse direction. That is, instead of the terms in the control response function requiring specific levels of covariate balance to ensure that $\mathcal{B}(T, C') = 0$, the levels of covariate balance indicate which terms can potentially be in the control response function without increasing $\mathcal{B}(T, C')$. Thus, a claim that balance is sufficient is equivalent to the claim that there are no terms in the control response function that require higher levels of covariate balance. If this claim is in doubt, then regression analysis can be used to estimate and adjust for the potential bias caused by additional terms that may be present (Ho et al., 2007; Abadie and Imbens, 2011).

BOSS provides a high level of specificity in determining the desired balance characteristics of the control group. This allows for the incorporation of prior knowledge into BOSS in a variety of ways. For example, certain covariates and distributions could be prioritized over others by including only them in $\mathbf{D}$, or by adding appropriate weights to $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$. In the absence of specific knowledge about the roles of each of the covariates, a good starting point is to use an imbalance measure that incorporates all of the covariate balance measures that would have been checked after running a matching method (e.g., marginal and pairwise joint distributions with balance assessed by a difference of means or a Kolmogorov-Smirnov test). If these covariate balance measures are used as the arbiter of success in matching, then they should also be sufficient for BOSS. Depending on the amount of residual imbalance in the control group identified by BOSS, it may be advantageous to modify the imbalance measure used by BOSS in order to remove to remove residual imbalance from a subset of the terms or ensure balance on additional ones.

3.5 Testing Assumptions

When dealing with observational data, it is desirable to justify an estimate of τ_T^1 by demonstrating the validity of the assumptions that are required for the estimate to be unbiased. This includes Assumption 2.3 at a minimum, and may include additional functional form assumptions depending on the level of covariate balance that is available in the identified control group. While these assumptions cannot be established in general, some progress can be made towards *corroborating* them by designing appropriate statistical tests to assess whether the available data conforms to either conditions that are directly established by the assumptions or to conditions that would be reasonably expected to hold if the assumptions were valid. While a positive outcome from these tests cannot guarantee that the assumptions are valid, a negative outcome can serve as a red flag to indicate potential problems.

3.5.1 Tests for Strong Ignorability

Recall from (2.8) that Assumption 2.3 ensures that

$$\mathbf{Pr}\left(Y_t^0 \leq y \mid \mathbf{X}_t = \mathbf{x}\right) = \mathbf{Pr}\left(Y_c^0 \leq y \mid \mathbf{X}_c = \mathbf{x}\right) \tag{3.13}$$

for all $y \in \mathbb{R}$ and $\mathbf{x} \in \mathcal{X}$. One way to test Assumption 2.3 is to use a Kolmogorov-Smirnov two-sample test with the null hypothesis that the two samples of control responses

$$\left\{Y_t^0 : t \in T,\ \mathbf{X}_t = \mathbf{x}\right\} \quad \text{and} \quad \left\{Y_c^0 : c \in C,\ \mathbf{X}_c = \mathbf{x}\right\}$$

are drawn from the same distribution at some $\mathbf{x} \in \{\mathbf{X}_u : u \in T \cup C\}$. The difficulty is that Y_t^0 is unobserved for all $t \in T$, and as such, there is nothing to which the observed control responses $\left\{Y_c^0 : c \in C,\ \mathbf{X}_c = \mathbf{x}\right\}$ can be compared.

Previous work focused on various ways to resolve the above difficulty. Rosenbaum (1984) proposes a method to test Assumption 2.3 under additional assumptions regarding the causal mechanisms that govern the treatment and control responses. As an example, if the treatment is known to have no effect (i.e., $y_u^1 = y_u^0$ for all $u \in U$), then under Assumption 2.3, the treatment responses $\left\{Y_t^1 : t \in U^1,\ \mathbf{X}_t = \mathbf{x}\right\}$ and the control responses $\left\{Y_c^0 : c \in U^0,\ \mathbf{X}_c = \mathbf{x}\right\}$ should be drawn from the same distribution. As both sets of responses are observed, they lead to an appropriate statistical test.

Rosenbaum (1987a) proposes another method of testing Assumption 2.3. Suppose the control population U^0 is split into two subpopulations U_1^0 and U_2^0 with different distributions of an unobserved covariate (and thus possibly different distributions of $\mathbf{X}$ and Y^0, as well). Then Assumption 2.3 implies that

$$\mathbf{Pr}\left(Y_c^0 \leq y \mid c \in U_1^0,\ \mathbf{X}_c = \mathbf{x}\right) = \mathbf{Pr}\left(Y_c^0 \leq y \mid c \in U_2^0,\ \mathbf{X}_c = \mathbf{x}\right) \tag{3.14}$$

for all $\mathbf{x} \in \mathcal{X}$ and all $y \in \mathbb{R}$. That is, units in subpopulation U_1^0 with covariate values $\mathbf{x}$ are equally likely to have a control response of at most y compared to units in U_2^0 with the same covariate values. Unlike (3.13), the values in (3.14) are all observed, and so they can be used in an appropriate statistical test. For example, if C_1 and C_2 are simple random samples from U_1^0 and U_2^0, respectively, then (3.14) can be checked at the points of contact between the samples.

The above two approaches work by ensuring that a comparison distribution exists. In the first case, this distribution comes from the treatment group T, while in the second case it comes from a second control group C_2. An alternative approach is to make an additional assumption regarding the distribution of the control responses at each $\mathbf{x} \in \mathcal{X}$, and then compare it to the empirical distribution at each observed $\mathbf{x}$.

For example, under the restated version of strong ignorability in Assumption 3.1, if it is additionally assumed that the error terms $\mathcal{E}^0$ are normally distributed at each $\mathbf{x} \in \mathcal{X}$, then the control responses $\left\{Y_c^0 : c \in C,\ \mathbf{X}_c = \mathbf{x}\right\}$ should be normally distributed at any $\mathbf{x} \in \mathcal{X}$ with at least two units. If instead it is assumed that the error terms are homoscedastic but not necessarily normal, then the variance of the control responses in each of the subpopulations $\{c \in C : \mathbf{X}_c = \mathbf{x}\}$ should be the same. In either case, an appropriate statistical test can be used to determine if the sampled units corroborate the additional assumption. While these tests focus on additional assumptions about the error terms instead of Assumption 2.3, in many cases it may be reasonable to assume that the error terms are well-behaved. If they are not, Assumption 2.3 is not directly invalidated but it may be called into question.

3.5.2 Tests for Functional Form Assumptions

Assumptions 3.2 and 3.6 are direct extensions of Assumption 2.3(a) (equivalently, Assumption 3.1), and as such, the tests in the preceding section can be used as an initial validation of them. If those tests fail, then any functional form assumption based on only the observed covariates will also be called into question. However, should Assumption 2.3 be corroborated by the initial tests, then it would be desirable to test the additional assumptions as well.

If the specific assumption includes covariate terms and their degree (e.g., Assumption 3.2 where all terms appear to the first degree), then one method of testing is to use linear regression to estimate the coefficients of the terms and then conduct a goodness-of-fit test on the resulting model. Such an approach may have difficulty without an additional assumption of homoscedasticity for the error terms. Additionally, this approach cannot be applied if general terms are assumed to be present in the response function (e.g., $h_i^0(x_i)$).

A general approach for testing the functional form assumptions is to indirectly analyze the control response error terms $\mathcal{E}^0$. Assumption 3.2 is used to illustrate this process. Suppose for the sampled control units $C \subset U^0$ that there are two disjoint subsets $C' \subset C$ and $C'' \subset C$ that satisfy $\bar{\mathbf{X}}_{C'} = \bar{\mathbf{X}}_{C''}$. Under

Assumption 3.2, the difference in average control responses between these groups is

$$\begin{aligned}
\bar{Y}^0_{C'} - \bar{Y}^0_{C''} &= \frac{1}{|C'|}\sum_{c\in C'} Y^0_c - \frac{1}{|C''|}\sum_{c\in C''} Y^0_c \\
&= \frac{1}{|C'|}\sum_{c\in C'} \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_c + \alpha + \mathcal{E}^0_c\right) - \frac{1}{|C''|}\sum_{c\in C''} \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_c + \alpha + \mathcal{E}^0_c\right) \\
&= \boldsymbol{\beta}^{\mathrm{T}}\left(\bar{\mathbf{X}}_{C'} - \bar{\mathbf{X}}_{C''}\right) + \frac{1}{|C'|}\sum_{c\in C'} \mathcal{E}^0_c - \frac{1}{|C''|}\sum_{c\in C''} \mathcal{E}^0_c \\
&= \frac{1}{|C'|}\sum_{c\in C'} \mathcal{E}^0_c - \frac{1}{|C''|}\sum_{c\in C''} \mathcal{E}^0_c .
\end{aligned}$$

This difference is zero in expectation by Assumption 2.4. It can also be computed in practice because $\bar{Y}^0_{C'}$ and $\bar{Y}^0_{C''}$ are both observed. If the difference is significant, it would raise concerns about Assumption 3.2. An additional assumption about the control response errors expands this.

Assumption 3.19. The variance of the random variable $\mathcal{E}^0$ is finite.

Lemma 3.20. *Let $c_1, c_2, \ldots, c_n$ be a random sample of control units drawn independently and identically from U^0, with corresponding control response error terms $\mathcal{E}^0_{c_1}, \mathcal{E}^0_{c_2}, \ldots, \mathcal{E}^0_{c_n}$. If Assumptions 3.1 and 3.19 are valid, so that $\boldsymbol{E}\left[\mathcal{E}^0_{c_i}\right] = 0$ and $\boldsymbol{Var}\left[\mathcal{E}^0_{c_i}\right] \equiv \sigma^2 < \infty$ for all $i \in \{1, 2, \ldots, n\}$, then the Central Limit Theorem shows that*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{E}^0_{c_i}\right) \xrightarrow{d} N(0, \sigma^2).$$

Under Assumption 2.5, the control units in C are drawn from U^0 independently and identically. Lemma 3.20 then shows that the average control response error across all sampled units, $\sum_{c\in C} \mathcal{E}^0_c/|C|$, is approximately normally distributed. Under the weaker conditions of Assumption 2.4, the control units in C are drawn independently but not necessarily identically. Because Assumption 3.1 only ensures that the errors are uncorrelated with the covariates, the error terms in C may not be identically distributed in this case. In particular, the conditional variance $\mathcal{E}^0$ given $\mathbf{X}$ is not guaranteed to be uniform across $\mathcal{X}$. However, if the error terms satisfy some additional conditions (e.g., Lyapunov's condition or Lindeberg's condition), then the central limit theorem can be applied to show that $\sum_{c\in C} \mathcal{E}^0_c/|C|$ is approximately normally distributed under Assumption 2.4, as well.

Lemma 3.20 motivates the following procedure. Select an $\mathbf{x} \in \mathcal{X}$ at random. From C, identify distinct (and preferably disjoint) subsets $C'_1, C'_2, \ldots, C'_k$ that satisfy $\bar{\mathbf{X}}_{C'_i} = \mathbf{x}$ for all $i \in \{1, 2, \ldots, k\}$. Examine the distributions of the average control responses $\bar{Y}^0_{C'_i}$ for each of the groups. If the resulting distribution of average control responses is not normal, then the validity of Assumption 3.2 should be questioned. A similar procedure can be applied for Assumption 3.6.

The above procedure is meant to motivate the use of multiple control groups in testing the assumptions and to serve as a starting point for further research. Potential problems that may need to be resolved include too much overlap among the control groups, a failure to find multiple groups with mean $\mathbf{x}$, or a failure to find groups that are sufficiently large for the averages of their errors to be approximately normal. Computational results in Chapter 5 demonstrate how the above procedure may be used in practice.

Chapter 4

Complexity and Approximation Results

The BOSS framework leads to the following discrete optimization problem: "Given sets of units T and C with associated covariate vectors $\{\mathbf{X}_t = \mathbf{x}_t\}_{t \in T}$ and $\{\mathbf{X}_c = \mathbf{x}_c\}_{c \in C}$ and imbalance measure $\mathcal{I} : \mathbb{N}^{\mathcal{X}} \times \mathbb{N}^{\mathcal{X}} \to [0, \infty)$, find a (non-empty) control group $C' \subseteq C$ that minimizes $\mathcal{I}(T, C')$." This chapter discusses the computational complexity of the above optimization problem by focusing on several cases that arise from specific imbalance measures. Approximation results are also discussed.

4.1 Complexity Results

The BOSS framework allows for the size of the control group to be specified as input. For any imbalance measure, a solution to the size-unconstrained optimization problem can be identified by solving the size-constrained problem for each possible size and then selecting the control group with minimal imbalance. As such, the results here consider the complexity of the size-constrained problem. The variant of BOSS in which the control units can be selected with replacement is not considered. The corresponding decision problem is: "Given sets of units T and C with associated covariate vectors, an imbalance measure $\mathcal{I}$, a target control group size s, and a value $\gamma \geq 0$, is there a $C' \subseteq C$ that satisfies $|C'| = s$ and $\mathcal{I}(T, C') \leq \gamma$?"

Lemma 4.1. *The decision problem for BOSS with a target size and an imbalance measure $\mathcal{I}$ that can be evaluated in time polynomial in $|T|$, $|C|$, and $|\mathcal{P}|$ is in **NP**.*

Proof. For a candidate control group C', the conditions in the lemma ensure that the two requirements $|C'| = s$ and $\mathcal{I}(T, C') \leq \gamma$ can be checked in polynomial time, so the problem is in **NP**. □

Lemma 4.1 applies to $\mathcal{I}_{\text{DOM}}$, $\mathcal{I}_{\text{SDOM}}$, $\mathcal{I}_{\text{KS}}$, $\mathcal{I}_{\text{KS:max}}$, $\mathcal{I}_{\text{Diff}}$, $\mathcal{I}_{\text{Diff}^2}$, and $\mathcal{I}_{\chi^2}$. While the coarsened distribution imbalance measures can be defined with an arbitrary number of bins, there can be at most $|T \cup C|$ occupied bins for each covariate that need to be checked. For the imbalance measures $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}$, the covariate

Some of the material in this chapter has been adapted from Sauppe et al. (2014), *INFORMS Journal on Computing* **26**(3), with the permission of the copyright holder.

clusters in $\mathbf{D}$ impact whether or not they can be evaluated in time polynomial in $|T|$, $|C|$, and $|\mathcal{P}|$. For example, if $\mathbf{D}$ contains a covariate cluster D with $|D| = \lfloor p/2 \rfloor$, then

$$|\mathcal{X}_D(T \cup C)| = \prod_{i \in D} |\mathcal{X}_i(T \cup C)| \in O\Big(|T \cup C|^{\lfloor p/2 \rfloor}\Big),$$

which is not polynomial in p. For $\mathcal{I}_{\text{Diff:}\mathbf{D}}$, this problem can be mitigated by only inspecting the occupied bins for cluster D, but $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$ potentially has to compute the empirical cumulative distribution functions for T and C' at all of these values (this can be improved to only consider the values in $\mathcal{X}_D(T \cup C')$, though the number of values is still exponential in p). A related problem is the number of clusters in $\mathbf{D}$, which can be up to $2^p - 1$. However, if $\mathbf{D}$ is restricted to contain only clusters with at most some constant number of covariates, then these functions can be evaluated in time polynomial in $|T|$, $|C|$, and p and so Lemma 4.1 applies.

4.1.1 Mean Imbalance Measures

This section considers BOSS with the imbalance measure $\mathcal{I}_{\text{DOM}}$.

Theorem 4.2. *The decision problem for BOSS with $\mathcal{I}_{DOM}$ and a target size is **NP-Hard**.*

Proof. The result can be shown using a reduction from the Subset Sum problem, which is **NP-Hard** (Garey and Johnson, 1979). The Subset Sum problem is: Given a set A of integers, does there exist an nonempty subset $A' \subseteq A$ such that $\sum_{a \in A'} a = d$?

For an arbitrary instance of Subset Sum, the reduction proceeds as follows. The constructed BOSS instance has one covariate, $\mathcal{P} \equiv \{1\}$. The treatment group T contains one unit t with $x_{t,1} = d$. The control pool C contains a unit c_a for each $a \in A$ with $x_{c_a,1} = a \cdot |A|$, along with $|A| - 1$ "placeholder" units with value 0 for covariate 1. The target control group size is $s = |A|$, and the desired imbalance level is $\gamma = 0$. This completes the reduction, which can be performed in time polynomial in the size of the Subset Sum instance. It remains to be shown that the answer to the Subset Sum instance is yes if and only if the answer to the constructed BOSS instance is yes.

($\Rightarrow$) Let C' be a control group for BOSS that satisfies $|C'| = |A|$ and $\mathcal{I}_{\text{DOM}}(T, C') \leq \gamma = 0$. This implies that

$$\frac{1}{|C'|} \sum_{c \in C'} x_{c,1} = \frac{1}{|T|} \sum_{t \in T} x_{t,1} = d$$

Let $C'_A = \{c \in C' : c = c_a \text{ for some } a \in A\}$. Since $|C'| = |A|$, $|C'_A| \geq 1$ since there are only $|A| - 1$ total placeholder units. Then

$$d = \frac{1}{|C'|}\sum_{c \in C'} x_{c,1} = \frac{1}{|A|}\left(\sum_{c_a \in C'_A} x_{c_a,1} + \sum_{c \in C' \setminus C'_A} x_{c,1}\right) = \frac{1}{|A|}\left(\sum_{c_a \in C'_A} a|A| + \sum_{c \in C' \setminus C'_A} 0\right) = \sum_{c_a \in C'_A} a,$$

so $A' \equiv \{a \in A : c_a \in C'_A\}$ is a nonempty subset of A that sums to d, as desired.

($\Leftarrow$) Let $A' \subseteq A$ be a nonempty set of integers such that $\sum_{a \in A'} a = d$. Let $C'_A = \{c_a : a \in A'\}$ be the corresponding set of control units, and let C' be the set C'_A padded with any additional units $c \in C$ satisfying $x_{c,1} = 0$ so that $|C'| = |A|$. Such a padding is possible because there are $|A| - 1$ placeholder units available and $|A'| \geq 1$. Then it is straightforward to verify that $\sum_{c \in C'} x_{c,1}/|C'| = d$, which ensures that $\mathcal{I}(T, C') = 0 \leq \gamma$, as desired. □

Corollary 4.3. *The decision problem for BOSS with $\mathcal{I}_{DOM}$ and a target size remains **NP-Hard** if restricted to a single covariate.*

The special case of Subset Sum in which the target sum is 0 can be reduced to BOSS with $\mathcal{I}_{\text{DOM}}$ and no target control group size. For this reduction, there is no need to introduce placeholder control units with $x_{c,1} = 0$.

4.1.2 Coarsened Distribution Imbalance Measures

This section considers BOSS with the imbalance measures $\mathcal{I}_{\text{Diff}}$ and $\mathcal{I}_{\text{Diff}:\mathbf{D}}$. The bin boundaries for the covariates and the covariate clusters $\mathbf{D}$ are specified with the problem.

Theorem 4.4. *The decision problem for BOSS with $\mathcal{I}_{Diff:\mathbf{D}}$ and a target size is **NP-Hard**.*

Proof. The result can be shown using a reduction from the 3-Dimensional Matching (3DM) problem, which is **NP-Hard** (Garey and Johnson, 1979). The 3DM problem is: Given disjoint sets Q, R, and W with $|Q| = |R| = |W| = d$ and $V \subseteq Q \times R \times W$, does there exist a $V' \subseteq V$ such that $|V'| = d$ and no two elements of V' agree in any coordinate?

For an arbitrary instance of 3DM with $Q \equiv \{q_1, q_2, \ldots, q_d\}$, $R \equiv \{r_1, r_2, \ldots, r_d\}$, $W \equiv \{w_1, w_2, \ldots, w_d\}$, and $V \subseteq Q \times R \times W$, the reduction proceeds as follows. It is assumed that $|V| \geq d$, otherwise the problem is infeasible. The constructed BOSS instance has three covariates, $\mathcal{P} \equiv \{1, 2, 3\}$. The treatment group T contains d units, indexed from 1 through d, with $x_{t_i1} = x_{t_i2} = x_{t_i3} = i$ for $i \in \{1, 2, \ldots, d\}$. For each $v \equiv (q_i, r_j, w_k) \in V$, the control pool C contains a unit c_v with $x_{c_v1} = i$, $x_{c_v2} = j$, and $x_{c_v3} = k$. Each covariate $i \in \mathcal{P}$ has $n_i = d$ histogram bins, with the bin boundaries given by

$$b_{i0} = 1,\ b_{i,n_i} = d,\ b_{ij} = j + 0.5\ \forall\ j \in \{1, 2, \ldots, n_i - 1\}.$$

Under these definitions, the units in each histogram bin are given by

$$B_{ij} \equiv \{t_j\} \cup \{c \in C : x_{ci} = j\}\,.$$

To complete the reduction, let $\mathbf{D} \equiv \{\{1\}, \{2\}, \{3\}\}$, set the target control group size as $s = d$, and set $\gamma = 0$. All steps for the reduction can be finished in time polynomial in the size of the 3DM instance. It remains to be shown that the answer to the 3DM instance is yes if and only if the answer to the constructed BOSS instance is yes.

($\Leftarrow$) Let C' be a control group for BOSS that satisfies $|C'| = d$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') \leq \gamma = 0$. This implies that $\eta_{ij}(T) = \eta_{ij}(C')$ for all $i \in \mathcal{P}$ and $j \in N_i$. By construction, $\eta_{ij}(T) = 1/|T|$, which means that $|C' \cap B_{ij}| = 1$ for all $i \in \mathcal{P}$ and $j \in N_i$. Now let $V' \equiv \{v \in V : c_v \in C'\}$. By construction, the number of tuples in V' having element $q_j \in Q$ is equal to the number of control units in C' with $x_{c1} = j$, for all $j \in \{1, 2, \ldots, d\}$. From the bin definitions, this second quantity is $|C' \cap B_{ij}|$, which equals one. Hence V' contains exactly one tuple with element q_j for each $j \in \{1, 2, \ldots, d\}$. An identical argument applies for the elements in R and W, so V' is a valid solution to the 3DM instance.

($\Rightarrow$) Let $V' \subseteq V$ be a solution to the 3DM instance. This means that there is exactly one tuple $v \in V'$ containing $q_j \in Q$, for each $j \in \{1, 2, \ldots, d\}$. Similar conditions hold for the elements in R and W. This ensures that $|V'| = d$. Now let $C' \equiv \{c_v : v \in V'\}$, so that $|C'| = |V'| = d$. By construction, C' contains exactly one control unit c with $x_{ci} = j$ for each $i \in \mathcal{P}$ and $j \in \{1, 2, \ldots, d\}$. Then by the bin definitions, $|C' \cap B_{ij}| = 1$ for all $i \in \mathcal{P}$ and $j \in N_i$, and by the construction of T,

$$\eta_{ij}(T) = \frac{|T \cap B_{ij}|}{|T|} = \frac{1}{d} = \frac{|C' \cap B_{ij}|}{|C'|} = \eta_{ij}(C')$$

for all $i \in \mathcal{P}$ and $j \in N_i$, which means that $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0 \leq \gamma$. □

The reduction used in the proof of Theorem 4.4 leads to the following result.

Corollary 4.5. *The decision problem for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size remains* ***NP-Hard*** *for instances with three covariates.*

The reduction in the proof of Theorem 4.4 fails in the case when there is no constraint on the size of the control group. To see why, observe that the structure of the treatment group requires $|T \cap B_{ij}|/|T| = 1/|T|$ for all $i \in \mathcal{P}$ and $j \in N_i$. A control group C' with $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0$ must then satisfy $|T||C' \cap B_{ij}|/|C'| = 1$, which can occur if $|C' \cap B_{ij}| = k$ and $|C'| = k|T|$ for $k \in \mathbb{Z}^+$. For $k \geq 2$, the BOSS solution cannot necessarily be used to construct a valid solution for 3DM. As an example, the 3DM instance $Q \equiv \{q_1, q_2\}$, $R \equiv \{r_1, r_2\}$, $W \equiv \{w_1, w_2\}$ and

$$V \equiv \{(q_1, r_1, w_1), (q_2, r_1, w_2), (q_1, r_2, w_2), (q_2, r_2, w_1)\}$$

has no valid solution, but the corresponding instance of BOSS satisfies $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0$ by letting $C' = C$, i.e., without a restriction on $|C'|$.

Lemma 4.1 and Theorem 4.4 together imply that the decision problem for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size is **NP-Complete**. More specifically, the decision problem for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size is strongly **NP-Complete** because Theorem 4.4 applies even if the numeric inputs for BOSS (i.e., the covariate vectors) are restricted to be polynomial in the size of $|T|$, $|C|$, and $|\mathcal{P}|$. Hence, the optimization problem for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size has no pseudo-polynomial time algorithm and no fully polynomial-time approximation scheme (FPTAS) unless $\mathbf{P} = \mathbf{NP}$. Analogous results apply to BOSS with $\mathcal{I}_{\text{Diff}}$.

There are two natural restrictions with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$: limit the number of covariates (and consequently the number of covariate clusters), or limit the number of bins for each covariate (i.e., bound $|N_i|$ for all $i \in \mathcal{P}$). In the case of limiting the number of covariates, Corollary 4.5 establishes that there are only a few remaining cases in which this might lead to problems that can be solved in polynomial time.

Before considering the restricted cases, some properties of $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ can be established. For a fixed treatment group T and set of histogram bins, define

$$\Delta^-_{Dj}(C') \equiv \max\left(0, |T \cap B_{Dj}|/|T| - |C' \cap B_{Dj}|/|C'|\right) \tag{4.1a}$$

$$\Delta^+_{Dj}(C') \equiv \max\left(0, |C' \cap B_{Dj}|/|C'| - |T \cap B_{Dj}|/|T|\right) \tag{4.1b}$$

as the *bin shortage* and *bin excess*, respectively, of control group C' (with respect to T) for any $D \subseteq \mathcal{P}$ and $j \in N_D$. By definition, the bin shortage and excess satisfy

$$\begin{aligned}\Delta^-_{Dj}(C') + \Delta^+_{Dj}(C') &= \max\left(\frac{|T \cap B_{Dj}|}{|T|} - \frac{|C' \cap B_{Dj}|}{|C'|}, \; \frac{|C' \cap B_{Dj}|}{|C'|} - \frac{|T \cap B_{Dj}|}{|T|}\right) \\ &= \left|\frac{|T \cap B_{Dj}|}{|T|} - \frac{|C' \cap B_{Dj}|}{|C'|}\right| \\ &= |\eta_{Dj}(T) - \eta_{Dj}(C')|\end{aligned}$$

for all $D \subseteq \mathcal{P}$ and $j \in N_D$. This allows $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ to be expressed as

$$\begin{aligned}
\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') &\equiv \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \left|\eta_{Dj}(T) - \eta_{Dj}(C')\right| \\
&= \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \left(\Delta^{-}_{Dj}(C') + \Delta^{+}_{Dj}(C')\right) \\
&= \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \Delta^{-}_{Dj}(C') + \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \Delta^{+}_{Dj}(C').
\end{aligned} \tag{4.2}$$

For any control group C' and covariate cluster D, the bin shortages and excesses are related by separating the bins in N_D into two categories (defined with respect to a fixed treatment group T):

$$\begin{aligned}
N^{-}_{D}(C') &\equiv \left\{j \in N_D : \left|T \cap B_{Dj}\right| / |T| \geq \left|C' \cap B_{Dj}\right| / |C'|\right\} \\
N^{+}_{D}(C') &\equiv \left\{j \in N_D : \left|T \cap B_{Dj}\right| / |T| < \left|C' \cap B_{Dj}\right| / |C'|\right\}.
\end{aligned}$$

These sets of bins are related through

$$\begin{aligned}
\sum_{j \in N^{-}_{D}(C')} \left(\eta_{Dj}(T) - \eta_{Dj}(C')\right) &= \sum_{j \in N^{-}_{D}(C')} \eta_{Dj}(T) - \sum_{j \in N^{-}_{D}(C')} \eta_{Dj}(C') \\
&= \left(1 - \sum_{j \in N^{+}_{D}(C')} \eta_{Dj}(T)\right) - \left(1 - \sum_{j \in N^{+}_{D}(C')} \eta_{Dj}(C')\right) \\
&= \sum_{j \in N^{+}_{D}(C')} \left(\eta_{Dj}(C') - \eta_{Dj}(T)\right),
\end{aligned} \tag{4.3}$$

which follows because $\sum_{j \in N_D} \eta_{Dj}(S) = 1$ for any $S \subseteq T \cup C$ and $D \subseteq \mathcal{P}$. By definition, the bin shortages satisfy

$$\begin{aligned}
\sum_{j \in N_D} \Delta^{-}_{Dj}(C') &= \sum_{j \in N^{-}_{D}(C')} \Delta^{-}_{Dj}(C') + \sum_{j \in N^{+}_{D}(C')} \Delta^{-}_{Dj}(C') \\
&= \sum_{j \in N^{-}_{D}(C')} \left(\frac{\left|T \cap B_{Dj}\right|}{|T|} - \frac{\left|C' \cap B_{Dj}\right|}{|C'|}\right) + \sum_{j \in N^{+}_{D}(C')} 0 \\
&= \sum_{j \in N^{-}_{D}(C')} \left(\eta_{Dj}(T) - \eta_{Dj}(C')\right).
\end{aligned} \tag{4.4}$$

A similar argument applies to the bin excesses, yielding

$$\sum_{j \in N_D} \Delta^{+}_{Dj}(C') = \sum_{j \in N^{+}_{D}(C')} \left(\eta_{Dj}(C') - \eta_{Dj}(T)\right). \tag{4.5}$$

Combining (4.3), (4.4), and (4.5) yields

$$\sum_{j \in N_D} \Delta^-_{Dj}(C') = \sum_{j \in N_D} \Delta^+_{Dj}(C'). \tag{4.6}$$

Then (4.2) and (4.6) show that

$$\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 2 \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \Delta^-_{Dj}(C') = 2 \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \Delta^+_{Dj}(C'). \tag{4.7}$$

Thus, minimizing $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ is equivalent to minimizing twice the total bin shortage or twice the total bin excess.

Theorem 4.6. *The optimization problem for BOSS with $\mathcal{I}_{\mathit{Diff}:\mathbf{D}}$ and a target size is in* ***P*** *when $|\mathcal{P}| \leq 2$.*

Proof. The result can be shown by using a reduction to the minimum cost network flow (MCNF) problem. Consider an arbitrary instance of BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$, target size s, two covariates $\mathcal{P} \equiv \{1, 2\}$, treatment group T, control pool C, a set of histogram bins for each covariate (with empty bins removed), and the full set of covariate clusters $\mathbf{D} \equiv \{\{1\}, \{2\}, \{1, 2\}\}$. For the remainder of the proof, let $D \equiv \{1, 2\}$. Let $r_{ij} \equiv s \cdot |T \cap B_{ij}|/|T|$ for each $i \in \mathcal{P}$ and $j \in N_i$, and let $r_{Dj} \equiv s \cdot |T \cap B_{Dj}|/|T|$ for each $j \in N_D$.

The reduction to MCNF proceeds as follows. Create a source vertex v^+ with supply s and a sink vertex v^- with supply $-s$. For each $j \in N_1$, create a transshipment node v_{1j}, an arc $a^-_{1j} \equiv (v^+, v_{1j})$ with cost 0 and capacity $\lfloor r_{1j} \rfloor$, an arc $a^+_{1j} \equiv (v^+, v_{1j})$ with cost 1 and capacity ∞, and an arc $a^=_{1j} \equiv (v^+, v_{1j})$ with cost $\lceil r_{1j} \rceil - r_{1j}$ and capacity $\lceil r_{1j} \rceil - \lfloor r_{1j} \rfloor$. Similarly, for each $j \in N_2$, create a transshipment node v_{2j}, an arc $a^-_{2j} \equiv (v_{2j}, v^-)$ with cost 0 and capacity $\lfloor r_{2j} \rfloor$, an arc $a^+_{2j} \equiv (v_{2j}, v^-)$ with cost 1 and capacity ∞, and an arc $a^=_{2j} \equiv (v_{2j}, v^-)$ with cost $\lceil r_{2j} \rceil - r_{2j}$ and capacity $\lceil r_{2j} \rceil - \lfloor r_{2j} \rfloor$. Finally, for each $j \equiv (j_1, j_2) \in N_D$, create a transshipment node v_{Dj}, an arc $a_{Dj} \equiv (v_{1,j_1}, v_{Dj})$ with cost 0 and capacity $|C \cap B_{Dj}|$, an arc $a^-_{Dj} \equiv (v_{Dj}, v_{2,j_2})$ with cost 0 and capacity $\lfloor r_{Dj} \rfloor$, an arc $a^+_{Dj} \equiv (v_{Dj}, v_{2,j_2})$ with cost 1 and capacity ∞, and an arc $a^=_{Dj} \equiv (v_{Dj}, v_{2,j_2})$ with cost $\lceil r_{Dj} \rceil - r_{Dj}$ and capacity $\lceil r_{Dj} \rceil - \lfloor r_{Dj} \rfloor$. Let V and A be the sets of all nodes and arcs, respectively, in the network. This completes the reduction, which can be done in time polynomial in the number of (occupied) bins, which is at most $|T \cup C|$. The costs in the MCNF instance are scaled by $s/2$ with respect to $\mathcal{I}_{\text{Diff:}\mathbf{D}}$. Figure 4.1 shows the network flow instance that arises from the above transformation.

By construction, the minimum cost network flow instance has integer capacities on all arcs and integer supply and demand values at the source and sink vertices. As such, it has an integral optimal solution that can be found in polynomial time (Ahuja et al., 1993). It remains to be shown that the BOSS instance has a solution C' satisfying $|C'| = s$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') \leq \gamma$ if and only if the corresponding minimum cost network flow instance has an integral solution with cost at most $s\gamma/2$. Some additional notation is used in this regard. Let $f : A \to [0, \infty)$, $\text{cap} : A \to \mathbb{Z}$, and $\text{cost} : A \to \mathbb{R}$ be the flow, capacity, and cost (per unit flow) functions, respectively, for the arcs in the network, and let $\sup : V \to \mathbb{R}$ be the supply function for the nodes in the network. A flow is feasible if it satisfies the capacity constraints $f(a) \leq \text{cap}(a)$ for all $a \in A$ and conservation of flow

$$\sum_{a \equiv (v', v) \in A} f(a) + \sup(v) = \sum_{a \equiv (v, v') \in A} f(a)$$

for all $v \in V$.

($\Rightarrow$) Let C' be a control group satisfying $|C'| = s$ and $\mathcal{I}(T, C') \leq \gamma$ for some $\gamma \geq 0$. A minimum cost flow solution f with integer flow can be constructed as follows. For each $j \in N_D$, let $f(a_{Dj}) \equiv |C' \cap B_{Dj}|$, which does not violate arc capacities because $|C' \cap B_{Dj}| \leq |C \cap B_{Dj}| = \text{cap}(a_{Dj})$. Flow conservation requires that

$$f(a^{-}_{Dj}) + f(a^{=}_{Dj}) + f(a^{+}_{Dj}) = f(a_{Dj}) = |C' \cap B_{Dj}|,$$

which can be satisfied by setting

$$\begin{aligned}
f(a^{-}_{Dj}) &= \min\left(\lfloor r_{Dj} \rfloor, |C' \cap B_{Dj}|\right), \\
f(a^{=}_{Dj}) &= \begin{cases} 0 & \text{if } |C' \cap B_{Dj}| \leq \lfloor r_{Dj} \rfloor \leq r_{Dj} \\ 1 & \text{if } |C' \cap B_{Dj}| > r_{Dj}, \end{cases} \qquad (4.8) \\
f(a^{+}_{Dj}) &= \max\left(0, |C' \cap B_{Dj}| - \lceil r_{Dj} \rceil\right).
\end{aligned}$$

By construction, these flows do not violate the arc capacities. Additionally, these flows ensure that the flow costs from the arcs $a^{=}_{Dj}$ and a^{+}_{Dj} satisfy

$$f(a^{=}_{Dj}) \cdot \text{cost}(a^{=}_{Dj}) + f(a^{+}_{Dj}) \cdot \text{cost}(a^{+}_{Dj}) = \begin{cases} 0 & \text{if } |C' \cap B_{Dj}| \leq r_{Dj} \\ \lceil r_{Dj} \rceil - r_{Dj} & \text{if } r_{Dj} < |C' \cap B_{Dj}| = \lceil r_{Dj} \rceil \\ |C' \cap B_{Dj}| - r_{Dj} & \text{if } |C' \cap B_{Dj}| > r_{Dj}. \end{cases}$$

With (4.1), the above costs simplify to

$$f(a^{=}_{Dj}) \cdot \text{cost}(a^{=}_{Dj}) + f(a^{+}_{Dj}) \cdot \text{cost}(a^{+}_{Dj}) = |C'| \cdot \Delta^{+}_{Dj}(C'). \qquad (4.9)$$

Flow conservation also requires that

$$f(a^{-}_{1,j_1}) + f(a^{=}_{1,j_1}) + f(a^{+}_{1,j_1}) = \sum_{j \equiv (j_1,j_2) \in N_D} f(a_{Dj}) = \sum_{j \equiv (j_1,j_2) \in N_D} |C' \cap B_{Dj}| = |C' \cap B_{1,j_1}| \quad \forall\ j_1 \in N_1,$$
$$f(a^{-}_{2,j_2}) + f(a^{=}_{2,j_2}) + f(a^{+}_{2,j_2}) = \sum_{j \equiv (j_1,j_2) \in N_D} f(a_{Dj}) = \sum_{j \equiv (j_1,j_2) \in N_D} |C' \cap B_{Dj}| = |C' \cap B_{2,j_2}| \quad \forall\ j_2 \in N_2,$$

which can be satisfied by using flow assignments comparable to those in (4.8). The resulting flow costs are also comparable to those in (4.9). These assignments satisfy flow conservation requirements at the source and sink nodes. For the source node, the outgoing flow is

$$\sum_{j \in N_1} \left(f(a^{-}_{1j}) + f(a^{=}_{1j}) + f(a^{+}_{1j}) \right) = \sum_{j \in N_1} |C' \cap B_{1j}| = |C'| = \sup(v^{+}),$$

while the sink node has a total incoming flow of

$$\sum_{j \in N_2} \left(f(a^{-}_{2j}) + f(a^{=}_{2j}) + f(a^{+}_{2j}) \right) = \sum_{j \in N_2} |C' \cap B_{2j}| = |C'| = -\sup(v^{-}).$$

This establishes that f is feasible and also integral. The total cost is computed with (4.7) and (4.9) as

$$\begin{aligned}
\sum_{a \in A} f(a) \cdot \mathrm{cost}(a) &= \sum_{j \in N_1} |C'| \cdot \Delta^{+}_{1j}(C') + \sum_{j \in N_2} |C'| \cdot \Delta^{+}_{2j}(C') + \sum_{j \in N_D} |C'| \cdot \Delta^{+}_{Dj}(C') \\
&= |C'| \left(\sum_{j \in N_1} \Delta^{+}_{1j}(C') + \sum_{j \in N_2} \Delta^{+}_{2j}(C') + \sum_{j \in N_D} \Delta^{+}_{Dj}(C') \right) \\
&= s \cdot \mathcal{I}_{\mathrm{Diff:}\mathbf{D}}(T, C')/2 \\
&\leq s\gamma/2,
\end{aligned}$$

which completes this direction.

($\Leftarrow$) Let f be a feasible integral flow solution to the minimum cost network flow instance with cost $s\gamma/2$ for some $\gamma \geq 0$. This ensures that $f(a_{Dj})$ is integral and satisfies $f(a_{Dj}) \leq |C \cap B_{Dj}|$ for all $j \in N_D$. Construct the control group C' by selecting an arbitrary set of $f(a_{Dj})$ units from $C \cap B_{Dj}$, for each $j \in D$. Because f is feasible, flow conservation requires that the source node v^{+} sends out s units of flow. As the arcs a_{Dj} for $j \in N_D$ form a cut in the network, there must be s units of flow crossing these arcs. By construction, $|C'| = s$. Flow conservation requirements can be combined with the results from (4.8) and (4.9) to show that

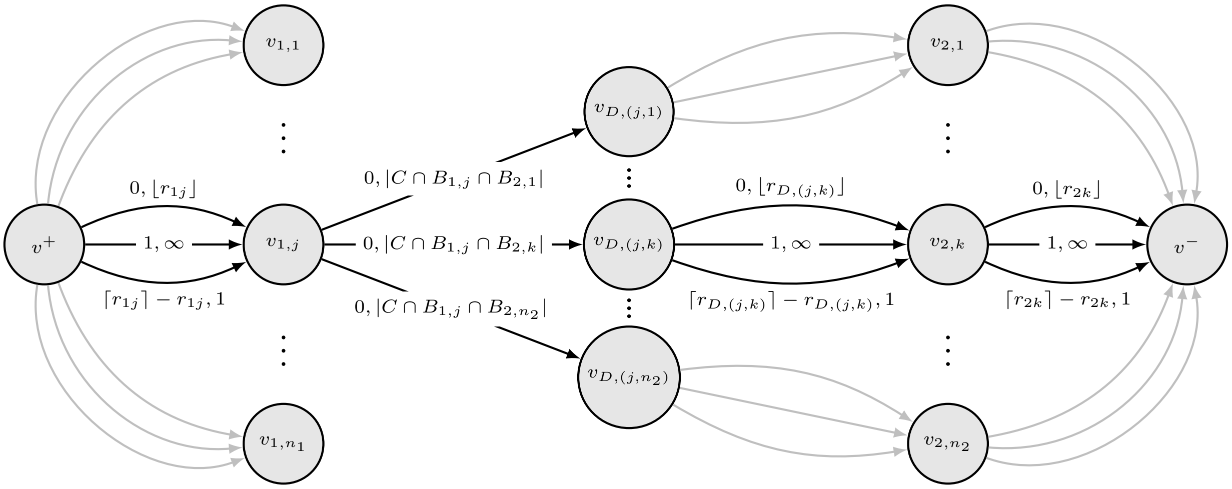
$$f(a^{\overline{=}}_{Dj}) \cdot \text{cost}(a^{\overline{=}}_{Dj}) + f(a^{+}_{Dj}) \cdot \text{cost}(a^{+}_{Dj}) \geq |C'| \cdot \Delta^{+}_{Dj}(C')$$

for all $j \in N_D$, with similar results holding for covariates 1 and 2. Hence,

$$|C'| \sum_{j \in N_1} \Delta^{+}_{1j} + |C'| \sum_{j \in N_2} \Delta^{+}_{2j} + |C'| \sum_{j \in N_D} \Delta^{+}_{Dj} \leq \sum_{a \in A} f(a) \cdot \text{cost}(a) = s\gamma/2,$$

and consequently $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') \leq \gamma$, as desired.

This completes the proof for the case when $\mathbf{D} = \{\{1\}, \{2\}, \{1, 2\}\}$. The other cases for two covariates with $\mathbf{D} \subset \{\{1\}, \{2\}, \{1, 2\}\}$ are handled by setting the arc costs to zero for any arc associated with an excluded covariate cluster. The case with one covariate is handled through an appropriate modification to the network. $\square$



$v_{1,1}$
$v_{D,(j,1)}$
$v_{2,1}$
$0, \lfloor r_{1j} \rfloor$
$0, |C \cap B_{1,j} \cap B_{2,1}|$
$0, \lfloor r_{D,(j,k)} \rfloor$
$0, \lfloor r_{2k} \rfloor$
v^{+}
$1, \infty$
$v_{1,j}$
$0, |C \cap B_{1,j} \cap B_{2,k}|$
$v_{D,(j,k)}$
$1, \infty$
$v_{2,k}$
$1, \infty$
v^{-}
$\lceil r_{1j} \rceil - r_{1j}, 1$
$0, |C \cap B_{1,j} \cap B_{2,n_2}|$
$\lceil r_{D,(j,k)} \rceil - r_{D,(j,k)}, 1$
$\lceil r_{2k} \rceil - r_{2k}, 1$
$v_{D,(j,n_2)}$
v_{1,n_1}
v_{2,n_2}

Figure 4.1: The minimum cost network flow transformation for solving BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ for $\mathcal{P} \equiv \{1, 2\}$. Only a subset of the vertices and edges in the network are shown. Labels on edges indicate cost and capacity, respectively. The source vertex v^{+} has s units of flow to send to the destination vertex v^{-}.

Lemma 4.7. *The optimization problem for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size is in **P** when $|\mathbf{D}| = 2$ or when $|\mathbf{D}| = 3$ with $\mathbf{D} \equiv \{D_1, D_2, D_3\}$ and $D_1 \cup D_2 = D_3$.*

Proof. The reduction in the proof of Theorem 4.6 can be adapted to the case with two covariate clusters $\mathbf{D} \equiv \{D_1, D_2\}$ by replacing the nodes corresponding to the bins from covariates 1 and 2 with nodes corresponding to the bins for the covariate clusters, and then adding arcs from each node corresponding to bin $j_1 \in N_{D_1}$ to each node corresponding to bin $j_2 \in N_{D_2}$ with cost 0 and capacity $|C \cap B_{D_1 j_1} \cap B_{D_2 j_2}|$. Empty bins are omitted.

The case with three covariate clusters follows from the above by first observing that $|C \cap B_{D_1 j_1} \cap B_{D_2 j_2}| = 0$ if $|D_1 \cap D_2| > 0$ and the components in j_1 and j_2 do not agree for the shared covariates. If j_1 and j_2 are compatible, meaning that they either have the same bin for each common covariate or that they have no covariates in common, then $|C \cap B_{D_1 j_1} \cap B_{D_2 j_2}| = |C \cap B_{D_3 j_3}|$, where the components of j_3 are determined by the values in j_1 and j_2. In this case, the network can be modified to penalize bin excess for covariate cluster D_3. □

Lemma 4.8. *The decision problem for BOSS with $\mathcal{I}_{Diff:\mathbf{D}}$ and a target size remains **NP-Hard** for instances with $|N_i| \leq 2$ for all $i \in \mathcal{P}$.*

Proof. The result can be shown by using a reduction from the Exact Cover by 3-Sets (X3C) problem, which is **NP-Hard** (Garey and Johnson, 1979). The X3C problem is: Given a set V with $|V| = 3q$ and a collection W of 3-element subsets of V, does there exist a $W' \subseteq W$ such that every element of V occurs in exactly one member of W'?

For an arbitrary instance of X3C with set $V \equiv \{1, 2, \ldots, 3q\}$ and a collection W of 3-element subsets of V, the reduction proceeds as follows. It is assumed that $|W| \geq q$, otherwise the problem is infeasible. The constructed BOSS instance has $3q$ covariates, $\mathcal{P} \equiv \{1, 2, \ldots, 3q\}$. The treatment group T contains q units, indexed from 1 through q, with $x_{t_1 i} = 1$ for all $i \in \mathcal{P}$ and $x_{t_j i} = 2$ for all $i \in \mathcal{P}$ and $j \in \{2, 3, \ldots, q\}$. For each three-set $w \equiv (i, j, k) \in W$, the control pool C contains a unit c_w with $x_{c_w i} = x_{c_w j} = x_{c_w k} = 1$ and $x_{c_w l} = 2$ for all $l \in \mathcal{P} \setminus \{i, j, k\}$. Each covariate $i \in \mathcal{P}$ has $n_i = 2$ histogram bins, with the bin boundaries defined as $b_{i0} = 1$, $b_{i1} = 1.5$, and $b_{i2} = 2$. Under these definitions, the units in each bin are

$$B_{i1} \equiv \{t_1\} \cup \{c_w \in C : i \in w\},$$
$$B_{i2} \equiv \{t_2, t_3, \ldots t_q\} \cup \{c_w \in C : i \notin w\},$$

for all $i \in \mathcal{P}$. To complete the reduction, let $\mathbf{D} \equiv \{\{i\} : i \in \mathcal{P}\}$, set the target control group size as $s = q$, and set $\gamma = 0$. All steps for the reduction can be finished in time polynomial in the size of the X3C instance. It remains to be shown that the answer to the X3C instance is yes if and only if the answer to the constructed BOSS instance is yes.

($\Leftarrow$) Let C' be a control group for BOSS that satisfies $|C'| = q$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') \leq \gamma = 0$. This implies that $\eta_{ij}(T) = \eta_{ij}(C')$ for all $i \in \mathcal{P}$ and $j \in N_i$. By the definition of T, this means that $|C' \cap B_{i1}| = 1$ for all $i \in \mathcal{P}$. Now let $W' \equiv \{w \in W : c_w \in C'\}$. By construction, the number of 3-sets in W' having element

$i \in V$ is equal to the number of control units in C' with $x_{ci} = 1$. From above, this second quantity equals one for all $i \in V$. This ensures that each element in V occurs in exactly one member of W', so W' is a valid solution to the X3C instance.

($\Rightarrow$) Let $W' \subseteq W$ be a solution to the X3C instance. This means that each element in V occurs in exactly one member of W'. Because W' contains 3-element subsets and $|V| = 3q$, it must be the case that $|W'| = q$. Now let $C' \equiv \{c_w \in C : w \in W'\}$, which satisfies $|C'| = q$. For each $i \in \mathcal{P}$, $|C' \cap B_{i1}|$ is equal to the number of units in C' with $x_{ci} = 1$. By construction, this second quantity is equal to the number of 3-sets in W' containing element i. From above, this quantity equals 1 for all $i \in V$. So $|C' \cap B_{i1}| = 1$ and $|C' \cap B_{i2}| = |C'| - |C' \cap B_{i1}| = q - 1$ for all $i \in \mathcal{P}$. The treatment group also satisfies this property. Because $|T| = |C'|$, it follows that $\mathcal{I}(T, C') = 0 \leq \gamma$. □

One final case that can be considered is if the number of covariates and the number of bins are both restricted.

Lemma 4.9. *The optimization problem for BOSS with $\mathcal{I}_{\text{Diff}}$ and a target size is in* ***P*** *if $|\mathcal{P}| \leq a_1$ and $|N_i| \leq a_2$ for all $i \in \mathcal{P}$ for constants a_1 and a_2.*

Proof. For each $c \in C$, the covariate vector $\mathbf{x}_c$ can be replaced with a bin vector satisfying $x_{ci} = j$ if $c \in B_{ij}$ for all $i \in \mathcal{P}$. The number of unique bin vectors is

$$\prod_{i \in \mathcal{P}} |N_i| \leq \prod_{i \in \mathcal{P}} a_2 \leq a_2^{a_1} \equiv a,$$

where a is a (potentially large) constant. The problem of selecting a control group $C' \subseteq C$ with target size s is then replaced by the problem of selecting a multiset of s bin vectors from among those that are available, with the units in C determining how many times each bin vector can be selected. In the worst case, each bin vector can appear at least s times in C. The multiset coefficient provides the number of ways in which s elements can be selected with repetition from a set of a unique elements, which leads to an upper bound of

$$\binom{s+a-1}{s} = \frac{(s+a-1)!}{s!(a-1)!} \in O\Big((s+a-1)^{a-1}\Big)$$

on the number of unique sets of bin vectors that can be used to form a valid solution for BOSS. Thus, a brute force search and evaluation of these solutions can be done in time proportional to a (large) polynomial in $s + a - 1$. □

The preceding result can be extended to $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ by observing that the number of covariate clusters is bounded by 2^{a_1}. Lemma 4.9 does not imply that BOSS with $\mathcal{I}_{\text{Diff}}$ is fixed-parameter tractable because the fixed parameters appear as an exponent for the free parameters (specifically, s). The algorithm provided in the proof is not particularly useful, either, even for very small values of a_1 and a_2. For example, with $|\mathcal{P}| = 3$ and $|N_i| \leq 2$ for all $i \in \mathcal{P}$, the bound on the number of unique multisets of bin vectors is $O\Big((s+7)^7\Big)$. Improving this bound is left as a direction for future research.

4.1.3 Other Distribution Imbalance Measures

Theorem 4.10. *The decision problem for BOSS with a target size and $\mathcal{I}_{KS}$, $\mathcal{I}_{KS:\max}$, $\mathcal{I}_{ecdf:\mathbf{D}}$, $\mathcal{I}_{Diff^2}$, or $\mathcal{I}_{\chi^2}$ is **NP-Hard**.*

Proof. The result can be shown using a two-step reduction from the 3-Dimensional Matching (3DM) problem to BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ to BOSS with one of $\mathcal{I}_{\text{KS}}$, $\mathcal{I}_{\text{KS:max}}$, $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$, $\mathcal{I}_{\text{Diff}^2}$, or $\mathcal{I}_{\chi^2}$. The first part of the reduction from an arbitrary instance of 3DM to BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size is described in the proof of Theorem 4.4. The constructed BOSS instance has three covariates, three covariate clusters (one for each of the covariates), a target control group size equal to the size of the treatment group, and a desired imbalance of $\gamma = 0$. By design, $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0$ if and only if $\eta_{ij}(T) = \eta_{ij}(C')$ for all $i \in \mathcal{P}$ and $j \in N_i$.

The reduction from BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ to BOSS with $\mathcal{I}_{\text{Diff}^2}$ is straightforward: everything is preserved except the imbalance measure. By design, $\mathcal{I}_{\text{Diff}^2}(T, C') = 0$ if and only if $\eta_{ij}(T) = \eta_{ij}(C')$ for all $i \in \mathcal{P}$ and $j \in N_i$. Hence, a control group C' satisfies $\mathcal{I}_{\text{Diff}^2}(T, C') = 0$ if and only if it satisfies $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0$. This demonstrates that BOSS with $\mathcal{I}_{\text{Diff}^2}$ can be used to solve the 3DM instance, and so it is **NP-Hard**. The reduction to BOSS with $\mathcal{I}_{\chi^2}$ is similar. By design, $\mathcal{I}_{\chi^2}(T, C') = 0$ if and only if $|C' \cap B_{ij}| = |T \cap B_{ij}|$. Because the target control group size equals the size of the treatment group, a feasible control group C' satisfies $\mathcal{I}_{\chi^2}(T, C') = 0$ if and only if it satisfies $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0$. Hence, BOSS with $\mathcal{I}_{\chi^2}$ is also **NP-Hard**.

The reduction from BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ to BOSS with $\mathcal{I}_{\text{KS}}$ is similar, except that the bin information is ignored while the units' covariate values are retained. For a 3DM instance with d elements per set, the corresponding BOSS instance has units with covariate values in $\{1, 2, \ldots, d\}$ for each covariate. By the definitions of the bins in the reduction, any $S \subseteq T \cup C$ satisfies

$$\begin{aligned}
&\eta_{i1}(S) = \widehat{F}_i(S, 1) \\
&\eta_{ij}(S) = \widehat{F}_i(S, j) - \widehat{F}_i(S, j-1) \quad \forall\, j \in \{2, 3, \ldots, d\},
\end{aligned}$$

for all $i \in \mathcal{P}$, which also ensures that

$$\widehat{F}_i(S, j) = \sum_{j' \in N_i : j' \leq j} \eta_{ij}(S) \quad \forall\ i \in \mathcal{P},\ j \in N_i.$$

So for any covariate $i \in \mathcal{P}$, $\widehat{F}_i(T, x) = \widehat{F}_i(C', x)$ for all $x \in \mathcal{X}_i(T \cup C')$ if and only if $\eta_{ij}(T) = \eta_{ij}(C')$ for all $j \in N_i$. By design, $\mathcal{I}_{\text{KS}}(T, C') = 0$ if and only if

$$\sup_{x \in \mathcal{X}_i(T \cup C)} \left| \widehat{F}_i(T, x) - \widehat{F}_i(C', x) \right| = 0$$

for all $i \in \mathcal{P}$, which occurs if and only if $\widehat{F}_i(T, x) = \widehat{F}_i(C', x)$ for all $i \in \mathcal{P}$ and $x \in \mathcal{X}_i(T \cup C)$. Hence, a control group C' can satisfy $\mathcal{I}_{\text{KS}}(T, C') = 0$ if and only if it satisfies $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 0$. Hence, BOSS with $\mathcal{I}_{\text{KS}}$ can be used to solve the 3DM instance and so it is **NP-Hard**. An analogous argument applies to $\mathcal{I}_{\text{KS:max}}$. The result for $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$ follows because $\mathcal{I}_{\text{KS}}$ is a special case of it. □

Corollary 4.11. *The decision problems for BOSS with a target size and $\mathcal{I}_{KS}$, $\mathcal{I}_{KS:\max}$, $\mathcal{I}_{ecdf:\mathbf{D}}$, $\mathcal{I}_{Diff^2}$, or $\mathcal{I}_{\chi^2}$ remain* ***NP-Hard*** *for instances with only three covariates or instances with at most two values per covariate.*

4.2 Approximation Results

Given the intractability of BOSS with a variety of imbalance measures, a study of approximation methods is reasonable. Specifically, one might look for an approximation algorithm that achieves a relative performance guarantee of α with respect to the optimal value. As the optimization problems associated with BOSS are naturally minimization problems, such an algorithm would be able to identify a control group C' that satisfies $\mathcal{I}(T, C') \leq \alpha\,\text{OPT}$, where $\text{OPT} \equiv \min_{C'' \subseteq C} \mathcal{I}(T, C'')$. However, such an algorithm is unlikely to exist.

Lemma 4.12. *Unless* ***P = NP****, there is no polynomial-time approximation algorithm with factor α for BOSS with $\mathcal{I}_{Diff:\mathbf{D}}$ and a target size, for any $\alpha > 1$.*

Proof. Consider the 3DM reduction from the proof of Theorem 4.4. For an instance of 3DM, the corresponding decision problem for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ asks if there is a control group C' with $|C'| = s$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') \leq \gamma = 0$. If such a group exists, then $\text{OPT} = 0$ and an α-approximation algorithm for BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and a target size would return a solution C' satisfying $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') \leq \alpha\,\text{OPT} = 0$. If no such group exists, then $\text{OPT} > 0$, and an α-approximation algorithm would return a solution C' satisfying

$0 < \mathrm{OPT} \leq \mathcal{I}_{\mathrm{Diff}:\mathbf{D}}(T, C') \leq \alpha\,\mathrm{OPT}$. In either case, the solution returned by the α-approximation algorithm could be used to determine whether or not the 3DM instance has a solution in polynomial time. Because 3DM is **NP-Hard**, this would imply that $\mathbf{P} = \mathbf{NP}$. □

Corollary 4.13. *Unless $\boldsymbol{P} = \boldsymbol{NP}$, there is no polynomial-time approximation algorithm with factor α for BOSS with a target size and $\mathcal{I}_{DOM}$, $\mathcal{I}_{KS}$, $\mathcal{I}_{KS:\max}$, $\mathcal{I}_{ecdf:\mathbf{D}}$, $\mathcal{I}_{Diff^2}$, or $\mathcal{I}_{\chi^2}$, for any $\alpha > 1$.*

Despite the negative result from Lemma 4.12, alternate imbalance measures may be more amenable to approximation. The remainder of this section investigates this possibility for a balance measure that is related to $\mathcal{I}_{\mathrm{Diff}:\mathbf{D}}$.

4.2.1 Maximizing Balance

The difficulty with approximating an imbalance measure is that the ideal case of no imbalance does not allow any room for an approximation algorithm with a relative performance guarantee to return a sub-optimal solution. To remedy this, one can use a balance measure that assesses similarity instead of dissimilarity, which shifts the goal from minimization to maximization.

One such balance measure can be created from $\mathcal{I}_{\mathrm{Diff}:\mathbf{D}}$ after a few observations. The first observation was mentioned earlier in (4.6): for any control group C', the total bin excess is equal to the total bin shortage for each $D \in \mathbf{D}$. Therefore, either the bin excess penalty or the bin shortage penalty can be removed without impacting the quality of the solutions.

The second observation builds on the first by shifting from a *penalty* for bin shortage to a *reward* for *bin coverage*. The following discussion focuses on BOSS with a target control group size, s. For each $D \in \mathbf{D}$ and $j \in N_D$, define the *demand* (with respect to a fixed treatment group T) as

$$r_{Dj} \equiv s \cdot \eta_{Dj}(T) = s \cdot |T \cap B_{Dj}|/|T|,$$

which represents the desired number of control units in C' that should belong to bin j of cluster D. For a control group C' of any size, define the *coverage* (with respect to a fixed treatment group T) as

$$w_{Dj}(C') \equiv \min\left(|C' \cap B_{Dj}|, r_{Dj}\right)$$

for all $D \in \mathbf{D}$ and $j \in N_D$.

Any control group C' with $|C'| = s$ satisfies

$$\begin{aligned} |C'| \cdot \Delta^-_{Dj}(C') &= |C'| \cdot \max\left(0, |T \cap B_{Dj}|/|T| - |C' \cap B_{Dj}|/|C'|\right) \\ &= \max\left(0, s \cdot |T \cap B_{Dj}|/|T| - |C' \cap B_{Dj}|\right) \\ &= \max\left(0, r_{Dj} - |C' \cap B_{Dj}|\right) \\ &= r_{Dj} - \min\left(r_{Dj}, |C' \cap B_{Dj}|\right) \\ &= r_{Dj} - w_{Dj}(C'), \end{aligned} \tag{4.10}$$

which can be seen by examining the relationship between r_{Dj} and $|C' \cap B_{Dj}|$. If $r_{Dj} \geq |C' \cap B_{Dj}|$, then

$$\max\left(0, r_{Dj} - |C' \cap B_{Dj}|\right) = r_{Dj} - |C' \cap B_{Dj}| = r_{Dj} - \min\left(r_{Dj}, |C' \cap B_{Dj}|\right).$$

In the case when $r_{Dj} < |C' \cap B_{Dj}|$, then

$$\max\left(0, r_{Dj} - |C' \cap B_{Dj}|\right) = 0 = r_{Dj} - r_{Dj} = r_{Dj} - \min\left(r_{Dj}, |C' \cap B_{Dj}|\right).$$

Then (4.7) and (4.10) can be used to show that any C' with $|C'| = s$ satisfies

$$\begin{aligned} \mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') &= 2 \cdot \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \Delta^-_{Dj}(C') \\ &= \frac{2}{|C'|} \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \left(r_{Dj} - w_{Dj}(C')\right) \\ &= \frac{2}{s} \sum_{D \in \mathbf{D}} \sum_{j \in N_D} \frac{s \cdot |T \cap B_{Dj}|}{|T|} - \frac{2}{s} \sum_{D \in \mathbf{D}} \sum_{j \in N_D} w_{Dj}(C') \\ &= 2 \cdot |\mathbf{D}| - \left(\frac{2}{s}\right) \sum_{D \in \mathbf{D}} \sum_{j \in N_D} w_{Dj}(C'). \end{aligned} \tag{4.11}$$

Because the coverage values are always nonnegative and $|\mathbf{D}|$ is constant, minimizing $\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C')$ over all control groups with size s is equivalent to maximizing the sum of the coverage values across all bins and clusters. Thus, BOSS with $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ can be reformulated as a maximization problem with the *coverage balance measure*:

$$\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C') \equiv \sum_{D \in \mathbf{D}} \sum_{j \in N_D} w_{Dj}(C'),$$

where $\mathcal{I}^+$ denotes a balance measure that should be maximized.

4.2.2 Approximability

The reformulated problem for BOSS with $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ can be viewed as a generalized set cover problem since it combines aspects of two well-known problems: Set Multicover, in which each element needs to be covered a specified number of times, and Maximum Coverage, in which the number of sets that can be used in the cover is constrained and the goal is to cover as many elements as possible. In this case, the elements correspond to the bins for each covariate cluster, and r_{Dj} specifies the number of times the associated bin (element) needs to be covered. The sets available for covering elements correspond to control units with their elements determined by the bins to which the control unit belongs.

Both Set Multicover and Maximum Coverage have been well-studied from an approximation perspective (Vazirani, 2001). A generalization of these two problems, which also captures BOSS with $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$, is the Submodular Maximum Coverage problem. This problem is: Given a universe V of elements, a monotone submodular function $f : 2^V \to [0, \infty)$, and an integer d, pick a set $V' \subseteq V$ of size at most d that maximizes $f(V')$. A function f is monotone if $f(Q) \leq f(R)$ whenever $Q \subseteq R$. A function f is submodular if and only if $f(Q)+f(R) \geq f(Q \cup R)+f(Q \cap R)$ for all $Q, R \subseteq V$; equivalently, $f(Q \cup \{v\}) - f(Q) \geq f(R \cup \{v\}) - f(R)$ for all $Q \subseteq R \subseteq V$ and $v \in V$. Note that since the number of subsets of V is exponential, the function f is specified as a value oracle; that is, f is specified as a polynomial-time subroutine that will return the function value for any given $V' \subseteq V$. The Submodular Maximum Coverage problem can be approximated using the standard greedy algorithm presented in Algorithm 1. Nemhauser et al. (1978) proved that Algorithm 1 achieves an approximation ratio of $1 - \frac{1}{e}$ for Submodular Maximum Coverage if $f(\emptyset) = 0$.

Algorithm 1 Greedy Algorithm for Submodular Maximum Coverage

```
procedure GreedyCover($V$, $f$, $d$)
    $V' \leftarrow \emptyset$
    while $|V'| < d$ do
        $v' \leftarrow \arg\max_{v \in V \setminus V'} \{f(V' \cup \{v\}) - f(V')\}$
        $V' \leftarrow V' \cup \{v'\}$
    end while
    return $v'$
end procedure
```

Lemma 4.14. *For a fixed treatment group T, control pool C, histogram bins, and covariate clusters, the function $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ is both monotone and submodular.*

Proof. To show that $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ is monotone, it suffices to observe that for any $C'' \subset C' \subseteq C$, $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}(T, C'') \leq \mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}(T, C')$ since $|C'' \cap B_{Dj}| \leq |C' \cap B_{Dj}|$ for all $D \in \mathbf{D}$ and $j \in N_D$.

To show that $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ is submodular, let $C'' \subset C' \subset C$ be arbitrary sets and let $c \in C \setminus C'$. Since

$C'' \subset C'$, it follows that $|C'' \cap B_{Dj}| \leq |C' \cap B_{Dj}|$ for all $D \in \mathbf{D}$ and $j \in N_D$. Now consider the marginal improvement that c can provide when added to C' compared to when c is added to C''. The unit c can improve $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C')$ by fulfilling some unsatisfied demand for some bin $j \in N_D$ of a cluster $D \in \mathbf{D}$. This requires that $|C' \cap B_{Dj}| < r_{Dj}$. But then unit c also fulfills unsatisfied demand for C'', since $|C'' \cap B_{Dj}| \leq |C' \cap B_{Dj}|$. Hence

$$\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C'' \cup \{c\}) - \mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C'') \geq \mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C' \cup \{c\}) - \mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C').$$

As this holds for all $C'' \subset C' \subset C$ and $c \in C \setminus C'$, it follows that $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ is submodular. □

As a result of Lemma 4.14, BOSS with $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ and a target control group size can be viewed as an instance of Submodular Maximum Coverage. This leads to the following theorem.

Theorem 4.15. *Algorithm 1 yields a $\left(1 - \frac{1}{e}\right)$-approximation for BOSS with $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ and a target size.*

Proof. Follows from Lemma 4.14 and the approximation ratio for Submodular Maximum Coverage (Nemhauser et al., 1978). □

The approximation ratio holds for $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ but not for $\mathcal{I}_{\text{Diff:}\mathbf{D}}$. However, (4.11) establishes that $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ are related by

$$\mathcal{I}_{\text{Diff:}\mathbf{D}}(T, C') = 2 \cdot |\mathbf{D}| - \left(\frac{2}{s}\right) \cdot \mathcal{I}^+_{\text{Cvg:}\mathbf{D}}(T, C').$$

This means that the relative quality of C' with respect to all other possible solutions remains the same under either $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ or $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ (i.e., if C' is the ith best solution for one objective, it is also the ith best solution for the other objective, ignoring ties).

4.2.3 Inapproximability

Theorem 4.15 demonstrates that BOSS with $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ and a target size can be approximated to within a factor of $\left(1 - \frac{1}{e}\right)$ by using the greedy algorithm for Submodular Maximum Coverage, but it does not say whether or not it is possible to improve upon this. The following result clarifies this point.

Theorem 4.16. *Unless $\mathbf{P} = \mathbf{NP}$, BOSS with $\mathcal{I}^+_{\text{Cvg:}\mathbf{D}}$ and a target size is inapproximable to within $\left(1 - \frac{1}{e} + \epsilon\right)$ for any fixed $\epsilon > 0$.*

Proof. The result can be shown by using an approximation-preserving reduction from the Maximum Coverage problem, which is inapproximable to within $\left(1 - \frac{1}{e} + \epsilon\right)$ for any fixed $\epsilon > 0$, unless $\mathbf{P} = \mathbf{NP}$ (Feige, 1998, Thm. 5.3). The Maximum Coverage problem is: Given a universe of elements V, a collection $\mathcal{W}$ of subsets of V, and an integer d, find a collection $\mathcal{W}' \subseteq \mathcal{W}$ of subsets with $|\mathcal{W}'| \leq d$ that maximizes $|\cup_{W \in \mathcal{W}'} W|$.

For an arbitrary instance of Maximum Coverage with elements $V \equiv \{1, 2, \ldots, m\}$, collection of subsets $\mathcal{W}$, and an integer d, the reduction proceeds as follows. Construct a BOSS instance with $\mathcal{P} \equiv \{1, 2, \ldots, m\}$. Construct the treatment group T with d units, indexed from 1 through d, with $x_{t_1 i} = 1$ for all $i \in \mathcal{P}$ and $x_{t_j i} = 2$ for all $i \in \mathcal{P}$ and $j \in \{2, 3, \ldots, d\}$. For each $W \in \mathcal{W}$, the control pool C contains a control unit c_V with $x_{c_V i} = 1$ if $i \in V$ and $x_{c_V i} = 3$ otherwise for each $i \in \mathcal{P}$. Each covariate $i \in \mathcal{P}$ has $n_i = 3$ histogram bins, with the bin boundaries given by $b_{i0} = 1$, $b_{i1} = 1.5$, $b_{i2} = 2.5$, $b_{i3} = 3.0$. Under these definitions, the units in each histogram bin are

$$\begin{aligned} B_{i1} &\equiv \{t_1\} \cup \{c_W \in C : i \in W\}, \\ B_{i2} &\equiv \{t_2, t_3, \ldots, t_d\}, \\ B_{i3} &\equiv \{c_W \in C : i \notin W\}. \end{aligned}$$

To complete the reduction, let $\mathbf{D} \equiv \{\{i\} : i \in \mathcal{P}\}$ and set the target control group size at $s = d$. By construction, the demand is $r_{D1} = 1$, $r_{D2} = d - 1$, and $r_{D3} = 0$ for all $D \in \mathbf{D}$.

All steps for the reduction can be finished in time polynomial in the size of the Maximum Coverage instance. To see that the reduction is approximation preserving, it suffices to observe that the BOSS instance has a solution C' satisfying $|C'| = s$ and $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}(T, C') = \gamma$ if and only if the Maximum Coverage instance has a solution covering γ elements.

($\Rightarrow$) Let $C' \subseteq C$ be a solution that satisfies $|C'| = s$ and $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}(T, C') = \gamma$. Because $r_{D3} = 0$ and $|C' \cap B_{D2}| = 0$ for all $D \in \mathbf{D}$, the only coverage that C' provides is in bin 1 of each cluster. Additionally, because $r_{D1} = 1$ for all $D \in \mathbf{D}$, C' must satisfy $|C' \cap B_{D1}| \geq 1$ for γ clusters in order to satisfy $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}(T, C') = \gamma$. For a cluster $D \equiv \{i\}$, in order for $|C' \cap B_{D1}| \geq 1$, there must exist a unit $c_W \in C'$ that satisfies $x_{c_W i} = 1$, which ensures that $i \in W$ by construction. Then $\mathcal{W}' \equiv \{W \in \mathcal{W} : c_W \in C'\}$ is a collection that satisfies

$$\left| \bigcup_{W \in \mathcal{W}'} W \right| = |\{i \in V : D \equiv \{i\}, |C' \cap B_{D1}| \geq 1\}| = \gamma.$$

($\Leftarrow$) Let $\mathcal{W}' \subseteq \mathcal{W}$ be a solution to the Maximum Coverage instance that covers γ elements. Without loss of generality, it is assumed that $|\mathcal{W}'| = d$, otherwise additional sets can be added to $\mathcal{W}'$ without decreasing the number of covered elements. Let $C' \equiv \{c_W \in C : W \in \mathcal{W}'\}$ be the control units corresponding to the sets in $\mathcal{W}'$. For each $i \in \cup_{W \in \mathcal{W}'} W$, there exists a $W \in \mathcal{W}'$ with $i \in W$. By construction, the corresponding control unit c_W satisfies $x_{c_W i} = 1$ and consequently $c_V \in B_{D1}$ for $D \equiv \{i\}$. This ensures that $|C' \cap B_{D1}| \geq 1$. As this holds for γ unique elements covered by $\mathcal{W}'$, it must be the case that $w_{Dj}(C') \geq 1$ for γ clusters. This means that $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}(T, C') = \gamma$, as desired.

The above approximation-preserving reduction implies that a $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for BOSS with $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ and a target size could be used to provide a $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for Maximum Coverage for any $\epsilon > 0$. Because Maximum Coverage cannot be approximated to this factor unless $\mathbf{P} = \mathbf{NP}$, no such approximation algorithm can exist for BOSS with $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ unless $\mathbf{P} = \mathbf{NP}$. □

An immediate consequence of Theorem 4.16 is that BOSS with $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ and a target size has no polynomial-time approximation scheme (PTAS) unless $\mathbf{P} = \mathbf{NP}$.

4.3 Comments

The hardness results in Section 4.1 rely on reductions from **NP-Hard** decision problems to BOSS instances that require zero imbalance. It would be useful to identify approximation-preserving reductions from **NP-Hard** optimization problems to BOSS instances with a minimization imbalance measure. However, as the results in Section 4.2 indicate, the structure of these imbalance measures is not immediately amenable to such reductions. One difficulty is that a reduction from an **NP-Hard** optimization problem to BOSS must encode both the constraints of the problem (with the exception of control group size) and the objective function information into the imbalance measure. Any imbalance measure that penalizes all imbalance equally would have difficulty avoiding control groups that violate some constraints but improve the objective function. This issue could be overcome by using an imbalance measure with separate penalties for each of the covariates.

Another possibility for making progress in identifying approximation-preserving reductions is to view BOSS as a vector problem. For example, BOSS with $\mathcal{I}_{\text{DOM}}$ and a target size can be restated in the following form: Given a set of vectors V in $\mathbb{R}^p$ and a target point $\mathbf{v}^* \in \mathbb{R}^p$, find a subset $V' \subseteq V$ of size s that minimizes $\left\|\mathbf{v}^* - \sum_{\mathbf{v} \in V'} \mathbf{v}\right\|_1$. Here the target point represents the mean covariate values for the treatment group after appropriate scaling. BOSS with $\mathcal{I}_{\text{Diff}}$ can be reformulated in a similar manner using binary vectors of dimension $\sum_{i \in \mathcal{P}} |N_i|$, as can $\mathcal{I}_{\text{Diff:}\mathbf{D}}$. If control units can be selected with repetition, then the resulting problems bear some similarity to lattice problems.

An additional possibility is to identify reductions from either **NP-Hard** decision problems or **NP-Hard** optimization problems to BOSS without a size constraint. As noted earlier, Subset Sum with a target sum of $d = 0$ can be reduced to BOSS with $\mathcal{I}_{\text{DOM}}$ and no size constraint, but no other reductions have been found.

Chapter 5

Computational Results

This chapter presents computational results for BOSS with various imbalance measures using both simulated and actual datasets. BOSS was solved with heuristics and exact algorithms. Comparisons with several matching approaches are also included.

5.1 Algorithms

Given the computational complexity of BOSS for various imbalance measures, heuristics and exact algorithms (e.g., branch-and-bound) are two ways in which BOSS can be solved in practice. Several approaches were investigated in order to solve the problem with a target control group size. Developing general techniques for BOSS without a target control group size is a direction for future research.

5.1.1 Simulated Annealing

Simulated annealing is a heuristic that transitions between solutions, or states, in a search space to seek a local minimum or maximum (Kirkpatrick et al., 1983). Beginning with a feasible solution, simulated annealing makes a series of random transitions from the current solution to a *neighboring* solution. An uphill transition is a move from a solution with a smaller objective function value to a neighboring solution with a larger objective value, while a downhill transition is a move that decreases the objective function value. For a minimization problem, uphill transitions are accepted randomly subject to a probability threshold. During the course of the search, the threshold for accepting an uphill transition increases, making them less likely to occur. This causes the search process to converge to a local optimum (where local is defined with respect to the neighbor function). Randomized restarts can be used to diversify the search by jumping to a new solution outside of the current neighborhood.

A simulated annealing heuristic for BOSS was implemented using the 1-exchange neighborhood defined

Some of the material in this chapter has been adapted from Sauppe et al. (2014), *INFORMS Journal on Computing* **26**(3), and Nikolaev et al. (2013), *Operations Research* **61**(2), with the permission of the copyright holders.

on a set of control units. Starting from an initial random control group C' with $|C'| = s$, simulated annealing attempts to swap a random unit $c_1 \in C'$ with a random unit $c_2 \in C \backslash C'$. The objective values of the solutions are their respective imbalances. Random restarts are employed if no progress has been made after a number of iterations or if a confirmed global optimum has been identified (i.e., if $\mathcal{I}(T, C') = 0$). The algorithm terminates after performing a pre-defined number of iterations. Details are provided in Algorithms 2 and 3.

5.1.2 Mathematical Programming Models

The BOSS framework with inputs T, C, a generic imbalance measure $\mathcal{I}$ and a target control group size s can be modeled as the following nonlinear integer program:

$$\min\ \mathcal{I}(\{\mathbf{x}_t : t \in T\}, \{\mathbf{x}_c : c \in C,\ v_c = 1\}) \quad \text{s.t.} \sum_{c \in C} v_c = s,\ v_c \in \{0, 1\}\ \forall\ c \in C. \tag{5.1}$$

In (5.1), each unit $c \in C$ has an associated binary decision variable v_c that indicates whether or not unit c is included in C'. For this formulation, the imbalance measure is considered to be a black box procedure, and as such, it is difficult to optimize in general. However, specific imbalance measures are more readily captured in mathematical programming formulations. For example, BOSS with $\mathcal{I}_{\mathrm{DOM}}$ and a target size can be expressed as the following mixed integer (linear) program (MIP):

$$\begin{aligned}
\min \quad & \sum_{i \in \mathcal{P}} w_i && (5.2a)\\
\text{s.t.} \quad & \sum_{c \in C} v_c = s && (5.2b)\\
& \frac{1}{s} \sum_{c \in C} v_c x_{ci} - \frac{1}{|T|} \sum_{t \in T} x_{ti} \leq w_i \qquad \forall\ i \in \mathcal{P} && (5.2c)\\
& \frac{1}{|T|} \sum_{t \in T} x_{ti} - \frac{1}{s} \sum_{c \in C} v_c x_{ci} \leq w_i \qquad \forall\ i \in \mathcal{P} && (5.2d)\\
& v_c \in \{0, 1\} \quad \forall\ c \in C. && (5.2e)
\end{aligned}$$

The binary decision variables v_c in (5.2e) indicate whether or not each $c \in C$ is included in C'. For each covariate $i \in \mathcal{P}$, the continuous variable w_i is constrained by (5.2c) and (5.2d) to be at least the difference in covariate means between the selected control group and the treatment group. The objective (5.2a) is to minimize the sum of the w variables. Appropriate scaling of the objective yields a model for BOSS with $\mathcal{I}_{\mathrm{SDOM}}$. The constraint (5.2b) ensures that exactly s control units are included in the control group. Solvers

Algorithm 2 Simulated annealing heuristic for BOSS.

```
procedure simulatedAnnealing($T$, $C$, $\mathcal{I}$, $s$)
    $C' \leftarrow$ a random subset of $C$ with size $s$
    $C^* \leftarrow C'$
    Initialize the search parameters:          ▷ Set by user; these values were determined through tests
        initialTemperatureFactor $\leftarrow 0.01$
        coolingRate $\leftarrow 0.975$
        totalCoolingPeriods $\leftarrow 800$
        coolingPeriodLength $\leftarrow 5000$
        restartAfterXFailures $\leftarrow 5000$
        restartAfterXOptSols $\leftarrow 1$
    Initialize the search counters and variables:
        iteration $\leftarrow 0$
        coolingPeriodsCompleted $\leftarrow 0$
        consecutiveFailedMoved $\leftarrow 0$
        zeroObjSols $\leftarrow 0$
        temperature $\leftarrow \mathcal{I}(T, C')\cdot$ initialTemperatureFactor
    do
        if (iteration  mod coolingPeriod) = 0 then
            temperature $\leftarrow$ temperature $\cdot$ coolingRate
            coolingPeriodsCompleted $\leftarrow$ coolingPeriodsCompleted $+1$
        end if
        $C' \leftarrow$ performRandomMove($T$, $C$, $\mathcal{I}$, $C'$, temperature)
        if move was accepted then
            consecutiveFailedMoves $\leftarrow 0$
            if $\mathcal{I}(T, C') < \mathcal{I}(T, C^*)$ then $C^* \leftarrow C'$ end if
            if $\mathcal{I}(T, C') = 0$ then zeroObjSols $\leftarrow$ zeroObjSols $+1$ end if
        else
            consecutiveFailedMoves $\leftarrow$ consecutiveFailedMoves $+1$
        end if
        if (consecutiveFailedMoves $\geq$ restartAfterXFailures)
          or (zeroObjSols $>$ restartAfterXOptSols) then
            $C' \leftarrow$ a random subset of $C$ with size $s$
            if $\mathcal{I}(T, C') < \mathcal{I}(T, C^*)$ then $C^* \leftarrow C'$ end if
            consecutiveFailedMoves $\leftarrow 0$
            zeroObjSols $\leftarrow 0$
        end if
        iteration $\leftarrow$ iteration $+1$
    while coolingPeriodsCompleted $<$ totalCoolingPeriods
    return $C^*$
end procedure
```

Algorithm 3 Transition step for simulated annealing.

```
procedure performRandomMove($T$, $C$, $\mathcal{I}$, $C'$, temperature)
    $c \leftarrow$ a random unit in $C \setminus C'$
    $c' \leftarrow$ a random unit in $C'$
    $C'' \leftarrow \{c\} \cup C' \setminus \{c'\}$
    $\delta \leftarrow \mathcal{I}(T, C') - \mathcal{I}(T, C'')$
    if $\delta > 0$ then                                              ▷ Downhill Move
        acceptanceProb $\leftarrow$ 1.0
    else if temperature $> 0.000001$ then
        acceptanceProb $\leftarrow exp\,(\delta/\text{temperature})$
    else
        acceptanceProb $\leftarrow$ 0.0
    end if
    if ($random[0,1] \leq$ acceptanceProb) then return $C''$        ▷ Move to the new solution
    else return $C'$ end if                                 ▷ Else stay at the original solution
end procedure
```

from integer and linear programming can be applied to these models to solve BOSS with $\mathcal{I}_{\text{DOM}}$ or $\mathcal{I}_{\text{SDOM}}$.

The problem of minimizing $\mathcal{I}_{\text{KS}}$ for BOSS with a target size can also be formulated as a mixed integer program. Recall that $\mathcal{X}_i(T \cup C)$ is the set of unique values that units in $T \cup C$ attain for covariate $i \in \mathcal{P}$. Then a MIP model for optimizing $\mathcal{I}_{\text{KS}}$ is

$$\min \quad \sum_{i \in \mathcal{P}} w_i \tag{5.3a}$$

$$\text{s.t.} \quad \sum_{c \in C} v_c = s \tag{5.3b}$$

$$\frac{1}{s} \sum_{c \in C : x_{ci} \leq x} v_c - \widehat{F}_i(T, x) \leq w_i \qquad \forall\ i \in \mathcal{P},\ x \in \mathcal{X}_i(T \cup C) \tag{5.3c}$$

$$\widehat{F}_i(T, x) - \frac{1}{s} \sum_{c \in C : x_{ci} \leq x} v_c \leq w_i \qquad \forall\ i \in \mathcal{P},\ x \in \mathcal{X}_i(T \cup C) \tag{5.3d}$$

$$v_c \in \{0, 1\} \quad \forall\ c \in C. \tag{5.3e}$$

The binary decision variables in (5.3e) are identical to those in (5.2e). For each covariate $i \in \mathcal{P}$, the continuous variable w_i is constrained by (5.3c) and (5.3d) to be at least the value of $\left|\widehat{F}_i(T, x) - \widehat{F}_i(C', x)\right|$ for all possible values $x \in \mathcal{X}_i(T \cup C)$. By minimizing (5.3a), an optimal solution will set each w_i to be exactly the maximum value of $\left|\widehat{F}_i(T, x) - \widehat{F}_i(C', x)\right|$ over all $x \in \mathcal{X}_i(T \cup C)$. Finally, (5.3b) ensures that the control group is of the desired size. Model (5.3) can be extended to BOSS with $\mathcal{I}_{\text{KS:max}}$ through minor modifications.

The number of constraints in (5.3c) and (5.3d) is

$$2\sum_{i\in\mathcal{P}} |\mathcal{X}_i(T \cup C)| \in O(|\mathcal{P}| \cdot |T \cup C|)$$

because there may be many unique values for each covariate (which is likely to be the case for continuous covariates). This can be quite large and result in significant memory requirements for MIP solvers. One way this can be addressed is to revise the constraints (5.3c) and (5.3d) to only assess imbalance on a smaller subset of the covariate values. For example, coarsening the values reduces the number of unique values per covariate and leads to a smaller MIP model. Such an approach was used by Zubizarreta (2012).

An alternate approach to handling the large number of constraints in Model (5.3) is to recognize that many of the constraints in (5.3c) and (5.3d) will not be active at any given solution, and thus, they may not ever be needed in the model. So instead of explicitly including all constraints in the model prior to optimization, the constraints can be included on an as-needed basis during the optimization process. This is similar to row generation techniques for linear and integer programming. As an example, Model (5.3) can be initialized with only a subset of the constraints in (5.3c) and (5.3d). Then during optimization, once an integer solution is found, the excluded constraints from (5.3c) and (5.3d) can be examined for violations. Violated constraint are added to the model, which is then re-optimized. This process is repeated until an integer solution with no violated constraints is identified. The mixed integer programming solver CPLEX provides this capability through the *Lazy Constraint Callback* feature.

In addition to the large number of constraints in Model (5.3), there is also a large number of nonzeros in the associated constraint matrix. This can dramatically increase memory requirements for MIP solvers. A reformulated model with an increased number of rows and columns but fewer nonzero elements can be used to decrease memory requirements. Let

$$\min\ \mathcal{X}_i(T \cup C) \equiv x_{i1} < x_{i2} < \ldots < x_{ik_i} \equiv \max\ \mathcal{X}_i(T \cup C)$$

be the k_i unique values in $\mathcal{X}_i(T \cup C)$ for covariate $i \in \mathcal{P}$. A sparse model can be created by adding the continuous variables q_{ij} for each $i \in \mathcal{P}$ and $j \in \{1, 2, \ldots, k_i\}$ and replacing the constraints (5.3c) and (5.3d) with

$$\begin{aligned}
\sum_{c \in C : x_{ci} = x_{i1}} v_c &= q_{i1} \quad \forall\ i \in \mathcal{P} \\
\sum_{c \in C : x_{ci} = x_{ij}} v_c + q_{i,j-1} &= q_{ij} \quad \forall\ i \in \mathcal{P},\ j \in \{2, 3, \ldots, k_i\} \\
\frac{q_{ij}}{s} - \widehat{F}_i(T, x_{ij}) &\le w_i \quad \forall\ i \in \mathcal{P},\ j \in \{1, 2, \ldots, k_i\} \\
\widehat{F}_i(T, x_{ij}) - \frac{q_{ij}}{s} &\le w_i \quad \forall\ i \in \mathcal{P},\ j \in \{1, 2, \ldots, k_i\}.
\end{aligned}$$

BOSS with $\mathcal{I}_{\text{Diff}}$ and a target size can be formulated as

$$\min \quad \sum_{i \in \mathcal{P}} \sum_{j \in N_i} w_{ij} \tag{5.4a}$$

$$\text{s.t.} \quad \sum_{c \in C} v_c = s \tag{5.4b}$$

$$\frac{1}{s} \sum_{c \in C \cap B_{ij}} v_c - \eta_{ij}(T) \le w_{ij} \qquad \forall\ i \in \mathcal{P},\ j \in N_i \tag{5.4c}$$

$$\eta_{ij}(T) - \frac{1}{s} \sum_{c \in C \cap B_{ij}} v_c \le w_{ij} \qquad \forall\ i \in \mathcal{P},\ j \in N_i \tag{5.4d}$$

$$v_c \in \{0, 1\} \quad \forall\ c \in C. \tag{5.4e}$$

In Model (5.4), there is a continuous variable w_{ij} that is constrained by (5.4c) and (5.4d) to be at least the imbalance between the treatment group and the selected control group for bin $j \in N_i$ of covariate $i \in \mathcal{P}$. Minimizing (5.4a) ensures that an optimal solution will set each w_{ij} to be exactly $|\eta_{ij}(T) - \eta_{ij}(C')|$. Model (5.4) requires $O(n \cdot |\mathcal{P}|)$ constraints, where $n \equiv \max_{i \in \mathcal{P}} n_i$. Depending on the granularity of the bins, this can be significantly smaller than the $O(|\mathcal{P}| \cdot |T \cup C|)$ constraints required by Model (5.3).

Models (5.3) and (5.4) can be extended to handle $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathcal{I}_{\text{Diff:}\mathbf{D}}$, respectively, for any set of covariate clusters $\mathbf{D}$. For $\mathcal{I}_{\text{Diff:}\mathbf{D}}$, constraints (5.4c) and (5.4d) are modified to include the covariate clusters and their associated bins. For $\mathcal{I}_{\text{ecdf:}\mathbf{D}}$, constraints (5.3c) are modified to

$$\frac{1}{s} \sum_{\substack{c \in C: \\ \mathbf{P}^D \mathbf{x}_c \le \mathbf{P}^D \mathbf{x}}} v_c - \widehat{F}_D(T, \mathbf{x}) \le w_D \quad \forall\ D \in \mathbf{D},\ \mathbf{x} \in \mathcal{X}_D(T \cup C), \tag{5.5}$$

with similar modifications for (5.3d). For a covariate cluster D, the number of constraints in (5.5) is

$$|\mathcal{X}_D(T \cup C)| = \prod_{i \in D} |\mathcal{X}_i(T \cup C)| \in O\Big(|T \cup C|^{|D|}\Big),$$

which can become quite large depending on the number of covariates in the cluster. In contrast, the total number of constraints in the MIP model for $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ is $O(|\mathbf{D}| \cdot |T \cup C|)$ because only the occupied bins need to be checked for each cluster.

BOSS with $\mathcal{I}_{\text{Diff}^2}$ can be formulated as a mixed integer quadratic program (MIQP) due to the quadratic penalty terms in the objective. This also allows the sign of the imbalance variables w_{ij} to be ignored.

$$\min \quad \sum_{i \in \mathcal{P}} \sum_{j \in N_i} w_{ij}^2 \tag{5.6a}$$

$$\text{s.t.} \quad \sum_{c \in C} v_c = s \tag{5.6b}$$

$$\frac{1}{s} \sum_{c \in C \cap B_{ij}} v_c - \eta_{ij}(T) = w_{ij} \qquad \forall\, i \in \mathcal{P},\ j \in N_i \tag{5.6c}$$

$$v_c \in \{0,1\} \quad \forall\, c \in C. \tag{5.6d}$$

A model for BOSS with $\mathcal{I}_{\chi^2}$ can be constructed through minor modifications.

5.1.3 Implementation

The above algorithms and models were implemented in C++ and interfaced with both R and CPLEX. The computational tests reported here were obtained across several Linux desktops with marginal differences in hardware characteristics. All machines had an Intel Core i7 2.67 GHz or 2.8 GHz quad core processor with hyper-threading, and all had either 6 or 12 GB of RAM. Time limits reported for CPLEX are specified in user time, while computing times reported by CPLEX are specified in CPU time. Due to CPLEX's ability to parallelize computations, many reported running times exceed the stated time limit.

In all experiments, the target control group size was set to the size of the treatment group. A preprocessing step was used to bin the data into uniform-width bins for any appropriate imbalance measure. Specifically, for covariate $i \in \mathcal{P}$ with n_i total bins, the bin boundaries were set to

$$b_{ij} \equiv \min \mathcal{X}_i(T \cup C) + j \cdot \left(\max \mathcal{X}_i(T \cup C) - \min \mathcal{X}_i(T \cup C)\right)/n_i \tag{5.7}$$

for each $j \in \{0, 1, \dots, n_i\}$. The values for n_i were varied in the experiments and are reported in the results. Unless otherwise specified, the covariate clusters were $\mathbf{D} \equiv \{\{i\} : i \in \mathcal{P}\}$.

For Models (5.3), (5.4), and (5.6), the implementations used a modified version of the imbalance constraints presented in Section 5.1.2. Specifically, for Model (5.4), the constraints (5.4c) and (5.4d) were written in the form

$$\sum_{c \in C \cap B_{ij}} v_c - s \cdot \eta_{ij}(T) \leq w_{ij} \quad \forall\, i \in \mathcal{P},\ j \in N_i \tag{5.8a}$$

$$s \cdot \eta_{ij}(T) - \sum_{c \in C \cap B_{ij}} v_c \leq w_{ij} \quad \forall\, i \in \mathcal{P},\ j \in N_i. \tag{5.8b}$$

This causes the objective function for the resulting model to be scaled by s with respect to the original objective (5.4a). When s is a multiple of $|T|$, the constraints in (5.8) have integer coefficients and the resulting objective function is integer-valued. This formulation was found to perform better in practice, as well. Similar modifications were applied to Models (5.3) and (5.6). In all cases, the original imbalance measures can be recovered after optimization by appropriate scaling (s^{-1} in the case of $\mathcal{I}_{\text{KS}}$ and $\mathcal{I}_{\text{Diff}}$, and s^{-2} in the case of $\mathcal{I}_{\text{Diff}^2}$).

5.2 Simulated Data

A variety of simulated datasets were created in order to test BOSS. By varying the properties of these datasets, the performance of BOSS with various imbalance measures was assessed in well-characterized settings. In particular, knowledge of the control response function allows the results from Chapter 3 to be demonstrated empirically.

5.2.1 Experiments with Heuristics

Preliminary computational tests were conducted to illustrate the potential for the BOSS framework. In these tests, the simulated annealing algorithm shown in Algorithm 2 was used as a first attempt at optimization.

Data Generation

Two samples (T, C) of 500 treatment units and 10,000 control units were created. The first set of samples included three covariates for each unit, while the second included ten covariates. Each sample was created by first randomly generating a pool of 5,000 potential treatment units and a pool of 10,000 control units, with the covariate values for each unit drawn from a normal distribution. Each treatment group of 500 units was drawn randomly but non-uniformly from the pool of potential treatment units. Units with covariate values in the tails of the covariate distribution were drawn with higher probability than those with values in the center of the distributions, ensuring that the resulting sets of treatment and control units had different covariate distributions.

Figure 5.1 shows the empirical cumulative distributions of covariate values in the treatment group and control pool in the sample with 3 covariates. The covariate distributions of the treatment group differ from those of the control pool, particularly for the first two covariates.

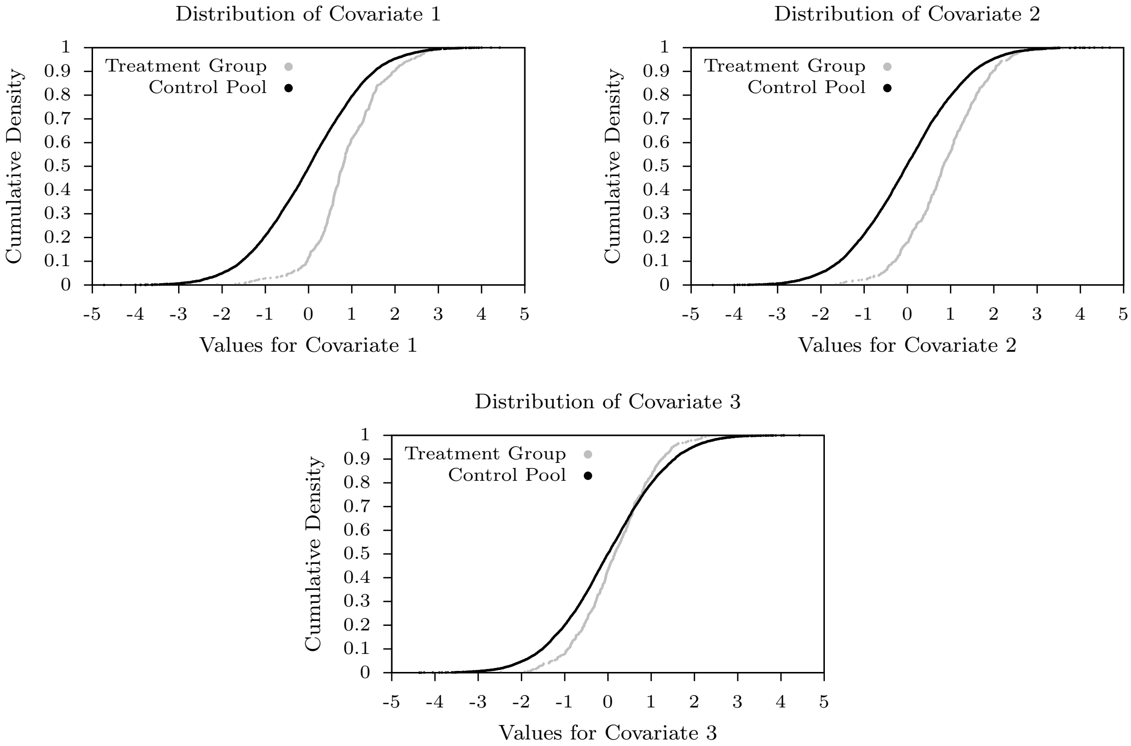


Figure 5.1: Initial covariate distributions of the treatment group and control pool for the 3-covariate sample.

After generating the set of treatment units in each sample, treatment and control responses were assigned to the units. The treatment effect was set to be zero, so that $y_u^1 = y_u^0$ for each unit $u \in T \cup C$. The control response errors ε_u^0 for each $u \in T \cup C$ were drawn from a normal distribution with mean zero and standard deviation 2. For the 3-covariate sample, two response functions were considered. The first response function was linear:

$$y_u^1 \equiv y_u^0 \equiv 10 + 7x_{u1} + 6x_{u2} + 5x_{u3} + \varepsilon_u^0. \tag{5.9}$$

The resulting dataset consisting of units with covariate values and observed responses was labeled *data3c10k*. The second response function was nonlinear:

$$y_u^1 \equiv y_u^0 \equiv 10 + \exp(x_{u1}) + x_{u2}^2 + 0.1x_{u3}^3 + \varepsilon_u^0. \tag{5.10}$$

The resulting dataset for (5.10) was labeled *data3c10kn*. For the 10-covariate sample, the dataset *data10c10k* was created from the following linear response function:

$$y_u^1 \equiv y_u^0 \equiv 10 + 7x_{u1} + 6x_{u2} + 5x_{u3} - 3x_{u4} + 3x_{u5} + 2x_{u6} + x_{u7} - x_{u8} + 0.5x_{u9} + 0.1x_{u10} + \varepsilon_u^0 \tag{5.11}$$

Theorem 3.3 implies that balance on the covariate means is sufficient to remove all bias in the estimate of the treatment effect for (5.9) and (5.11), while Theorem 3.7 implies that balance on the marginal distributions of the covariates is required for (5.10).

Results for BOSS with Coarsened Distribution Imbalance

Several experiments were conducted on the datasets *data3c10k* and *data10c10k* using BOSS with $\mathcal{I}_{\chi^2}$. Several different sets of bins were considered, with $n_i = 4$, 8, 16, and 32 for each $i \in \mathcal{P}$. This sequence was chosen because it forms a bin scheme where each successive set of bins simply subdivides the previous set of bins in half, creating a telescopic increase in the number of bins. As demonstrated by Theorem 3.15, this cuts the bound on the total bias in half when Assumptions 3.6 and 3.14 are valid.

For each dataset and bin scheme, 25 runs of the simulated annealing algorithm were performed, with a different random seed used for each run, to generate a set of control groups for analysis. Throughout a run, every 50th identified control group or control group with $\mathcal{I}_{\chi^2}(T, C') = 0$ was processed and stored, along with Kolmogorov-Smirnov (KS) two-sample goodness-of-fit test statistics for the treatment and control covariate distributions. For datasets with multiple covariates, the KS test statistic values were averaged over all the covariates. Upon completion of the experiments, any duplicated control groups were removed.

Because the simulated annealing algorithm uses 1-exchanges, each successive reported control group has a high degree of overlap with its predecessor (at most 50 out of a total of 500 units could have been changed). To reduce overlap among the solutions with $\mathcal{I}_{\chi^2}(T, C') = 0$, random restarts were performed after each such control group was identified.

Table 5.1 summarizes the features of optimal solutions obtained from the *data3c10k* dataset. The *Bins* column specifies the number of bins used (per covariate), and the *Observations* column reports the number of identified control groups satisfying $\mathcal{I}_{\chi^2}(T, C') = 0$. The remaining columns list the mean and standard deviation of the treatment effect estimates and the KS two-sample test statistics (averaged over the covariates). No results are presented for *data10c10k* because no solutions with $\mathcal{I}_{\chi^2}(T, C') = 0$ were found if more than four bins per covariate were used.

Table 5.1: Optimal solutions for *data3c10k* with respect to $\mathcal{I}_{\chi^2}$.

		$\widetilde{\tau}_T^1$		KS	
Bins	Observations	μ	σ	μ	σ
4	25,214	2.2904	0.2684	0.1155	0.0090
8	17,404	1.1434	0.1605	0.0825	0.0072
16	7,689	0.2380	0.1098	0.0369	0.0038
32	833	0.0122	0.0900	0.0274	0.0027
64	0	N/A	N/A	N/A	N/A

Table 5.1 shows that as the number of bins for each covariate increases, the estimate of the treatment effect tends toward the true value of zero. The KS test statistic values also indicate an increasingly higher level of balance in the covariate distributions of the treatment and control groups.

Table 5.2 shows the difference in covariate means for the treatment group and control pool, as well as the difference in covariate means for the treatment group and an optimized control group obtained by solving BOSS with $\mathcal{I}_{\chi^2}$ and $n_i = 32$ for all $i \in \mathcal{P}$. The initial mean imbalance in the treatment group and control pool is largely removed through optimization.

Table 5.2: Difference of covariate means before and after optimization with $\mathcal{I}_{\chi^2}$ and $n_i = 32$ for all $i \in \mathcal{P}$.

		Difference of Means	
Set	Covariate	Before Optimization	After Optimization
data3c10k	1	0.869	0.009
	2	0.862	0.001
	3	0.160	0.007
data10c10k	1	0.539	0.007
	2	0.553	0.014
	3	0.420	0.001
	4	−0.355	0.002
	5	0.446	0.028
	6	0.346	0.007
	7	0.407	0.010
	8	−0.180	0.005
	9	0.208	0.002
	10	0.152	0.009

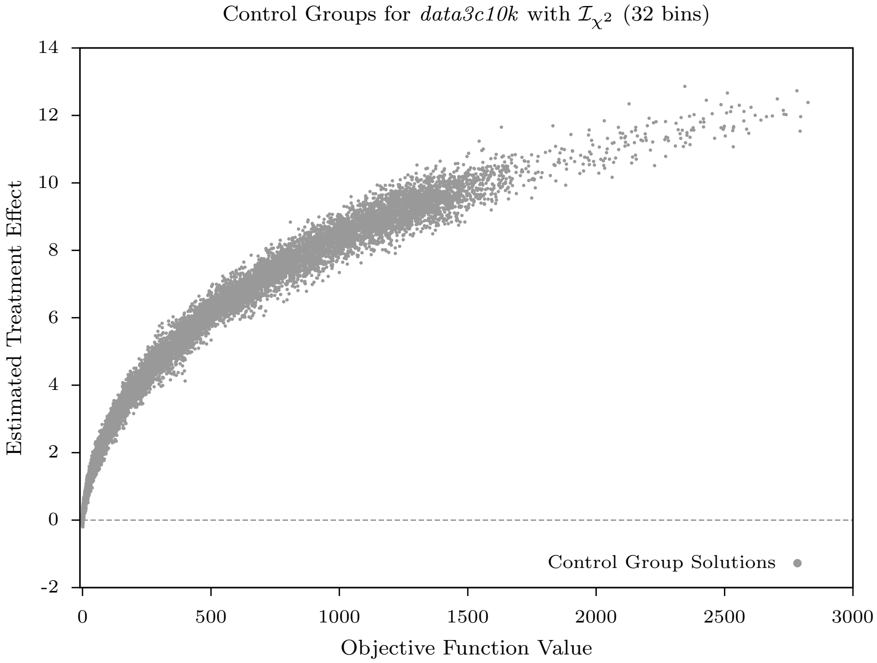


Figure 5.2: Scatter plot of the estimated treatment effect against imbalance (measured by $\mathcal{I}_{\chi^2}$ with 32 bins) for control groups from *data3c10k*.

The solutions with residual imbalance for both *data3c10k* and *data10c10k* were also analyzed. For each dataset, the identified control groups were sorted by their imbalance measures with respect to $\mathcal{I}_{\chi^2}$. The range of imbalances was then subdivided into windows of fixed width. For each window, all control groups with an imbalance within the window were grouped together and their estimated treatment effects and other relevant statistic values were averaged. Tables A.1 and A.2 display these average values obtained with $n_i = 32$ for all $i \in \mathcal{P}$. Figures 5.2 and 5.3 provide scatter plots of the estimated treatment effect against imbalance for all control groups identified during a single run of the simulated annealing algorithm on *data3c10k* and *data10c10k*, respectively. In general, as $\mathcal{I}_{\chi^2}(T, C')$ approaches zero, the estimated treatment effect tends toward zero, the true value for τ_T^1. Despite the inability to obtain solutions without residual imbalance for *data10c10k*, accurate estimates are still obtained when $\mathcal{I}_{\chi^2}$ is close to zero.

Results for BOSS with Mean Imbalance

Given the difficulty of obtaining control groups satisfying $\mathcal{I}_{\chi^2}(T, C') = 0$ for the *data10c10k* dataset, the simulated annealing algorithm was used with $\mathcal{I}_{\text{DOM}}$ in order to analyze how relaxing the balance requirements affects the solution quality. Table A.3 shows aggregate estimates of the treatment effect and other solution information from the identified control groups, split into several different imbalance measure ranges. As

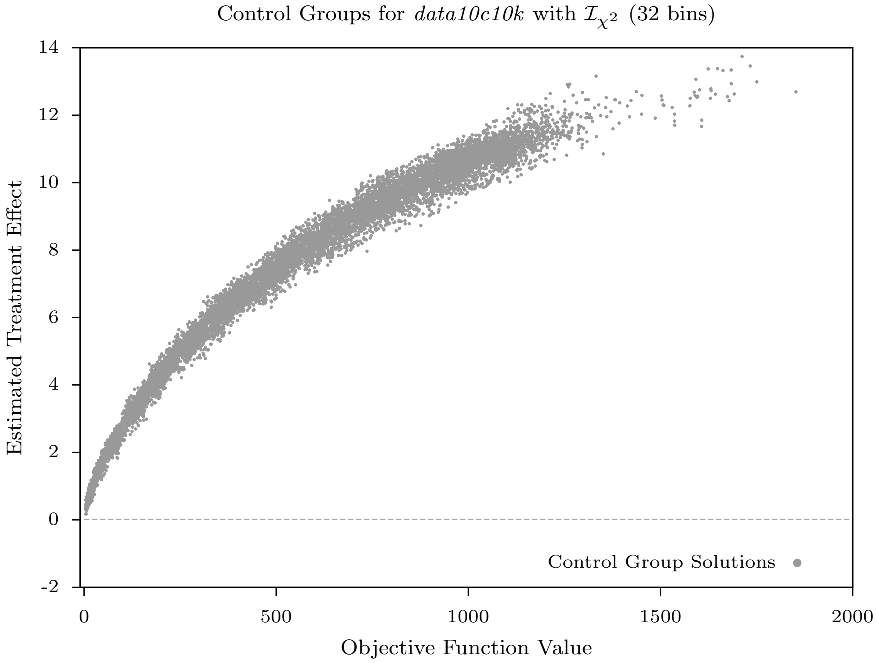


Figure 5.3: Scatter plot of the estimated treatment effect against imbalance (measured by $\mathcal{I}_{\chi^2}$ with 32 bins) for control groups from *data10c10k*.

$\mathcal{I}_{\mathrm{DOM}}$ approaches zero, the estimated treatment effect tends toward the true treatment effect of zero, which is as expected from Theorem 3.3 and the linear nature of the response function (5.11).

The results from Tables A.2 and A.3 indicate that $\mathcal{I}_{\mathrm{DOM}}$ is more effective than $\mathcal{I}_{\chi^2}$ in providing an accurate estimate of τ_T^1 for this particular dataset. However, the KS scores are worse with $\mathcal{I}_{\mathrm{DOM}}$, indicating that the covariate distributions are not as balanced compared to the control groups identified by $\mathcal{I}_{\chi^2}$. If the control response function is nonlinear, the distributional imbalance may result in residual bias in the estimate of the treatment effect.

To demonstrate this point, five runs of the simulated annealing algorithm were performed with the *data3c10kn* dataset, using both $\mathcal{I}_{\chi^2}$ with $n_i = 32$ for all $i \in \mathcal{P}$ and $\mathcal{I}_{\mathrm{DOM}}$. The best solutions obtained from these runs are reported in the first two rows of Table 5.3. In this case, the best solutions obtained with $\mathcal{I}_{\chi^2}$ lead to better estimates of τ_T^1 than those obtained with $\mathcal{I}_{\mathrm{DOM}}$. In this case, the nature of the response function (5.10) ensures that Assumption 3.2 is invalid while 3.6 is valid. As such, Theorem 3.7 applies but Theorem 3.3 does not.

The imbalance measure $\mathcal{I}_{\mathrm{DOM}}$ can be extended by incorporating higher moments, either raw or centered,

Table 5.3: Best solutions for *data3c10kn* for various imbalance measures.

	Permitted		$\widetilde{\tau}_T^1$		KS	
Objective	Imbalance	Observations	μ	σ	μ	σ
$\mathcal{I}_{\chi^2}$ $(n_i = 32)$	1.0×10^{-7}	156	-0.0170	0.0875	0.0804	0.0078
$\mathcal{I}_{\text{DOM}}$	1.0×10^{-3}	7,086	-1.3889	0.3395	0.2770	0.0226
$\mathcal{I}_{\text{DOMV}}$	1.0×10^{-3}	357	0.0392	0.0959	0.1669	0.0179
$\mathcal{I}_{\text{DOM2V}}$	1.0×10^{-3}	403	0.0986	0.1057	0.1435	0.0121

of the distributions. For these experiments, two extensions were considered. Let

$$\mu_i(S) \equiv \frac{1}{|S|} \sum_{u \in S} x_{ui}$$

and

$$s_i^2(S) \equiv \frac{1}{|S|-1} \sum_{u \in S} \left(x_{ui} - \mu_i(S)\right)^2$$

be the sample mean and sample variance for covariate $i \in \mathcal{P}$ across the units in $S \subseteq T \cup C$. Two additional imbalance measures are

$$\mathcal{I}_{\text{DOMV}}(T, C') \equiv \mathcal{I}_{\text{DOM}}(T, C') + \sum_{i \in \mathcal{P}} \left| s_i^2(T) - s_i^2(C') \right|$$

and

$$\mathcal{I}_{\text{DOM2V}}(T, C') \equiv \sum_{i \in \mathcal{P}} \left(\mu_i(T) - \mu_i(C')\right)^2 + \sum_{i \in \mathcal{P}} \left| s_i^2(T) - s_i^2(C') \right| .$$

These two imbalance measures can be used to find control groups with the first and second moments of the covariate distribution as close as possible to those of the treatment group. These two measures differ in the weight that they place on the means. For *data3c10kn*, the results from optimizing these two imbalance measures with the simulated annealing algorithm are shown in the third and fourth rows of Table 5.3. The mean estimates from solutions identified by $\mathcal{I}_{\text{DOMV}}$ and $\mathcal{I}_{\text{DOM2V}}$ improve over the mean estimate from $\mathcal{I}_{\text{DOM}}$, but they are worse than the mean estimate from $\mathcal{I}_{\chi^2}$. In addition, both $\mathcal{I}_{\text{DOMV}}$ and $\mathcal{I}_{\text{DOM2V}}$ are less successful than $\mathcal{I}_{\chi^2}$ at balancing the covariate distributions.

Comparison with Matching Methods

To demonstrate the performance of BOSS with respect to existing matching methods, the *Matching* package (Sekhon, 2011) was used. The package allows for matching based on propensity score, matching directly on

Table 5.4: Comparison of single best solutions for BOSS and matching for *data3c10kn*.

	Objective		KS	
Objective	Value	$\widetilde{\tau}_T^1$	μ	max
$\mathcal{I}_{\chi^2}$ $(n_i = 32)$	0.0	-0.1142	0.025	0.026
$\mathcal{I}_{\text{DOM}}$	1.50×10^{-5}	-0.9877	0.093	0.118
$\mathcal{I}_{\text{DOMV}}$	3.77×10^{-4}	0.0271	0.062	0.088
$\mathcal{I}_{\text{DOM2V}}$	2.69×10^{-4}	0.1154	0.045	0.060
Prop. Score	N/A	-1.3434	0.125	0.158
Cov. Matching	N/A	0.0943	0.025	0.034

the values of the covariates, or some combination of the two. For the purposes of testing, a standard logistic regression model was used to estimate the propensity score as a linear function of the covariates.

Table 5.4 compares the best solutions (as defined by the imbalance measure, with ties broken arbitrarily) obtained by BOSS with $\mathcal{I}_{\chi^2}$ using $n_i = 32$ for all $i \in \mathcal{P}$, $\mathcal{I}_{\text{DOM}}$, $\mathcal{I}_{\text{DOMV}}$, and $\mathcal{I}_{\text{DOM2V}}$ to the solutions returned by both propensity score matching and matching on the covariates for the *data3c10kn* dataset. The *Objective* column lists the method used to obtain the solution, the *Objective Value* column lists the residual imbalance of the best solution for BOSS (no objective score is available for matching), the $\widetilde{\tau}_T^1$ column lists the estimate of the treatment effect computed from the best solution, and the *KS* columns list the average and maximum values of the KS test statistic for the marginal covariate distributions in the treatment group and the best control group.

For these results, the propensity score model fares the worst in producing accurate estimates of the treatment effect, while covariate matching and BOSS with $\mathcal{I}_{\chi^2}$, $\mathcal{I}_{\text{DOMV}}$, and $\mathcal{I}_{\text{DOM2V}}$ produce reasonable results. The poor performance of the propensity score approach might be due to the use of a linear model to estimate it, while the actual response function is nonlinear. A better model for estimating the propensity score would potentially improve these results. Additionally, the propensity score approach produces the worst balance as measured by the KS statistic, followed by BOSS with $\mathcal{I}_{\text{DOM}}$. BOSS with $\mathcal{I}_{\chi^2}$ and covariate matching perform the best in this regard.

One difficulty for matching on the covariates is that close matches become less likely as the number of covariates increases. To demonstrate this problem, the matching procedures were run on the *data10c10k* dataset. Table 5.5 shows the best solutions obtained by the various BOSS and matching approaches. Since *data10c10k* uses a linear response function (5.11), both propensity score matching and BOSS with $\mathcal{I}_{\text{DOM}}$ perform better than they did previously. This improvement occurs because balancing covariate means for a linear response function produces good estimates, as Theorem 3.3 indicates. Estimating the propensity

Table 5.5: Comparison of single best solutions for BOSS and matching for *data10c10k*.

	Objective		KS	
Objective	Value	$\widetilde{\tau}_T^1$	μ	max
$\mathcal{I}_{\chi^2}$ $(n_i = 32)$	2.9502	0.2168	0.026	0.036
$\mathcal{I}_{\text{DOM}}$	0.0029	0.1294	0.039	0.056
$\mathcal{I}_{\text{DOMV}}$	0.0157	0.1857	0.037	0.048
$\mathcal{I}_{\text{DOM2V}}$	0.0158	0.1947	0.045	0.052
Prop. Score	N/A	-0.1148	0.066	0.114
Cov. Matching	N/A	2.818	0.067	0.088

score with a linear model tends to balance the means indirectly, while optimizing $\mathcal{I}_{\text{DOM}}$ balances the means directly. On the other hand, the effectiveness of covariate matching is reduced due to the difficulty of finding close matches on ten different covariates. Finally, BOSS with $\mathcal{I}_{\chi^2}$ is seen to produce the best covariate balance as measured by the KS test statistic, while the matching approaches produce the worst covariate balance.

5.2.2 Experiments with Exact Algorithms

A second set of computational tests used the mixed integer programming (MIP) models in Section 5.1.2 to explore how the number of covariates affects the ability to balance their marginal distributions. These tests compared BOSS with $\mathcal{I}_{\text{Diff}}$ to several matching methods. By using exact algorithms instead of heuristics, it is possible to verify if solutions with residual imbalance are optimal. This is a key benefit of the BOSS framework as compared to matching methods, which generally do not have any way to assess if the balance for an identified matched-pair sample can be improved.

Data Generation and Setup

Three datasets with 25 covariates each were created for these tests. Covariate values for control units were normally distributed with mean 0 and standard deviation 3 for all covariates. For the treatment units, each covariate $i \in \mathcal{P}$ had mean $\mu_i \sim U(-2, 2)$, and was normally distributed about its mean with a standard deviation of 2. This construction ensured that sufficient overlap existed on the covariate values between the treatment group and control pool. If there is insufficient overlap in an observational study, then other methods of adjustment such as regression may be used, though such methods tend to be model-dependent. Each dataset contained 100 treatment units, and 1000, 5000, and 10,000 control units; the datasets were designated as *data25c1k*, *data25c5k*, and *data25c10k*, respectively. No response functions were used to initialize the treatment and control responses of the units in these datasets. This allows for different response functions to be applied after optimization in order to assess how the estimate of the treatment effect from an optimal solution changes under different types of functions.

Model (5.4) with the modifications in (5.8) was used to formulate BOSS with $\mathcal{I}_{\text{Diff}}$. For all tests, the bins were constructed using the the boundaries in (5.7) with $n_i = 20$ for all $i \in \mathcal{P}$. For matching, both the Mahalanobis distance and the propensity score distance were used as distance metrics. The Mahalanobis distance was computed using the *StatMatch* package in R (D'Orazio, 2012), and the propensity scores were estimated as a linear function of the covariates using a standard logistic regression model. For each distance metric, the optimal matched pairs were identified by formulating and solving a matching model in CPLEX.

The MIP models were solved using CPLEX with a time limit of 60 seconds. All matching problems solved within a few seconds, so the time limits had no impact on solution quality. For BOSS with $\mathcal{I}_{\text{Diff}}$, CPLEX typically found good solutions within the first 60 seconds and made only minor improvements after that. The greedy approximation algorithm for BOSS with $\mathcal{I}^{+}_{\text{Cvg:}\mathbf{D}}$ presented in Section 4.2 was implemented (with a few heuristic improvements) and used to generate initial solutions which were passed to CPLEX prior to optimization. In almost all cases, CPLEX was able to quickly improve upon these solutions.

Comparisons on Covariate Balance

BOSS with $\mathcal{I}_{\text{Diff}}$ and matching methods with the Mahalanobis metric and propensity scores were run on each of the datasets while varying the number of covariates from one to 25. A comparison of the covariate balance levels (assessed through Kolmogorov-Smirnov two-sample test statistics on the marginal distributions) for solutions identified by BOSS and matching on the datasets *data25c1k*, *data25c5k*, and *data25c10k* is presented in Tables A.4, A.5, and A.6, respectively. Figures 5.4 and 5.5 provide graphical displays of these results, with all possible numbers of covariates (from one up to 25). In each of the figures, the top two charts are for *data25c1k*, the middle two are for *data25c5k*, and the bottom two are for *data25c10k*. In Figure 5.4, the charts on the left side show the average values of the Kolmogorov-Smirnov test statistic across the covariates for the best matching and BOSS solutions. The charts on the right show the maximum values. In Figure 5.5, the charts on the left side show the average p-values from the KS tests while the charts on the right side show the minimum p-values. The *MM* lines correspond to matching with the Mahalanobis metric and the *PS* lines correspond to matching with propensity scores.

The trends in Figures 5.4 and 5.5 show that both Mahalanobis matching and BOSS yielded solutions with good covariate balance for a small number of covariates, but as more covariates were included, BOSS

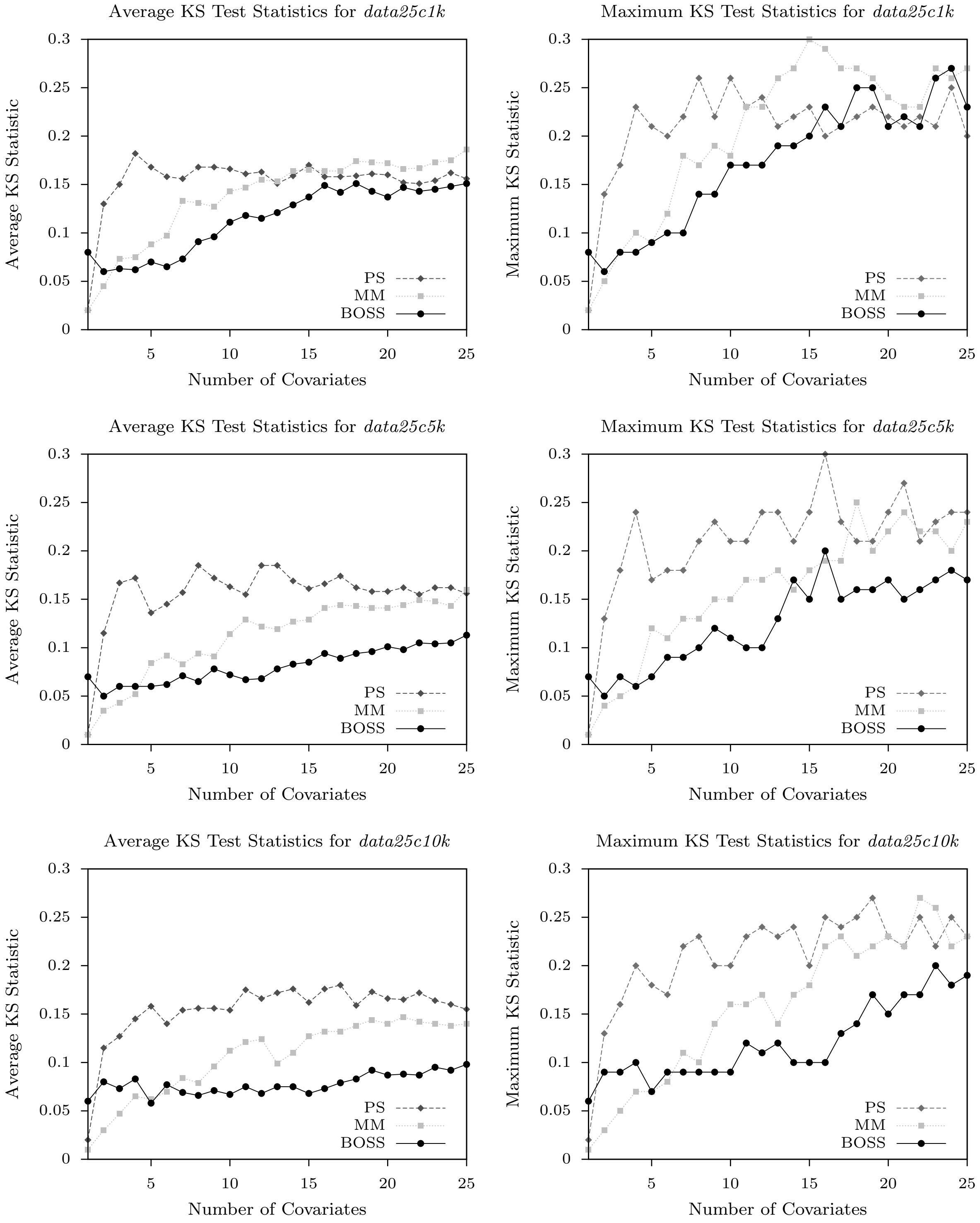
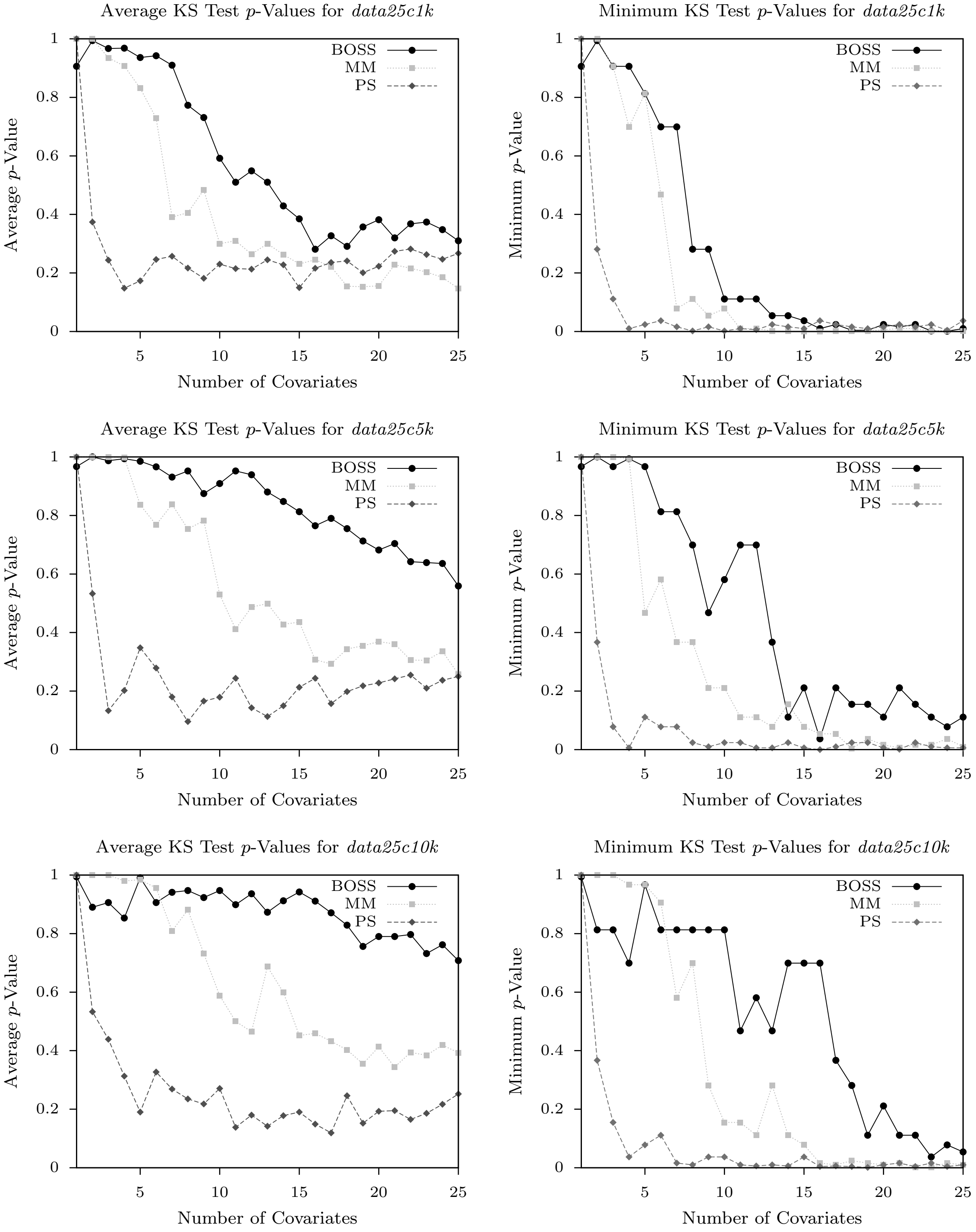


Figure 5.4: Trends for balance quality of matching and BOSS solutions. Trend lines indicate the average and maximum values of the Kolmogorov-Smirnov test statistic.


Average KS Test *p*-Values for *data25c1k*

Minimum KS Test *p*-Values for *data25c1k*

Average KS Test *p*-Values for *data25c5k*

Minimum KS Test *p*-Values for *data25c5k*

Average KS Test *p*-Values for *data25c10k*

Minimum KS Test *p*-Values for *data25c10k*

Figure 5.5: Trends for balance quality of matching and BOSS solutions. Trend lines indicate the average and minimum *p*-values from the Kolmogorov-Smirnov test across all covariates.

showed a consistent improvement over the other approaches with respect to the KS test. Propensity score matching had worse performance as measured by the KS test for more than one covariate, but the balance degraded at a slower rate as more covariates were included. In all cases, BOSS outperformed both matching methods if there were five or more covariates. Additionally, for sets *data25c5k* and *data25c10k*, minimum p-values were above 0.05 for all but two BOSS solutions across all numbers of covariates. In contrast, the minimum p-values dropped below 0.05 after 18 and 16 covariates for Mahalanobis matching and after 8 and 7 covariates for propensity score matching on the respective datasets. In general, all methods performed better with the larger control pools, as more potential controls increased the amount of overlap between the covariate distributions.

For the maximum KS values and minimum p-values in Figures 5.4 and 5.5, the trend lines are more erratic. One reason for this is that BOSS with $\mathcal{I}_{\text{Diff}}$ attempts to minimize the average covariate imbalance, which may increase imbalance on some covariates in order to decrease imbalance on others. However, BOSS can also minimize the maximum imbalance across all covariates, which should smooth out these trend lines.

BOSS with $\mathcal{I}_{\text{KS}}$ should provide KS results that are at least as good as those for BOSS with $\mathcal{I}_{\text{Diff}}$ because it is directly optimizing the measure of interest. However, in practice Model (5.3) solves significantly slower than Model (5.4), even with the improvements to reduce memory usage (for the larger datasets, such improvements were required in order to fit the model in memory). Because of this issue, BOSS with $\mathcal{I}_{\text{Diff}}$ was used for these tests.

Comparisons with Response Functions

As a further comparison between matching and BOSS with $\mathcal{I}_{\text{Diff}}$, three separate scenarios were considered for the form of the control response function. In all three scenarios, the treatment and control responses for each unit were identical so that there was no treatment effect. The responses in Scenario A were computed as a linear function of the covariates, the responses in Scenario B were computed as a separable but nonlinear function of the covariates, and the responses in Scenario C were computed as a nonlinear function of the covariates with several covariate interaction terms. The full response functions for all 25 covariates are presented in (5.12), (5.13), and (5.14), respectively, with $\mathcal{E}_u^0$ drawn from a normal distribution with mean zero and standard deviation 1 for all units. When fewer than 25 covariates were used during optimization, the remaining terms in the response function were dropped (e.g., for *data25c1k* with 10 covariates, all response function terms for x_{ui} for $i > 10$ were omitted from the computation of the responses).

$$\begin{aligned} y_u^1 \equiv y_u^0 \equiv 1.4x_{u1} &+ 1.3x_{u2} + 0.9x_{u3} - 0.9x_{u4} + 1.4x_{u5} - 1.2x_{u6} + 0.4x_{u7} \\ &- 1.3x_{u8} + 0.6x_{u9} + 1.2x_{u10} + 0.5x_{u11} + 0.4x_{u12} + 1.2x_{u13} - 0.9x_{u14} \\ &+ 1.5x_{u15} + 0.8x_{u16} - 0.9x_{u17} + 1.5x_{u18} - 1.2x_{u19} + 0.7x_{u20} - 0.5x_{u21} \\ &- 1.3x_{u22} + 1.1x_{u23} - 1.2x_{u24} + 0.5x_{u25} + \varepsilon_u^0 \end{aligned} \tag{5.12}$$

$$\begin{aligned} y_u^1 \equiv y_u^0 \equiv\ & 0.8x_{u1}(1.0 - x_{u1}) + 0.5x_{u2}(0.7 + x_{u2}) + 0.9x_{u3} - 0.9x_{u4} \\ &+ 0.7x_{u5}^2(0.5 + x_{u5}) - 0.6x_{u6}^2 + 0.4x_{u7} - 0.8x_{u8} \\ &+ 0.6x_{u9}(0.9 - x_{u9}) + 0.2x_{u10}^2(0.3 - x_{u10}) + 0.5x_{u11}^2 - 1.4x_{u12} \\ &- 0.8x_{u13} - 0.9x_{u14}^2 + 0.5x_{u15}^2(0.1 + x_{u15}) + 0.8x_{u16} \\ &- 0.9x_{u17}(0.2 - x_{u17}) + 1.5x_{u18} - 1.2x_{u19}(1.0 + x_{u19}) + 0.7x_{u20}^2(0.8 - x_{u20}) \\ &- 0.5x_{u21} - 1.3x_{u22}(1.0 + x_{u22}) + 1.1x_{u23} - 1.2x_{u24}(1.0 + x_{u24}) \\ &+ 0.4x_{u25}^2(0.6 - x_{u25}) + \varepsilon_u^0 \end{aligned} \tag{5.13}$$

$$\begin{aligned} y_u^1 \equiv y_u^0 \equiv\ & 0.8x_{u1}(1.0 - x_{u1}) + 0.5x_{u2}(0.7 + x_{u1}) + 0.27x_{u3}x_{u2} - 0.9x_{u4}^2 \\ &+ 0.7x_{u5}(0.5 + x_{u5})x_{u2} - 0.6x_{u6}x_{u1} + 0.4x_{u7} - 0.8x_{u8} \\ &+ 0.6x_{u9}(0.9 - x_{u9}) + 0.2x_{u10}^2(0.3 - x_{u7}) + 0.5x_{u11}^2 - 1.4x_{u12} \\ &- 0.8x_{u13} - 0.9x_{u14}^2 + 0.5x_{u15}^2(0.1 + x_{u15}) + 0.8x_{u16} \\ &- 0.9x_{u17}(0.2 - x_{u13}) + 1.5x_{u18} - 1.2x_{u19}(1.0 + x_{u11}) + 0.7x_{u20}^2(0.8 - x_{u20}) \\ &- 0.5x_{u21} - 1.3x_{u22}(1.0 + x_{u22}) + 1.1x_{u23} - 1.2x_{u24}(1.0 + x_{u23}) \\ &+ 0.4x_{u25}^2(0.6 - x_{u25}) + \varepsilon_u^0 \end{aligned} \tag{5.14}$$

For each of the three datasets and each possible number of covariates, the three different response functions were applied to the corresponding matching and BOSS solutions and an estimate of the treatment effect was computed as the difference in average responses for the treatment and control groups. Tables A.7, A.8, and A.9 present these results for datasets *data25c1k*, *data25c5k*, and *data25c10k*. Figure 5.6 provides a graphical representation of these results. The top three charts are for *data25c1k*, the middle three are for *data25c5k*, and the bottom three are for *data25c10k*. The charts in the first column are for Scenario *A*, the charts in the second column are for Scenario *B*, and the charts in the third column for Scenario *C*. The charts show the estimated treatment effects obtained from the matching and BOSS solutions as the number of covariates increases. The solid horizontal line represents the actual treatment effect of zero, and the gray trend line labeled *Init* represents the initial (biased) estimate derived from the treatment group and control pool.

With the linear response function model of Scenario A, propensity score matching performed the best, while BOSS and Mahalanobis matching performed comparably for *data25c1k* and *data25c5k* and BOSS beat Mahalanobis matching on *data25c10k*. This is because propensity score matching typically achieves a high level of mean balance on the covariates, and this is sufficient if the response function is linear. As earlier results demonstrated, BOSS with $\mathcal{I}_{\text{DOM}}$ would most likely perform at least as well as propensity score matching.

For Scenario B, the performance of propensity score matching quickly degraded due to the nonlinearity of the response function. In some cases, the estimates were close to the true value, but in other cases, there was a large gap. No clear pattern was evident. In contrast, BOSS and Mahalanobis matching both exhibited more predictable performance where the estimate quality degraded as more covariates were added. The two methods were comparable for *data25c1k* and *data25c5k*, and BOSS slightly outperformed Mahalanobis matching on *data25c10k*.

In Scenario C, all methods had difficulty, though BOSS and Mahalanobis matching provided reasonable estimates if there were fewer than fifteen covariates. Even though BOSS did not directly optimize balance on any of the joint distributions, it produced reliable estimates if good balance was obtained on the marginals despite the presence of interaction terms in the response function. As marginal balance deteriorated, the estimates did as well. Under all scenarios, the performance of all methods generally improved as control pool size increased and weakened as more covariates were included.

The objective values for the BOSS solutions shown in Tables A.4, A.5, and A.6 exhibit a clear relationship with the quality of the estimates reported in Tables A.7, A.8, and A.9. Specifically, as the residual imbalance measured by $\mathcal{I}_{\text{Diff}}$ increases, the quality of the estimated treatment effect decreases. This indicates the importance of identifying BOSS solutions with imbalance that is as small as possible.

5.2.3 Experiments with Alternate Optima

A third set of computational tests explored the differences in estimates from alternate optima. A new dataset with 10 covariates, 100 treatment units, and 10,000 control units was constructed using a process similar to that of Section 5.2.2. The tests used $\mathcal{I}_{\text{Diff}}$ with 20 bins and solved the associated MIP model using CPLEX while limiting the number of covariates included in the model to 1, 3, 5, and 10. Upon verification of an optimal solution, CPLEX was used to generate a set $\mathcal{S}$ of up to 5000 alternate optima, with a total time

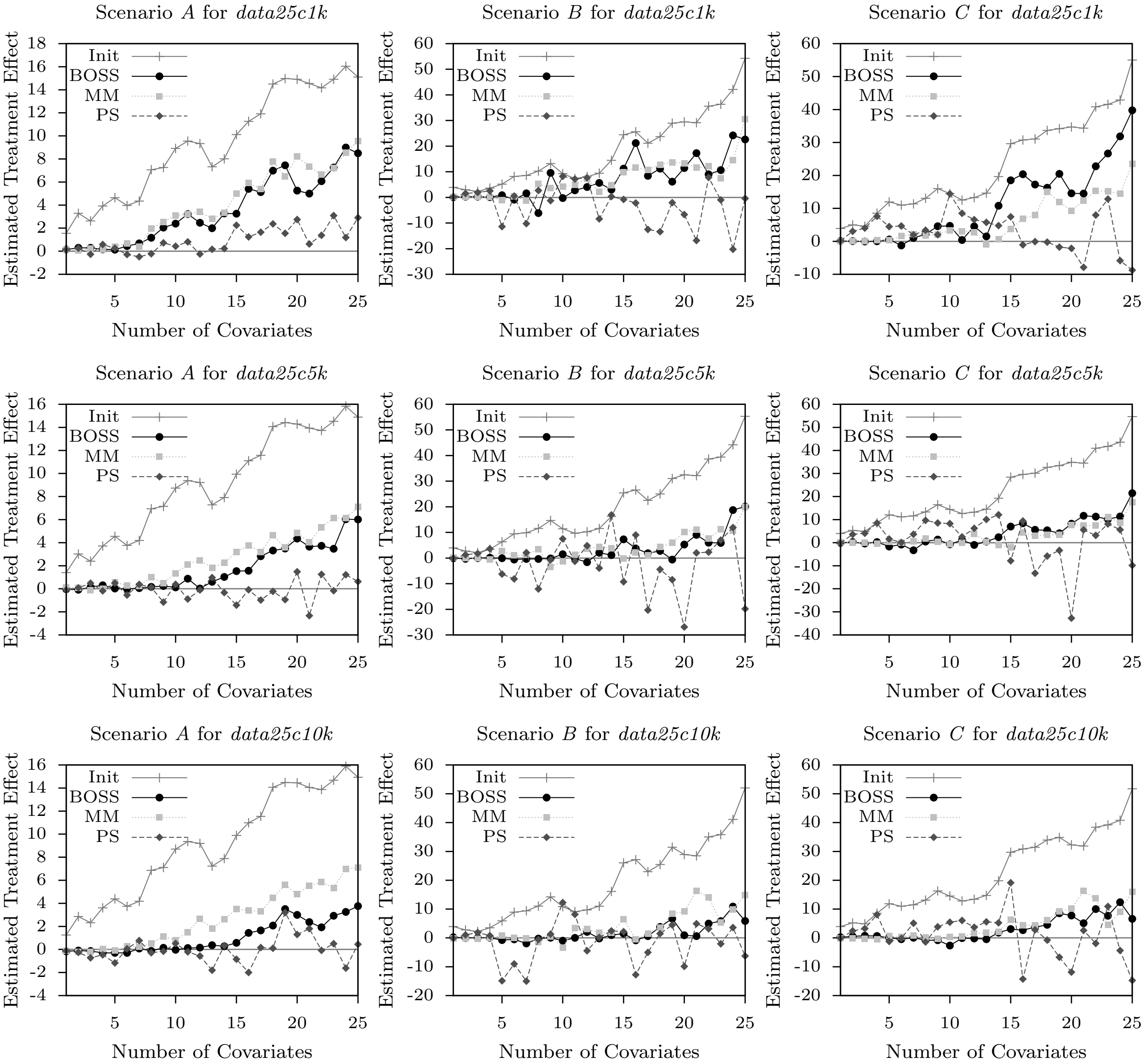


Figure 5.6: Trends for estimated treatment effects for matching and BOSS solutions.

Table 5.6: Initial estimates of the treatment effect for the different scenarios.

$\lvert\mathcal{P}\rvert$	A	B	C	D
1	1.603	4.389	4.389	-59.370
3	3.557	3.180	5.793	-67.104
5	5.777	8.702	13.302	-65.071
10	11.038	11.093	17.073	128.878

limit of 1800 seconds.

For 1, 3, and 5 total covariates, CPLEX found solutions with $\mathcal{I}_{\text{Diff}}(T, C') = 0$, thus allowing it to verify optimality. The total numbers of solutions identified by CPLEX were 5148 for 1 covariate, 5016 for 3 covariates, and 5012 for 5 covariates. For the 10 covariate case, the best solution identified by CPLEX had an (unscaled) residual imbalance of 8, and only one alternate solution with the same level of imbalance was identified. The lower bound reported by CPLEX was 0, and it failed to verify optimality within the time limit.

Results for Response Functions

Four different scenarios were considered for the control response function: Scenario A used the linear function in (5.12), Scenario B used the separable nonlinear function in (5.13), Scenario C used the nonlinear function in (5.14), and Scenario D used the highly nonlinear function in (5.15). The error terms for the units were distributed normally with mean zero and standard deviation 1. The treatment and control responses were identical so that there was no treatment effect.

$$\begin{aligned} y_u^1 \equiv y_u^0 \equiv\ & 0.3 \cdot \exp\Big((x_{u1})^2 / 10.0 \Big) + 0.5 \cdot \exp(x_{u1} x_{u2} / 7.0) + 0.9 \left| x_{u1} x_{u2} x_{u3} \right| - 0.9 \left(x_{u4} \right)^2 \\ & + 0.7 \left| x_{u5} \right| \cdot (0.5 + x_{u5}) \cdot \log(|x_{u2} + 0.1|) - 0.6 x_{u6} x_{u1} + 0.4 x_{u7} - 0.8 \cdot \exp(|x_{u8}|) \\ & + 0.6 x_{u9} (0.9 - x_{u9}) + 0.2 \left(x_{u10} \right)^2 (0.3 - x_{u7}) + \varepsilon_u^0 \end{aligned} \tag{5.15}$$

Table 5.6 shows the initial biased estimates of the treatment effect produced by comparing the average treatment response for the treatment units with the average control response across the entire sample C under each of the four scenarios. The $\lvert\mathcal{P}\rvert$ column lists the number of covariates used to compute the response function (e.g., for 1, the response function was computed from only the terms associated with the first covariate).

Theorems 3.7 and 3.15 indicate that control groups with no residual imbalance as measured by $\mathcal{I}_{\text{Diff}}$

Table 5.7: Average estimates of the treatment effect for alternate optima under different response function scenarios.

			A		B		C		D	
$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{S}\rvert$	$\mathcal{I}_{\text{Diff}}$ (20)	μ	σ	μ	σ	μ	σ	μ	σ
1	5148	0	−0.040	0.029	0.284	0.054	0.284	0.054	−0.093	0.050
3	5016	0	0.092	0.096	0.346	0.123	0.434	0.192	0.839	0.447
5	5012	0	0.126	0.111	0.573	0.467	1.442	0.715	0.475	0.592
10	2	8	0.283	0.071	1.394	0.221	2.815	0.472	−9.936	2.144

should produce estimates of the treatment effect with negligible bias under Scenarios A and B, but not necessarily under Scenarios C and D (the one exception is the single-covariate case, which does not have any interaction terms). Additionally, any difference in estimates between two optimal control groups should be solely due to the error terms under Scenarios A and B, whereas in Scenarios C and D differences may be a result of imbalance on joint distributions that were not taken into account.

From Lemma 3.20, if the optimal control groups were each random samples from U^0, then the distribution of average errors across the groups should be approximately normal. Because the standard deviation of the error terms is 1, the average error for a group of 100 units should be distributed normally with mean 0 and standard deviation $\sqrt{1/100} = 0.1$. Thus, under Scenarios A and B, the standard deviation in the estimates across all optimal control groups should be approximately 0.1. However, due to issues of overlap among optimal control groups and insufficient diversity among the set of optimal solutions, this may be unrealistic.

To test these theories, an estimate of the average treatment effect was computed for each identified control group under each of the four scenarios. Table 5.7 shows the mean (μ) and standard deviation (σ) of these treatment effect estimates computed across all solutions under each of the four scenarios and covariate sizes. The $|\mathcal{P}|$ column lists the number of covariates, and the $|\mathcal{S}|$ column lists the number of alternate optima that were identified.

The results in Table 5.7 largely confirm the expected behavior. In Scenarios A and B, the average estimate of the treatment effect is close to the true value of zero with the exception of the 10-covariate case for Scenario B. Additionally, the standard deviation of the treatment effect estimates is generally close to the expected value of 0.1. These observations also apply to Scenarios C and D with only one covariate. However, when multiple covariates are included, the averages of the estimates are farther from zero and the standard deviations are larger than would be expected if the differences between the estimates were due solely to the error terms. Across all scenarios, the estimates are a considerable improvement compared to the initial estimates in Table 5.6.

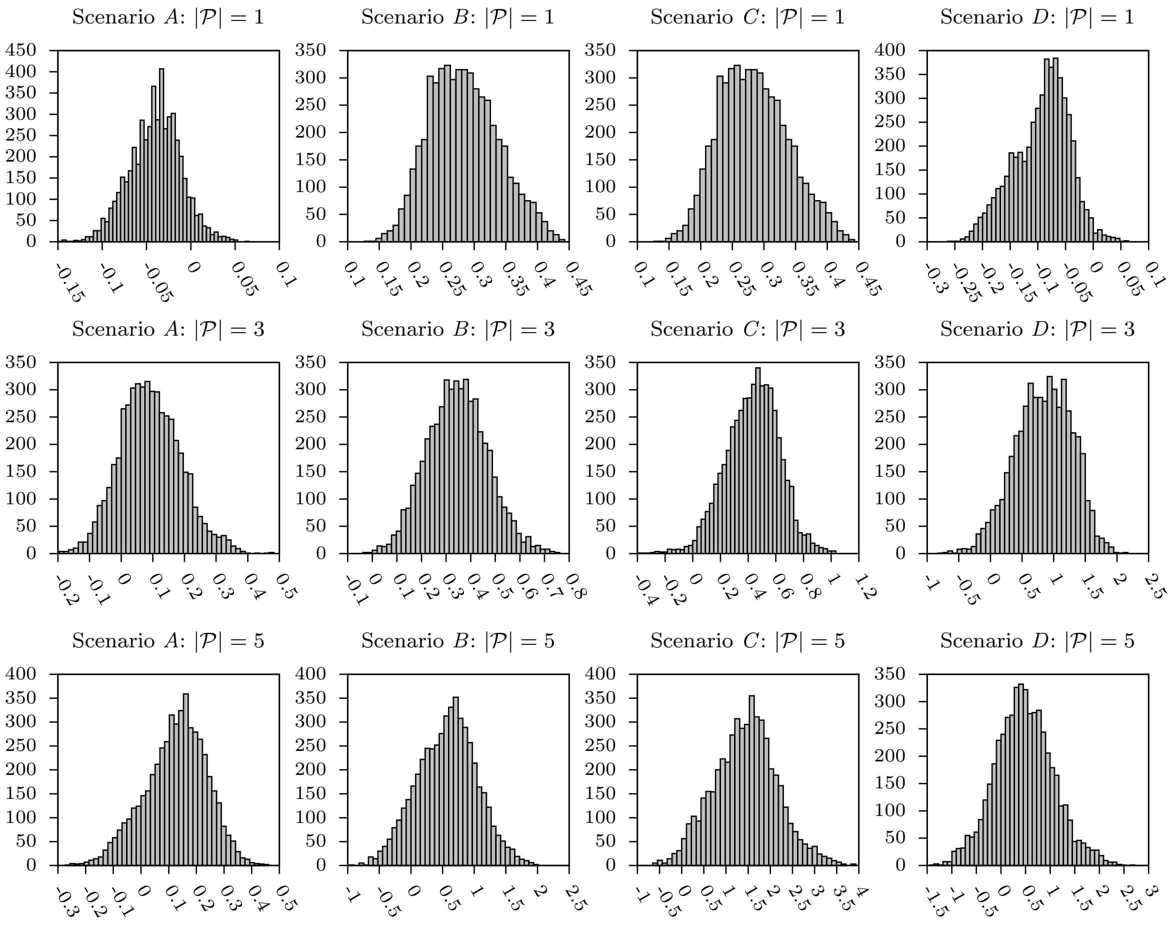


Figure 5.7: Distributions of estimated treatment effects for optimal solutions. The binned values for the treatment effect estimates are reported on the horizontal axis and the frequencies are reported on the vertical axis.

Figure 5.7 shows histograms for the distributions of the treatment effect estimates for each of the scenarios and number of covariates. The columns contain plots for Scenarios A, B, C, and D, respectively, while the rows contain the plots for the different covariate sizes (1, 3, and 5; 10 is excluded because only two solutions were found). In each case, the bin width in the plots was set using the Freedman-Diaconis rule.

Results for Unobserved Covariates

Control groups were identified previously by minimizing imbalance over a subset of the covariates, and response functions were then computed on the subset of covariates that were included in the optimization model. However, the response functions can be computed on all of the covariates, regardless of whether or not those covariates were included in the imbalance measure. This process mimics the situation in which

Table 5.8: Average estimates of the treatment effect for alternate optima in the presence of unobserved covariates.

			A		B		C		D	
$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{S}\rvert$	$\mathcal{I}_{\text{Diff}}$ (20)	μ	σ	μ	σ	μ	σ	μ	σ
1	5148	0	9.288	0.266	4.459	2.144	13.650	1.307	49.469	3.605
3	5016	0	8.004	0.655	6.625	6.926	2.202	2.318	97.571	518.255
5	5012	0	5.174	0.571	3.744	2.083	6.080	1.235	46.279	71.423
10	2	8	0.283	0.071	1.394	0.221	2.815	0.472	−9.936	2.144

unobserved covariates are present within the data.

Table 5.8 shows the mean (μ) and standard deviation (σ) for the estimates computed from the optimal control groups using all covariates under each of the four scenarios. The results for the 10-covariate case are the same as those in Table 5.7 because there are no unobserved covariates. For the cases with 1, 3, and 5 covariates, which have 9, 7, and 5 unobserved covariates, respectively, there is quite a bit more bias in the estimates. As more covariates are included and fewer are unobserved, these estimates get marginally better, but they are still far from the true value of zero. However, balancing on any number of covariates is still an improvement over the initial estimates from T and C given in the row for 10 covariates in Table 5.6. Additionally, the standard deviation of the estimates is considerably larger compared to the earlier tests without unobserved covariates.

Figure 5.8 shows histograms for the distributions of the treatment effect estimates for each of the scenarios in the presence of unobserved covariates. The columns again contain plots for Scenarios A, B, C, and D, respectively, while the rows contain the plots for the different covariate sizes (1, 3, and 5).

5.3 Real Dataset

Additional computational tests were conducted with BOSS on a well-studied dataset from the literature. The results presented here extend preliminary results presented by Cho et al. (2013) and focus on solving the optimization problems for BOSS exactly using CPLEX and the models from Section 5.1.2.

5.3.1 The LaLonde Dataset

Randomized experimental studies possess desirable statistical properties such as unbiasedness that generally lead to confidence in the causal estimates. Observational studies lack these properties, so instead researchers draw conclusions from the data through post-processing. One method of testing the effectiveness of these

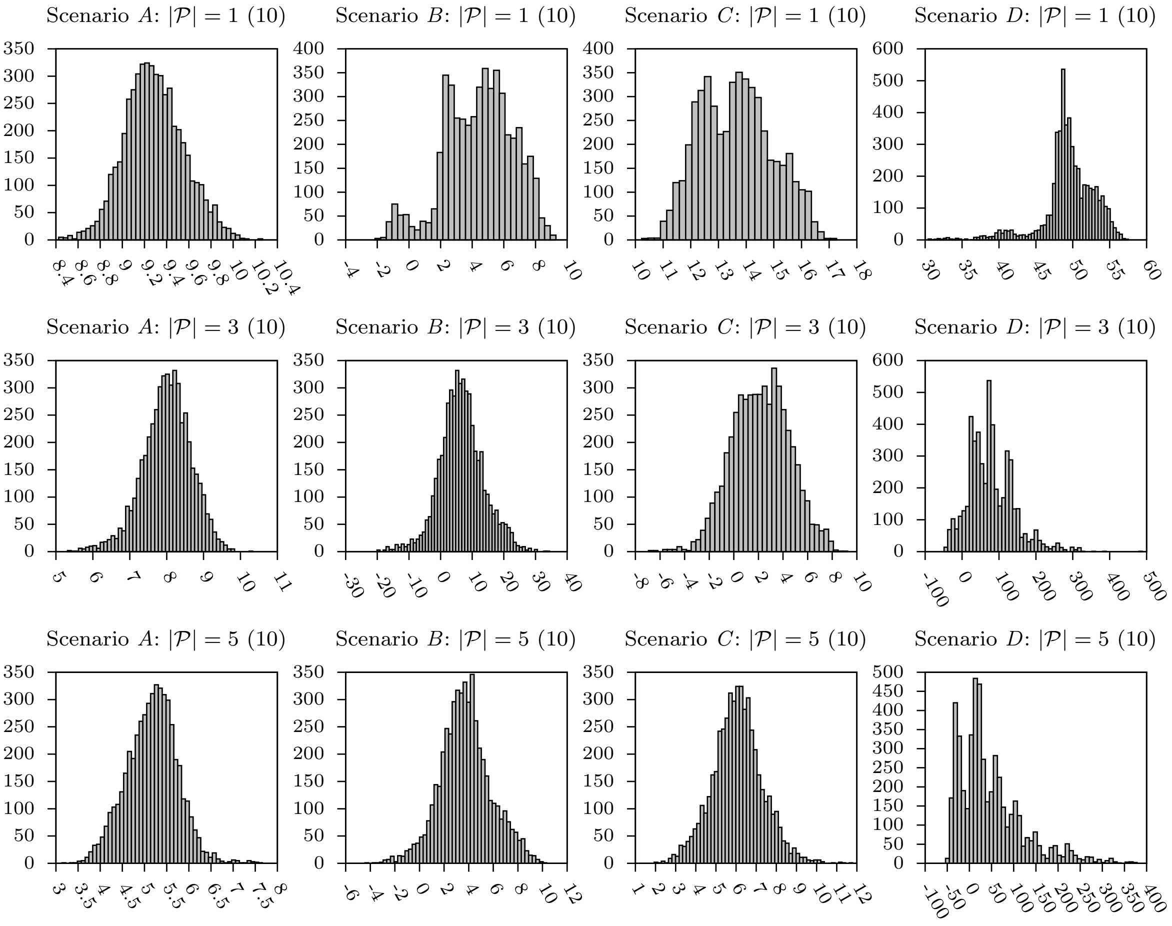


Figure 5.8: Distributions of estimated treatment effects for optimal solutions in the presence of unobserved covariates. The binned values for the treatment effect estimates are reported on the horizontal axis and the frequencies are reported on the vertical axis. The plot for Scenario D with 3 covariates omits seven outliers with treatment effect estimates in excess of 13850.

post-processing techniques is by applying them to a study in which the outcome is already known, either by design or through an equivalent experimental study, and determining whether or not the correct conclusions can be reached. This process was used by LaLonde (1986), and it has resulted in numerous subsequent papers in the causal inference literature.

LaLonde (1986) attempted to use observational data techniques to recreate the results of a randomized experimental study, the National Supported Work Demonstration (NSW). The NSW was a temporary employment program that sought to provide job skills and experience to unemployed workers in order to help them find long-term work opportunities. Qualified candidates were randomly chosen for training positions, with those unselected forming the control group. The outcome of interest was the income level for treated and control individuals several years after the training program. The experimental study found a net positive effect for training: the treated individuals had an average income larger than the control individuals.

After the experimental study, LaLonde (1986) proposed the use of the NSW study in an observational study setting. LaLonde took the treatment group from the NSW study and two large pools of control individuals from two datasets (the *Panel Study of Income Dynamics* (PSID), and the *Current Population Survey* (CPS)), and used econometric methods to derive an estimate of the treatment effect. Under a variety of models and methods, LaLonde was unable to reliably recover the experimental estimate of the treatment effect, and he ultimately concluded that the existing procedures for post-processing observational data were inadequate.

Later work by Dehejia and Wahba (1999, 2002), Dehejia (2005), and Smith and Todd (2001, 2005a,b) expanded on the original work of LaLonde (1986), with conflicting conclusions. More recent work by Diamond and Sekhon (2013) emphasized the importance of covariate balance in order to achieve reliable treatment effect estimates, both by showing that existing methods that got the wrong estimates failed to good balance and by showing that new methods that obtain good balance obtain the right estimates. The importance of balance in the LaLonde dataset makes it a good test for BOSS.

The covariates in the LaLonde dataset are: age, education (in years), income in 1974 (RE74), income in 1975 (RE75), and indicator variables for Black, Hispanic, married, and high school degree. The outcome of interest is income in 1978 (RE78). The original datasets used by LaLonde only included income in 1975, not in 1974. Dehejia and Wahba (1999) argued that it was important to account for two years of pre-treatment income, and so they restricted the original datasets to those individuals for whom income in 1974 was known. The original experimental dataset is designated *nswexp*, and the experimental dataset restricted to individuals for whom income in 1974 is known is designated as *nswre74exp*. There are two potential control pools, CPS

Table 5.9: Initial values for the LaLonde datasets.

Objective	*nswexp*	*nswcps*	*nswpsid*	*nswre74exp*	*nswre74cps*	*nswre74psid*
$\mathcal{I}_{\text{DOM}}$	3.9901×10^1	1.0597×10^4	1.6011×10^4	2.7785×10^2	2.4051×10^4	3.4877×10^4
$\mathcal{I}_{\text{KS}}$	0.2926	3.1428	3.3026	0.5727	3.7261	3.9824
$\mathcal{I}_{\text{KS:max}}$	0.0835	0.7278	0.6979	0.1265	0.7697	0.7736
$\mathcal{I}_{\text{Diff}}$ (20)	0.7962	6.3305	6.6091	1.4331	7.4799	7.9761
$\mathcal{I}_{\text{Diff}^2}$ (20)	0.0291	2.4222	2.6993	0.0783	2.9656	3.4294
$\widetilde{\tau}_T^1$ ($)	886.30	−8870.30	−15577.57	1794.34	−8497.51	−15204.78
$\lvert T\rvert$	297	297	297	185	185	185
$\lvert C\rvert$	425	15992	2490	260	15992	2490

and PSID, which leads to a total of four observational datasets: the experimental treatment individuals with the CPS and PSID groups, designated as *nswcps* and *nswpsid*, and those same sets restricted to individuals for whom income in 1974 is known, designated as *nswre74cps* and *nswre74psid*.

Table 5.9 shows the values for the various imbalance measures, the treatment effect estimates, and the number of treatment and control units for the experimental datasets (*nswexp* and *nswre74exp*) and the constructed observational datasets (*nswcps*, *nswpsid*, *nswre74cps*, and *nswre74psid*). The *Objective* column gives the imbalance measure with a parenthesized number indicating the number of bins, if appropriate (e.g., "$\mathcal{I}_{\text{Diff}}$ (20)" uses 20 uniform-width bins for each covariate, with empty bins dropped during optimization). The remaining columns give the values for the datasets, computed using the treatment group and the entire control pool.

The values in Table 5.9 indicate that the observational datasets have a much greater level of covariate imbalance than the experimental datasets. Additionally, the estimates of the treatment effect given in the row labeled $\widetilde{\tau}_T^1$ *($)* show that the observational datasets produce significantly biased estimates as compared to the experimental benchmarks, which are unbiased due to randomization. A difference is expected given the nature of observational data, and the result illustrates the need to control for biases if attempting to obtain treatment effect estimates. The goal of BOSS is to remove potential bias due to covariate imbalance between the treatment group and control pool.

5.3.2 Preliminary Results

The mathematical programming models for BOSS were solved on the LaLonde datasets using CPLEX with a time limit of 600 seconds. Results are shown in Table 5.10. The *Set* column gives the name of the dataset, the *Objective* column gives the imbalance measure, the $\widetilde{\tau}_T^1$ *($)* column gives the estimate of the treatment effect

Table 5.10: Best solutions for the LaLonde datasets with various imbalance measures.

			Objective	Lower		CPU	Improvement	
Set	Objective	$\widetilde{\tau}_T^1$ ($)	Value	Bound	Gap (%)	Time (s)	Observ.	Exper.
nswcps	$\mathcal{I}_{\text{DOM}}$	-598.50	0.0009	0.0000	100.00	1988.97	1.15×10^7	4.32×10^4
	$\mathcal{I}_{\text{SDOM}}$	-806.25	0.0000	0.0000	100.00	2073.97	N/A	N/A
	$\mathcal{I}_{\text{KS}}$	-556.95	0.0236	0.0198	15.87	361.13	1.33×10^2	1.24×10^1
	$\mathcal{I}_{\text{KS:max}}$	-562.35	0.0067	0.0064	4.52	502.76	1.08×10^2	1.24×10^1
	$\mathcal{I}_{\text{Diff}}$ (20)	-732.43	0.0404	0.0404	0.00	0.83	1.57×10^2	1.97×10^1
	$\mathcal{I}_{\text{Diff}^2}$ (20)	-575.93	0.0001	0.0001	31.92	2055.80	1.78×10^4	2.14×10^2
nswpsid	$\mathcal{I}_{\text{DOM}}$	-3250.06	3.8791	3.8787	0.01	95.51	4.13×10^3	1.03×10^1
	$\mathcal{I}_{\text{SDOM}}$	-5475.02	1.1690	1.1690	0.00	0.20	N/A	N/A
	$\mathcal{I}_{\text{KS}}$	-4616.25	0.9529	0.9529	0.00	0.52	3.47	0.31
	$\mathcal{I}_{\text{KS:max}}$	-3842.72	0.1886	0.1886	0.00	0.62	3.70	0.44
	$\mathcal{I}_{\text{Diff}}$ (20)	-5011.55	1.7778	1.7778	0.00	0.28	3.72	0.45
	$\mathcal{I}_{\text{Diff}^2}$ (20)	-3932.18	0.1800	0.1800	0.01	0.59	1.50×10^1	0.16
nswre74cps	$\mathcal{I}_{\text{DOM}}$	1538.85	0.0088	0.0000	100.00	2062.98	2.74×10^6	3.16×10^4
	$\mathcal{I}_{\text{SDOM}}$	1128.73	0.0000	0.0000	100.00	2128.25	N/A	N/A
	$\mathcal{I}_{\text{KS}}$	1710.31	0.0811	0.0811	0.00	287.32	4.60×10^1	7.06
	$\mathcal{I}_{\text{KS:max}}$	1979.44	0.0216	0.0213	1.32	405.93	3.56×10^1	5.85
	$\mathcal{I}_{\text{Diff}}$ (20)	1767.22	0.1189	0.1189	0.00	1.16	6.29×10^1	1.21×10^1
	$\mathcal{I}_{\text{Diff}^2}$ (20)	1522.39	0.0010	0.0008	21.37	2049.34	2.99×10^3	7.88×10^1
nswre74psid	$\mathcal{I}_{\text{DOM}}$	1043.61	5.4521	5.4453	0.12	1472.79	6.40×10^3	5.10×10^1
	$\mathcal{I}_{\text{SDOM}}$	-2706.21	0.8509	0.8509	0.00	0.54	N/A	N/A
	$\mathcal{I}_{\text{KS}}$	-2360.15	1.2432	1.2432	0.00	0.85	3.20	0.46
	$\mathcal{I}_{\text{KS:max}}$	-1549.89	0.2054	0.2054	0.00	1.05	3.77	0.62
	$\mathcal{I}_{\text{Diff}}$ (20)	-1486.13	1.6973	1.6973	0.00	0.29	4.70	0.84
	$\mathcal{I}_{\text{Diff}^2}$ (20)	-1228.73	0.1302	0.1302	0.01	1.10	2.63×10^1	0.60

obtained from the best solution, the *Objective Value* column gives the objective function value of the best solution found by CPLEX, the *Lower Bound* column gives the best lower bound computed by CPLEX upon termination (verification of optimality or time limit), the *Gap (%)* column gives the relative gap between the best solution and the lower bound computed by CPLEX, and the *CPU Time (s)* column gives the CPU time, in seconds, that CPLEX spent solving the problem. If the objective value equals the lower bound (or equivalently, where the gap is zero), then CPLEX was able to verify the optimal solution within the time limit. Finally, the *Improvement* columns provide the ratio of improvement in objective function value achieved by CPLEX compared to the initial observational dataset using all members of the control pool (the *Observ.* column) and the experimental dataset (the *Exper.* column). A ratio larger than one indicates that the solution identified by CPLEX has better balance (as measured by the objective function) than the corresponding dataset, while a ratio less than one indicates that CPLEX found a solution featuring worse balance.

There are several observations that can be made. For the *nswpsid* and *nswre74psid* datasets, CPLEX is able to terminate with an optimal or near-optimal solution in all cases, often in a relatively small amount of time. This is likely due to the small size of the control pool, which produces MIP models with a relatively small number of decision variables. For the *nswcps* and *nswre74cps* datasets which have a much larger control pool, CPLEX requires more time to solve the models and is only able to verify optimality in a few cases. However, with the exception of $\mathcal{I}_{\text{DOM}}$ and $\mathcal{I}_{\text{SDOM}}$, the gaps are relatively small. (For $\mathcal{I}_{\text{SDOM}}$, the difference between the objective value and lower bound is not shown in Table 5.10 due to round-off.)

The second observation is that in all cases the BOSS solutions have better covariate balance than the initial control pool, as evidenced by the fact that all values in the *Improvement: Observ.* column are greater than one (and in some cases this improvement is several orders of magnitude). This is not surprising, since the goal of BOSS is to improve covariate balance. More interesting is the fact that in some cases these solutions also have better covariate balance than the experimental datasets. This is particularly evident for the *nswcps* and *nswre74cps* datasets, where the improvement is almost always an order of magnitude or more above the experimental baseline. So despite the fact that the solutions obtained for the *nswpsid* and *nswre74psid* datasets are provably optimal in most cases, the solutions obtained for the *nswcps* and *nswre74cps* datasets feature better covariate balance.

The third observation concerns the treatment effect estimates, which is ultimately the quantity of interest. For both the *nswcps* and *nswpsid* datasets, the treatment effect estimates are all negative. This does not agree with the experimental estimate from *nswexp* which is positive ($886). This bias could be due to residual imbalance on the observed covariates or their joint distributions, imbalance on one or more unobserved covariates, or some combination of these factors.

For the *nswcps* dataset, there is little remaining imbalance on the existing covariates, which supports the existence of an important unobserved covariate that is not balanced. For the *nswpsid* dataset, a significant amount of imbalance remains on the observed covariates, but this is not due to insufficient optimization because the identified solutions are all optimal or nearly optimal with respect to the corresponding objectives. Given that the treatment effect estimates for the *nswpsid* dataset are further from the experimental benchmark than the estimates for the *nswcps* dataset, it seems likely that bias in the estimates for *nswpsid* is caused by both residual imbalance on the observed covariates and imbalance on unobserved covariates. Dehejia and Wahba (1999) reached a similar conclusion regarding the presence of an important unobserved covariate in these two datasets, which is why they created the *nswre74cps* and *nswre74psid* datasets.

The treatment effect estimates obtained from BOSS solutions for the *nswre74cps* and *nswre74psid* datasets are a substantial improvement upon the *nswcps* and *nswpsid* datasets, particularly for the *nswre74cps* dataset. All treatment effect estimates for *nswre74cps* are positive and larger than \$1100, which is more in line with the experimental estimate of \$1794. Additionally, the distribution-based imbalance measures ($\mathcal{I}_{\mathrm{KS}}$, $\mathcal{I}_{\mathrm{KS:max}}$, $\mathcal{I}_{\mathrm{Diff}}$, and $\mathcal{I}_{\mathrm{Diff}^2}$) are within about \$270 of the experimental estimate.

The estimates for *nswre74psid*, while closer to the experimental benchmark than those obtained for *nswpsid*, are still significantly negative for all but one imbalance measure. However, the solutions obtained for the distribution-based imbalance measures feature lower covariate balance than the experimental dataset, and relatively large residual imbalances compared to the corresponding *nswre74cps* solutions. Thus, it is likely that the bias in these treatment effect estimates is due to residual imbalance on the observed covariates.

One interesting observation from the *nswre74psid* results is that the difference of means imbalance measure $\mathcal{I}_{\mathrm{DOM}}$ produces a better estimate of the treatment effect than the other objectives. One can look at the imbalance levels for each individual covariate in the resulting control group solutions to understand why this is the case. Table A.11 provides a summary of the balance levels (difference of means, Kolmogorov-Smirnov test statistic and corresponding p-value) for each covariate in the solutions for both *nswre74cps* and *nswre74psid* for each objective (details for the *nswcps* and *nswpsid* solutions are included in Table A.10). The most noticeable difference between the *nswre74psid* solutions is that the solution produced by $\mathcal{I}_{\mathrm{DOM}}$ has a small difference in means for both the RE74 and RE75 covariates, while the remaining objectives produce solutions with large difference of means scores for these two covariates. Additionally, with the exception of $\mathcal{I}_{\mathrm{DOM}}$, all of the solutions for *nswre74cps* have much lower difference of means scores than the corresponding solutions for *nswre74psid*. Thus, it seems likely that the poor estimates from *nswre74psid* for all objectives except $\mathcal{I}_{\mathrm{DOM}}$ stem from the fact that the identified solutions feature poor balance on the RE74 and RE75 covariates.

A possible reason why these objectives produce poor balance on the RE74 and RE75 covariates is that the smaller size of the control pool for *nswre74psid* makes it more difficult to achieve balance on the distributions of the covariates, as evidenced by the poor imbalance scores. However, it is possible to balance the means of the covariates without balancing the distributions, which is what $\mathcal{I}_{\mathrm{DOM}}$ does. In the case of $\mathcal{I}_{\mathrm{SDOM}}$, the scaling factor reduces the penalty due to imbalance on RE74 and RE75, leading to solutions that favor balancing the means of the other covariates first. In the case of the distribution-based objectives, they attempt to balance the covariate distributions but are unable to do so effectively, leading to poor balance on the covariate means.

5.3.3 Alternate Optima

A natural question to ask is how sensitive are the treatment effect estimates produced by BOSS to changes in the control group. One way to assess sensitivity is to use BOSS to produce multiple optimally balanced control groups and look at the distribution of treatment effect estimates from each control group. A similar procedure to that in Section 5.2.3 was employed to gather multiple solutions for the datasets and imbalance measures that had solutions with better balance than the experimental control group (i.e., the imbalance measures for which *Improvement: Exper.* is greater than 1).

In an attempt to obtain a diverse set of control groups featuring minimal mutual overlap, an iterative procedure was used to seek alternate optima with minimal overlap with those that had been previously identified. Let $\mathcal{S}$ be a pool of alternate optima (i.e., a set of subsets of C). For each unit $c \in C$, define

$$\kappa(\mathcal{S}, c) \equiv |\{C' \in \mathcal{S} : c \in C'\}|$$

as the number of times unit c has appeared in a previous solution. Then the objective

$$\min \sum_{c \in C} 2\kappa(\mathcal{S}, c) v_c \tag{5.16}$$

can be used in the MIP models to identify a control group that is distinct from those that were previously identified.

The above concepts were used to identify a diverse set of alternate optima for each imbalance measure through the following procedure:

1. Solve the associated model for $\mathcal{I}$; let C^* be the best solution identified within the time limit.

2. Initialize the solution pool $\mathcal{S} \equiv \{C^*\}$.

3. Add the constraint $\mathcal{I}(T, C') \leq \mathcal{I}(T, C^*)$ and the objective (5.16) to the model.

4. Repeatedly solve the revised model, updating the solution pool after each iteration.

The time limit used in Step 1 was 300 seconds. A total time limit of 1800 seconds was used for Step 4. Each iteration of this step had a time limit of 150 seconds. If a new solution was identified by CPLEX, then it was added to the pool and the next iteration was started. If there was no progress, CPLEX was allowed to continue searching as long as the total time limit was not exceeded. Step 4 was repeated up to a total of 500 times within the time limit.

Table 5.11: Estimates from alternate solutions for the LaLonde datasets.

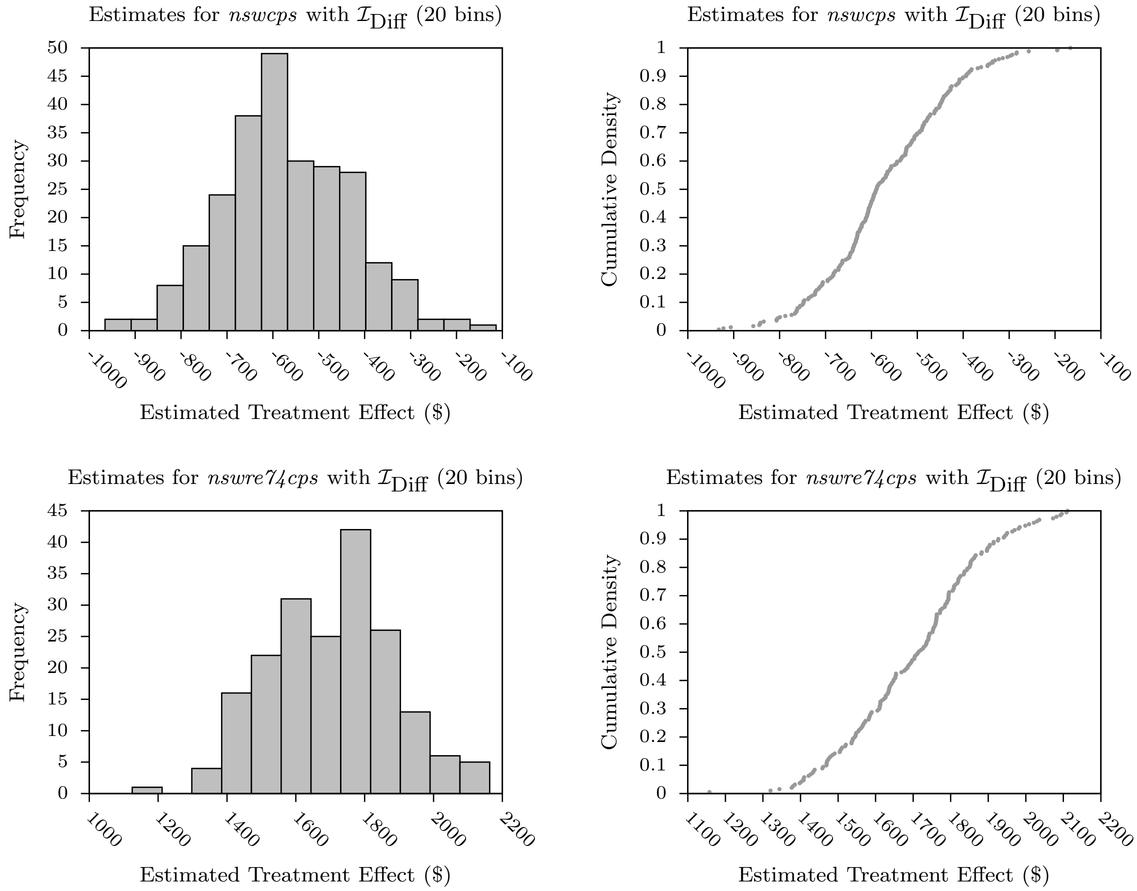
			Permitted	$\widetilde{\tau}_T^1 \ (\$)$				
Set	Objective	$\lvert\mathcal{S}\rvert$	Imbalance	μ	σ	min	max	Common
nswcps	$\mathcal{I}_{\text{DOM}}$	5	0.0009	−771.16	142.96	−906.33	−598.50	0.1046
	$\mathcal{I}_{\text{KS}}$	18	0.0236	−475.20	145.97	−803.64	−241.92	0.5455
	$\mathcal{I}_{\text{KS:max}}$	20	0.0067	−654.69	130.01	−949.30	−477.52	0.5993
	$\mathcal{I}_{\text{Diff}}$ (20)	251	0.0404	−570.52	138.01	−932.46	−166.47	0.6027
	$\mathcal{I}_{\text{Diff}^2}$ (20)	1	0.0001	−499.98	0.00	−499.98	−499.98	1.0000
nswpsid	$\mathcal{I}_{\text{DOM}}$	10	3.8791	−3320.05	51.08	−3417.33	−3250.06	0.9596
nswre74cps	$\mathcal{I}_{\text{DOM}}$	1	0.0088	1538.85	0.00	1538.85	1538.85	1.0000
	$\mathcal{I}_{\text{KS}}$	12	0.0811	1825.98	155.12	1534.53	1996.89	0.7081
	$\mathcal{I}_{\text{KS:max}}$	12	0.0216	1842.19	183.16	1508.24	2280.16	0.7081
	$\mathcal{I}_{\text{Diff}}$ (20)	191	0.1189	1702.86	179.02	1157.52	2111.34	0.7081
	$\mathcal{I}_{\text{Diff}^2}$ (20)	1	0.0010	1601.74	0.00	1601.74	1601.74	1.0000
nswre74psid	$\mathcal{I}_{\text{DOM}}$	12	5.4833	955.19	216.96	682.97	1229.21	0.8757

With the above procedure, the solutions included in $\mathcal{S}$ are not necessarily optimal because the time limit in Step 1 may stop CPLEX prematurely. However, tests indicated that CPLEX was unable to make much progress after this time limit, and so it was determined to be sufficient for the purposes of identifying a near-optimal solution.

Table 5.11 shows the alternate optima results. The *Set* column lists the dataset; the *Objective* column reports the imbalance measure; the $|\mathcal{S}|$ column reports the number of identified alternate solutions; the *Permitted Imbalance* column reports the residual imbalance that was allowed for the solutions in $\mathcal{S}$; the μ, σ, min, and max columns report the mean, standard deviation, minimum, and maximum of the estimates for τ_T^1 computed from the control groups in $\mathcal{S}$; and the *Common* column reports the proportion of units in each control group that were common across all control groups, computed as $\left|\bigcap_{C' \in \mathcal{S}} C'\right| / s$.

Cases in which alternate solutions were identified did not exhibit a large amount of variability in the estimates. All estimates from *nswre74cps* remain above \$1100, while all estimates from *nswcps* remain below -\$150. Similarly, the estimates from *nswre74psid* with $\mathcal{I}_{\text{DOM}}$ remain slightly positive while those from *nswpsid* with $\mathcal{I}_{\text{DOM}}$ remain significantly negative (though the control groups identified in the latter case are all comprised of a very large common set of control units making up almost 96% of each group). No alternate optima were identified for $\mathcal{I}_{\text{Diff}^2}$, which is likely due to the fact that the modifications in Step 3 transform Model (5.6) from a mixed integer quadratic program to a mixed integer quadratically constrained program (MIQCP); for the other imbalance measures, the modified models remain mixed integer programs.

Figure 5.9 shows the distributions of treatment effect estimates from *nswcps* (first row) and *nswre74cps*


Estimates for *nswcps* with $\mathcal{I}_{\text{Diff}}$ (20 bins)

Estimates for *nswcps* with $\mathcal{I}_{\text{Diff}}$ (20 bins)

Estimates for *nswre74cps* with $\mathcal{I}_{\text{Diff}}$ (20 bins)

Estimates for *nswre74cps* with $\mathcal{I}_{\text{Diff}}$ (20 bins)

Figure 5.9: Distributions of treatment effect estimates from *nswcps* and *nswre74cps* using $\mathcal{I}_{\text{Diff}}$ with 20 bins.

(second row) using $\mathcal{I}_{\text{Diff}}$ (other imbalance measures are not included due to the small number of solutions found for them). The plots in the first column are histograms for the distributions with bin widths set using the Freedman-Diaconis rule, and the plots in the second column are the cumulative distribution functions for the estimates. Shapiro-Wilk normality tests for these distributions yield p-values of 0.576 for *nswcps* and 0.543 for *nswre74cps*.

In the absence of any knowledge about the experimental estimate of the treatment effect, it would be difficult to discredit the estimates from *nswcps* using the results from Table 5.11 and Figure 5.9. Knowledge about the importance of an extra year of pre-treatment income (the RE74 covariate) allowed researchers to construct the *nswre74cps* dataset, which emphasizes the necessity of including all covariates in the dataset. Concern about missing covariates necessitates a fairly comprehensive search through the space of potential control group solutions, which might assist analysis. This task is complicated by the size of the solution space. While exhaustive enumeration can identify all control group solutions, in practice it is likely to be too time-consuming. An alternate approach is to search through these control groups in a more targeted manner; such an approach is explored in the next section.

5.3.4 Extreme Estimates

Instead of searching for a diverse subset of optimal control groups, an alternate approach is to identify the range of treatment effect estimates across the optimal control groups. Specifically, one can look at the smallest and largest possible estimates across all optimally balanced control groups. The MIP models can be modified in a similar manner as before, except with the new objective

$$\min\ (\max)\ \sum_{c \in C} y_c^0 v_c. \tag{5.17}$$

The range of estimates can be computed by first performing Step 1 from the procedure in Section 5.3.3 to identify a near-optimal control group C^*, and then either minimizing or maximizing objective (5.17) subject to the constraint that $\mathcal{I}(T, C') \leq \mathcal{I}(T, C^*)$.

If the range of estimates is small, then confidence is gained in both the estimates and the assumptions that were made. A small range also suggests that the error terms are reasonably well-behaved (e.g., small variance). On the other hand, if the range of estimates is large, then one or more of the following is likely: (1) any residual imbalance is still significant and can be exploited to amplify differences in the estimate of the treatment effect across control groups; (2) the assumptions for the imbalance measure are violated, either due to omitted terms in the control response function (e.g., squared terms, covariate interactions) or unobserved covariates, resulting in bias in the estimates; or (3) the error terms have a large variance. Provided that the optimization process is able to remove most of the imbalance, it should be possible to rule out (1). Definitively attributing a large range to either (2) or (3) is difficult but can be explored by adding additional balance requirements to see if they reduce the range.

One potential issue with the above procedure is that it looks at the control responses of the units in order to identify a control group. Within the matching literature, this is generally regarded as undesirable because it can lead to the selection of a control group on the basis of the estimate that it yields (e.g., if a researcher believes that an effect should be present, he or she may be inclined to select a control group that confirms this). However, provided both extremes are identified and reported, this should not be an issue.

Another potential issue is that the procedure essentially searches for outliers within the data. One would normally expect to find a few outliers in any dataset, so one could argue that one or two outliers within the set

Table 5.12: Range of estimates for the LaLonde datasets.

		Permitted	$\widetilde{\tau}_T^1$ ($)				
Set	Objective	Imbalance	LP Min	MIP Min	MIP Max	LP Max	Common
nswcps1	$\mathcal{I}_{\text{DOM}}$	0.0009	−7455.08	−608.58	4793.96	4803.31	0.4646
	$\mathcal{I}_{\text{KS}}$	0.0236	−2224.15	−2180.28	1084.40	1086.72	0.7609
	$\mathcal{I}_{\text{KS:max}}$	0.0067	−2038.99	−2038.99	552.77	555.15	0.7946
	$\mathcal{I}_{\text{Diff}}$ (20)	0.0404	−2074.27	−2071.69	907.71	909.30	0.7576
	$\mathcal{I}_{\text{Diff}^2}$ (20)	0.0001	−2354.60	−499.98	−499.98	998.85	1.0000
nswpsid1	$\mathcal{I}_{\text{DOM}}$	3.8791	−3592.96	−3422.30	−3202.79	−2988.19	0.9764
nswre74cps1	$\mathcal{I}_{\text{DOM}}$	0.0088	−6522.48	1305.72	1598.40	6091.83	0.9730
	$\mathcal{I}_{\text{KS}}$	0.0811	1129.38	1190.38	2257.64	2282.39	0.8486
	$\mathcal{I}_{\text{KS:max}}$	0.0216	248.25	253.05	3277.09	3283.30	0.7514
	$\mathcal{I}_{\text{Diff}}$ (20)	0.1189	457.94	487.79	2823.35	2826.94	0.7730
	$\mathcal{I}_{\text{Diff}^2}$ (20)	0.0010	−329.76	1601.74	1601.74	3080.77	1.0000
nswre74psid1	$\mathcal{I}_{\text{DOM}}$	5.4521	445.19	684.16	1053.99	1640.57	0.9892

of estimates from optimally balanced control groups is not necessarily an indication that the assumptions are invalid, as in (2). However, because $\widetilde{\tau}_T^1$ is computed as an average of control responses, in order to attribute a large estimate range primarily to (3), the outlier estimates must be computed from a large number of error terms that are moderate outliers, a handful of error terms that are extreme outliers, or possibly some combination of both. Either case is likely to be cause for concern, so it is useful to be able to identify such occurrences.

The above procedure was performed for the datasets and imbalance measures in Table 5.11. A time limit of 300 seconds was used to find the initial optimal or near-optimal solution C^*, and then the revised models with objective (5.17) and constraint $\mathcal{I}(T, C') \leq \mathcal{I}(T, C^*)$ were solved with a time limit of 300 seconds each.

The results from these tests are shown in Table 5.12. The *Set* column lists the dataset; the *Objective* column lists the imbalance measure used for optimization; the *Permitted Imbalance* column lists the allowed imbalance imposed as a constraint in the MIP model; the $\widetilde{\tau}_T^1$ *($)* columns give various values of the estimated treatment effect across all control group solutions at or below the imbalance threshold, with *LP Min* and *LP Max* reporting the minimum and maximum values of the linear programming relaxation of the model and *MIP Min* and *MIP Max* reporting the minimum and maximum estimates from integer solutions identified by CPLEX (these solutions are not necessarily optimal due to time limits); and the *Common* column gives the proportion of overlap between the extreme solutions identified by CPLEX, computed as $\left|C'_{\max} \cap C'_{\min}\right| / s$.

In general, the estimate ranges reported in Table 5.12 contain the estimates from alternate optima in Table 5.11. The exception is $\mathcal{I}_{\text{DOM}}$, where several of the minimum and maximum values reported in Table

5.12 are inferior to those reported in Table 5.11. This indicates that the model for maximizing or minimizing the treatment effect estimate was unable to identify an "extreme" control group that was identified during the search for alternate optima.

The results in the *Common* column in Table 5.12 highlight one difficulty with this approach. If the permitted imbalance is small, there are only a few available control groups that meet the balance requirement, and thus the overlap between these solutions is generally large. A natural consequence of large overlap is a small range, so it becomes difficult to determine if a small range is due to too much overlap or due to the balance requirements effectively removing bias from the covariates. Ideally, one would be able to identify extreme solutions with no imbalance and little to no overlap. In this case, a large estimate range would be an indication of either (2) or (3), while a small range would be an indication that the balance requirements are sufficient to remove bias in the estimate.

Unfortunately, the ideal case of small overlap and small range does not occur within any of the LaLonde data. However, a few observations can still be made. First, for *nswre74cps* the ranges for the treatment effect estimates are always positive, though the ranges are relatively large ($\mathcal{I}_{\text{KS:max}}$ has the largest range of approximately \$3000). While the large ranges may indicate that the corresponding functional form assumptions are invalid, including additional balance requirements (e.g., joint imbalance is minimal) while maintaining the specified levels of marginal imbalance can only reduce the estimate ranges (while potentially increasing overlap). This strengthens the case for a beneficial effect from the treatment (the job training program), even if the magnitude of that effect is still not immediately clear from the data.

On the other hand, for *nswcps*, the ranges for $\mathcal{I}_{\text{DOM}}$, $\mathcal{I}_{\text{KS}}$, $\mathcal{I}_{\text{KS:max}}$, and $\mathcal{I}_{\text{Diff}}$ all span zero. For $\mathcal{I}_{\text{DOM}}$ in particular, the range is over \$5000 and the overlap between the extreme solutions is less than 50%. The wide ranges may be caused by incorrect assumptions about the control response function for each of the imbalance measures. One way in which the preceding issue can be explored is to use an imbalance measure that includes joint distributions such as $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ with $\mathbf{D} = \binom{\mathcal{P}}{2}$. For the *nswcps* dataset, however, all potential control groups have a large residual imbalance with respect to $\mathcal{I}_{\text{Diff:}\mathbf{D}}$, which makes it difficult to test the corresponding assumptions on the form of the control response function and to determine whether the treatment is beneficial, detrimental, or has no effect. Further explorations might consider the possibility of unobserved covariates as Dehejia and Wahba (1999) did, or expand the dataset in order to find control groups featuring little residual imbalance for $\mathcal{I}_{\text{Diff:}\mathbf{D}}$ and other imbalance measures that use the joint distributions of the covariates.

5.3.5 Control Pool Size

One of the difficulties with using the *nswre74psid* dataset compared to *nswre74cps* is the difference in control pool size: *nswre74psid* contains 2490 control units while *nswre74cps* contains 15992. The larger pool of units assists in minimizing residual imbalance and consequently ensures a more accurate estimate of the treatment effect, as evidenced by the results in Table 5.10. To explore the question of how large the control pool needs to be in order to achieve good balance, a set of experiments were performed to randomly drop units from the control pool. After units were dropped, BOSS identified optimal solutions from the reduced control pool and constructed estimates from them. The number of dropped units and the random seed were varied over a set of experiments.

Matching methods were also tested. The specific methods that were tested were propensity score matching with a logistic regression model to estimate the propensity score, covariate matching with unit distance defined using the Euclidean distance metric, and covariate matching using the Mahalanobis distance metric. The *optmatch* package was used to construct the solutions for matching (Hansen, 2007).

The results are shown in Figure 5.10. The plots show the estimated treatment effect computed from solutions identified by BOSS (left column) and matching (right column) as the size of the control pool grows from 250 units to 15000. For BOSS, the plots from top to bottom are for $\mathcal{I}_{\text{DOM}}$, $\mathcal{I}_{\text{KS}}$, and $\mathcal{I}_{\text{Diff}}$ with $n_i = 20$ for all covariates, while for matching, the plots from top to bottom are Mahalanobis matching, Euclidean distance matching, and propensity score matching (using a Euclidean metric to define distance between propensity scores).

From Figure 5.10, the methods can be ranked by how large the control pool needs to be before the estimates are above zero on average. BOSS with $\mathcal{I}_{\text{DOM}}$ and Euclidean distance matching accomplish this after approximately 750 units; BOSS with $\mathcal{I}_{\text{Diff}}$ and $\mathcal{I}_{\text{KS}}$ accomplish this between 1500 and 2000 units; propensity score matching accomplishes this between 2000 and 4000 units; and Mahalanobis matching accomplishes this after 8000 units. Another ranking can be computed by looking at the number of units needed to produce an estimate above the experimental estimate of \$1794. Euclidean distance matching accomplishes this after 1500 units; $\mathcal{I}_{\text{KS}}$ accomplishes this between 4000 and 5000 units; propensity score matching accomplishes this after 6000 units; $\mathcal{I}_{\text{DOM}}$ and $\mathcal{I}_{\text{Diff}}$ accomplish this between 7000 and 8000 units; and Mahalanobis matching never accomplishes this.

Euclidean distance matching appears to be the clear winner based on the preceding discussion. However, while it may be the first method to report an estimate above \$1794, the majority of its estimates remain below this level, particularly as the size of the control pool grows. Similarly, Mahalanobis matching appears

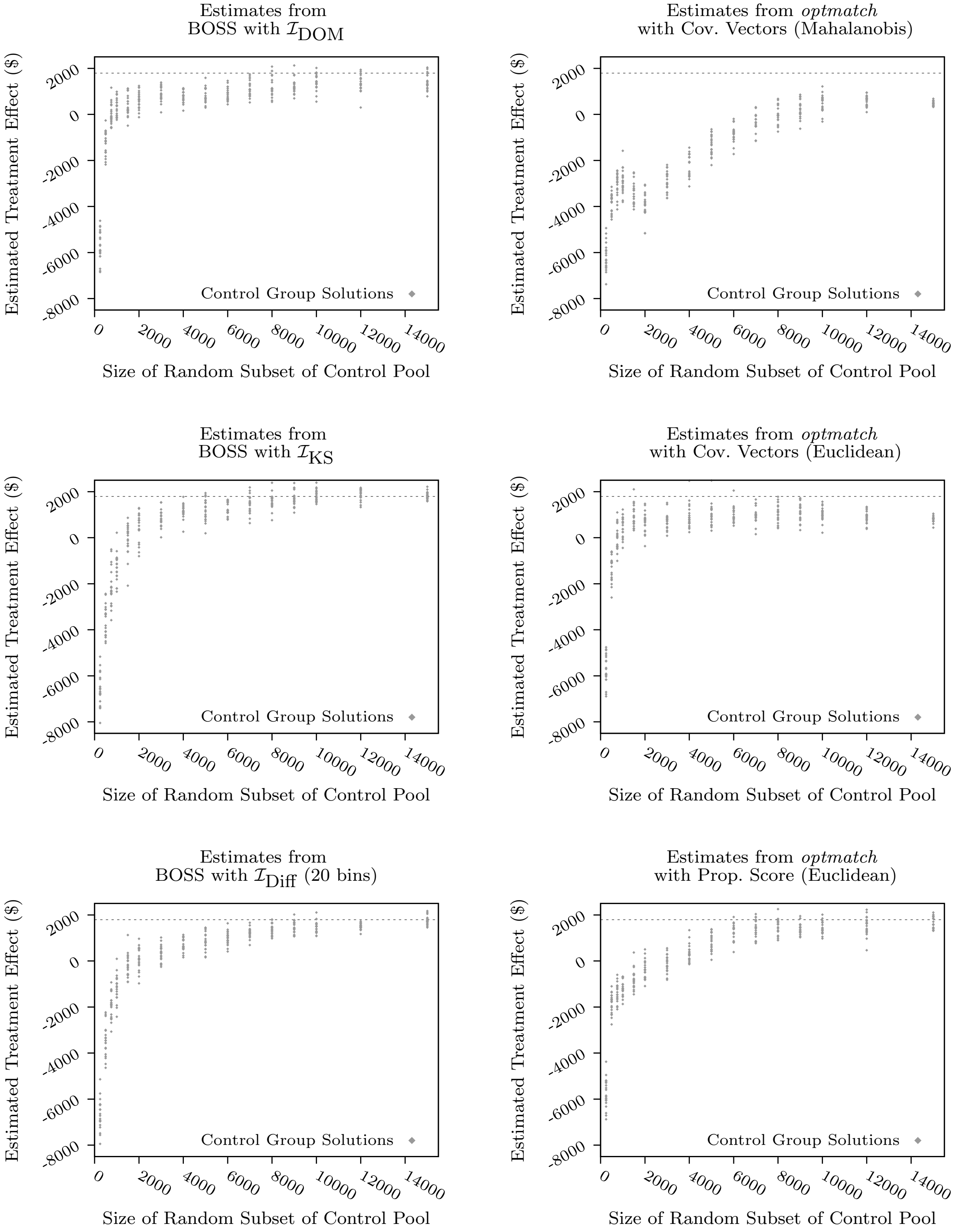


Figure 5.10: Estimates of the treatment effect for BOSS and matching on a subset of the control pool for *nswre74cps*.

to be the clear loser, with its estimates getting worse at one point as the control pool increases in size. Propensity score matching takes some time to ramp up but performs consistently as the size of the control pool increases. In comparison, BOSS with $\mathcal{I}_{\text{KS}}$ and $\mathcal{I}_{\text{Diff}}$ ramp up more quickly and also produce estimates close to the experimental benchmark in the long run. BOSS with $\mathcal{I}_{\text{DOM}}$ does better if the size of the control pool is small, but fails to keep pace with the distribution-based imbalance measures as the control pool increases in size. This is unsurprising given the nature of balancing means compared to marginal distributions.

5.4 Discussion

The tests in this chapter illustrate how the BOSS framework can be used to identify control groups for constructing treatment effect estimates. For the computational tests with simulated data, the accurate estimates of τ_T^1 produced by BOSS show that it is a viable approach for estimating treatment effects in observational data. In addition, the comparisons with matching methods illustrate the primary motivation for BOSS: obtaining covariate balance. In particular, the tests demonstrate that by minimizing an imbalance measure directly, it is possible to identify solutions with better covariate balance than those found by standard matching methods. The tests also demonstrate how a failure to remove all necessary imbalance can leave residual bias in the estimate of the treatment effect.

The tests with the LaLonde data illustrate that BOSS can handle real data in addition to simulated data. However, like traditional methods for dealing with observational data, BOSS assumes that all covariates that influence the outcome are known and included. If this is not the case, imbalances in unobserved covariates may introduce bias. Additionally, a small control pool can potentially limit the effectiveness of BOSS in removing imbalance, particularly for distribution-based imbalance measures.

Several computational tests illustrate how alternate optima help analyze the assumptions that are made. In particular, a non-normal distribution of the estimates or a wide range in the maximum and minimum estimate might call into question the functional form assumption. However, alternate optima are not always sufficiently diverse, and overlap between the different control groups remains a significant concern if attempting to draw conclusions from the estimates of the treatment effect.

Most of the tests focused on achieving balance on the marginal distributions of the covariates and not the joint distributions. However, if a method fails to produce reliable balance on the marginal distributions, then it generally will not produce balance on higher-order distributions either. By focusing on the marginal distributions first, one can determine whether sufficient marginal balance is possible before proceeding to search for solutions with balance on higher-order distributions. For example, with the LaLonde datasets, almost all imbalance could be removed on the marginal distributions but not on the joint distributions. Because the current theory from Chapter 3 only applies if there is no residual imbalance, focusing on the marginals allowed the theory to be tested with the LaLonde data. Extending the current theory to quantify the impact of residual imbalance on the estimate of the treatment effect is a direction for future research.

Chapter 6

Conclusion

This dissertation presented a comprehensive overview of the Balance Optimization Subset Selection (BOSS) framework which is used to derive causal estimates from observational (non-random) data. For an observational dataset consisting of a treatment group and a large pool of control units, BOSS has as its goal the identification of a subset of control units that minimizes a measure of imbalance with respect to the treatment group. The BOSS framework encourages: (1) the specification of assumptions regarding the interaction between the covariates and the control responses; (2) the determination of sources of imbalance that need to be removed in order to provide an unbiased estimate of the treatment effect; (3) the minimization of this imbalance using an appropriate optimization model; and (4) the construction of an estimate of the treatment effect if residual imbalance is minimal or nonexistent. In addition, BOSS can identify alternate optima to help assess the sensitivity of the estimates to changes in the control group. Alternate optima can also be used to alert the researcher to potential problems with the assumptions.

Three major aspects of the BOSS framework were discussed in this dissertation. The first aspect focused on the role of covariate balance in causal inference for observational data. Specifically, it explored the relationship between the covariates, the control response function, and bias in the estimate of the treatment effect. Under appropriate assumptions on the function that relates the covariates to the control responses, it was demonstrated that certain levels of covariate balance are necessary and sufficient in order to construct an unbiased estimate of the treatment effect. As the assumptions become weaker and the control response function becomes more general, the covariate balance requirements become more strict. The weakest assumption is *strong ignorability*, which requires exact balance on the full joint distribution of the covariates between the treatment and control groups. Collectively, the assumptions unify existing matching and regression methods for causal inference within the *balance hierarchy*. BOSS is designed to accommodate these assumptions by searching for a control group that satisfies the appropriate balance requirements.

The second aspect of the BOSS framework was the computational complexity of the associated decision and optimization problems. In particular, BOSS with several imbalance measures is **NP-Hard** except in a few restricted cases with a limited number of covariates. Approximating several of the optimization problems for BOSS with minimization imbalance measures to within any factor of the optimal solution was shown to be impossible unless $\mathbf{P} = \mathbf{NP}$ because such an approximation algorithm would lead to a polynomial-time algorithm for deciding an **NP-Complete** problem. However, a constant-factor approximation algorithm was constructed for a maximization problem resulting from a reformulation of one imbalance measure.

Finally, the third aspect of BOSS considered its application in practice. A simulated annealing heuristic and several mathematical programming models were presented for solving the optimization problems associated with BOSS. Despite the complexity of these problems, computational results demonstrated that the heuristic and the mixed integer programming solver CPLEX were able to find near-optimal solutions for moderately sized instances in reasonable time. Additionally, the solutions identified by BOSS featured improved covariate balance compared to traditional matching methods and they also produced accurate estimates of the treatment effect on simulated data. Further computational tests demonstrated the suitability of BOSS with several potential imbalance measures on the well-studied dataset of LaLonde (1986) and also considered how alternate optima can be used to assess the assumptions made by BOSS.

There are many avenues for future research. Of primary importance is an investigation into the impact of residual imbalance on the bias in the treatment effect estimate, because it is unlikely that all imbalance can be removed from most real-world datasets. Identifying how this bias can be characterized and bounded is essential to constructing accurate estimates of the treatment effect if residual imbalance cannot be removed. For example, if marginal distribution balance cannot be fully achieved but marginal moment balance can, then a process for distinguishing the better form of balance is needed.

Residual imbalance also complicates the combination of all covariate imbalance measures into a single objective function. If the values of two covariates are on different scales, then $\mathcal{I}_{\text{DOM}}$ will prioritize removing imbalance on the covariate with larger values. This presents no problem if imbalance on both covariates can be removed but not when some residual imbalance is present. Scaling measures like $\mathcal{I}_{\text{SDOM}}$ can accommodate this to some extent, but further development in this area is warranted. More generally, if a combination of moment-based imbalance measures and distribution-based imbalance measures are used, then it is necessary to ensure that they are appropriately scaled for combination into a single objective.

In the presence of residual imbalance, BOSS can be formulated as a multi-objective optimization problem, where each component of the objective is an imbalance measure for a single covariate or a subset of the covariates. In this manner, the trade-offs associated with minimizing one source of imbalance at the expense of others can be explored and potentially quantified. A sensitivity analysis using this formulation could provide information on the impacts of each of the covariates on the control responses.

The current sensitivity analysis discussed for BOSS can also be expanded in many ways. One possibility is to incorporate a measure of solution overlap into the search for extreme estimates. Finding two distinct solutions with a difference in estimates that is slightly less than the maximum range would be more useful than two extreme solutions with a large amount of overlap. Another possibility is the development of a better procedure for identifying a diverse set of alternate optima. A related direction is to investigate the error terms in order to characterize their impact on the estimates of the treatment effect arising from optimally balanced solutions. Determining what range and variation can be expected in practice given the possibility of overlap and developing statistical tests to check these expectations are both important goals for expanding the utility of the sensitivity analysis. Another possibility is the exploration of parallels between bootstrapping techniques and estimates from alternate optima for BOSS.

While the majority of the imbalance measures considered here can be modeled as mixed integer (linear or quadratic) programs and solved with CPLEX, work needs to be done on developing specialized exact algorithms both for identifying good solutions quickly and for verifying optimality. Such algorithms will most likely be necessary for handling larger datasets. Future work should also focus on the development of tighter mathematical programming formulations and problem-specific cuts. Additionally, methods should be developed for solving the optimization problems without the constraint on the size of the control group.

Another possibility is to search for approximation-preserving reductions from **NP-Hard** optimization problems to BOSS with various imbalance measures. More generally, this could include the identification of problems whose structure is similar to that of BOSS in order to potentially use existing techniques for those problems. Additional research into approximation algorithms for reformulations of the current imbalance measures will help identify initial solutions for use in exact algorithms.

From the application perspective, BOSS should be extended to handle the case of a non-binary treatment. This could be used to assess the effectiveness of treatment at different dosage levels. Another possibility is the adaptation of BOSS to construct estimates of unit-level, or personalized, treatment effects. This would be particularly suitable for applying BOSS to comparative effectiveness research (Concato et al., 2010). Updating the BOSS framework to handle the possibility of missing covariate data is also important for dealing with many real-world datasets.

Finally, identifying additional datasets on which BOSS can be tested remains an important direction for future work and validation. In conjunction with this, the development of a thorough list of guidelines and suggested best practices for using BOSS would be beneficial to researchers and practitioners in their work.

References

Abadie, A., G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**(1) 235–267.

Abadie, A., G. W. Imbens. 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* **29**(1) 1–11.

Ahuja, R. K., T. L. Magnanti, J. B. Orlin. 1993. *Network Flows: Theory, Algorithms and Applications*. Prentice Hall.

Althauser, R. P., D. B. Rubin. 1970. The computerized construction of a matched sample. *Amer. J. Sociol.* **76**(2) 325–346.

Cho, W. K. T., J. J. Sauppe, A. G. Nikolaev, S. H. Jacobson, E. C. Sewell. 2013. An optimization approach for making causal inferences. *Stat. Neerl.* **67**(2) 211–226.

Cochran, W. G., D. B. Rubin. 1973. Controlling bias in observational studies. *Sankhya Ser. A* **35**(4) 417–446.

Concato, J., E. V. Lawler, Lew R. A., J. M. Gaziano, M. Aslan, G. D. Huang. 2010. Observational methods in comparative effectiveness research. *Amer. J. Med.* **123**(12, Supplement) 16–23.

da Veiga, P. V., R. P. Wilder. 2008. Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health Journal* **12**(2) 194–203.

Dawid, A. P. 1979. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B* **41**(1) 1–31.

Dehejia, R. 2005. Practical propensity score matching: A reply to Smith and Todd. *J. Econometrics* **125**(1-2) 355–364.

Dehejia, R. H., S. Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Stat. Assoc.* **94**(448) 1053–1062.

Dehejia, R. H., S. Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* **84**(1) 151–161.

Diamond, A., J. S. Sekhon. 2013. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* **95**(2) 932–945.

D'Orazio, M. 2012. *StatMatch: Statistical Matching*. URL `http://CRAN.R-project.org/package=StatMatch`. R package version 1.1.0.

Edmonds, J. 1965. Maximum matchings and a polyhedron with 0,1-vertices. *J. Res. Natl. Bureau Standards* **69**(B) 125–130.

Fasano, G., A. Franceschini. 1987. A multidimensional version of the kolmogorov-smirnov test. *Monthly Notices of the Royal Astronomical Society* **225** 155–170.

Feige, U. 1998. A threshold of $\ln n$ for approximating set cover. *J. ACM* **45**(4) 634–652.

Garey, M. R., D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, New York, NY, USA.

Greenberg, B. G. 1953. The use of analysis of covariance and balancing in analytical surveys. *American Journal of Public Health* **43**(6) 692–699.

Hainmueller, J. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**(1) 25–46.

Hansen, B. B. 2007. Optmatch: Flexible, optimal matching for observational studies. *R News* **7**(2) 18–24.

Hansen, B. B., S. O. Klopfer. 2006. Optimal full matching and related designs via network flows. *J. Comput. Graph. Stat.* **15**(3) 609–627.

Herron, M. C., J. Wand. 2007. Assessing partisan bias in voting technology: The case of the 2004 New Hampshire recount. *Electoral Studies* **26**(2) 247–261.

Ho, D. E., K. Imai, G. King, E. A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15**(3) 199–236.

Holland, P. W. 1986. Statistics and causal inference. *J. Amer. Stat. Assoc.* **81**(396) 945–960.

Iacus, S. M., G. King, G. Porro. 2011. Multivariate matching methods that are monotonic imbalance bounding. *J. Amer. Stat. Assoc.* **106**(493) 345–361.

Iacus, S. M., G. King, G. Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* **20**(1) 1–24.

Imai, K. 2005. Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *Amer. Polit. Sci. Rev.* **99**(2) 283–300.

Imai, K., G. King, E. A. Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A-STA* **171**(2) 481–502.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86**(1) 4–29.

Justel, A., D. Peña, R. Zamar. 1997. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Stat. Prob. Lett.* **35**(3) 251–259.

Kelz, R. R., C. E. Reinke, J. R. Zubizarreta, M. Wang, P. Saynisch, O. Even-Shoshan, P. P. Reese, L. A. Fleisher, J. H. Silber. 2013. Acute kidney injury, renal function, and the elderly obese surgical patient: A matched case-control study. *Ann. Surg.* **258**(2) 359–363.

King, G., L. Zeng. 2006. The dangers of extreme counterfactuals. *Polit. Anal.* **14**(2) 131–159.

Kirkpatrick, S., C. C. Gelatt, M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* **220**(4598) 671–680.

Kuhn, H. 1955. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2**(1-2) 83–97.

LaLonde, R. J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *Amer. Econ. Rev.* **76**(4) 604–620.

Marko, N. F., R. J. Weil. 2010. The role of observational investigations in comparative effectiveness research. *Value in Health* **13**(8) 989–997.

Nemhauser, G. L., L. A. Wolsey, M. L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions – I. *Math. Programming* **14**(1) 265–294.

Nikolaev, A. G., S. H. Jacobson, W. K. T. Cho, J. J. Sauppe, E. C. Sewell. 2013. Balance Optimization Subset Selection (BOSS): An alternative approach for causal inference with observational data. *Oper. Res.* **61**(2) 398–412.

Peacock, J. A. 1983. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society* **202** 615–627.

Ratkovic, M. 2012. Identifying the largest balanced subset of the data under general treatment regimes. Tech. rep., Department of Politics, Princeton University.

Reinke, C. E., R. R. Kelz, J. R. Zubizarreta, M. Lanyu, P. Saynisch, O. Even-Shoshan, L. A. Fleisher, J. H. Silber. 2012. Obesity and readmission in elderly surgical patients. *Surgery* **152**(3) 355–362.

Rosenbaum, P. R. 1984. From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Amer. Stat. Assoc.* **79**(385) 41–48.

Rosenbaum, P. R. 1987a. The role of a second control group in an observational study. *Stat. Sci.* **2**(3) 292–306.

Rosenbaum, P. R. 1987b. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74**(1) 13–26.

Rosenbaum, P. R. 1989. Optimal matching for observational studies. *J. Amer. Stat. Assoc.* **84**(408) 1024–1032.

Rosenbaum, P. R. 1991. A characterization of optimal designs for observational studies. *J. R. Stat. Soc. Ser. B-M* **53**(3) 597–610.

Rosenbaum, P. R., R. N. Ross, J. H. Silber. 2007. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Stat. Assoc.* **102**(477) 75–83.

Rosenbaum, P. R., D. B. Rubin. 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B-M* **45**(2) 212–218.

Rosenbaum, P. R., D. B. Rubin. 1983b. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1) 41–55.

Rosenbaum, P. R., D. B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Stat. Assoc.* **79**(387) 516–524.

Rosenbaum, P. R., D. B. Rubin. 1985. Constructing a control group using multivariate matched sampling models that incorporate the propensity score. *Amer. Stat.* **39**(1) 33–38.

Rosenbaum, Paul R. 2012. Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Stat.* **21**(1) 57–71.

Rubin, D. B. 1973a. Matching to remove bias in observational studies. *Biometrics* **29**(1) 159–183.

Rubin, D. B. 1973b. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29**(1) 185–203.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**(5) 688–701.

Rubin, D. B. 1976a. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics* **32**(1) 109–120.

Rubin, D. B. 1976b. Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics* **32**(1) 121–132.

Rubin, D. B. 1978. Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* **6**(1) 34–58.

Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Stat. Assoc.* **74**(366) 318–328.

Rubin, D. B. 1980. Bias reduction using Mahalanobis-metric matching. *Biometrics* **36**(2) 293–298.

Rubin, D. B. 1991. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47**(4) 1213–1234.

Sauppe, J. J., S. H. Jacobson, E. C. Sewell. 2014. Complexity and approximation results for the Balance Optimization Subset Selection model for causal inference in observational studies. *INFORMS J. Comput.* **26**(3) 547–566.

Sekhon, J. S. 2004. Quality meets quantity: Case studies, conditional probability and counterfactuals. *Perspectives on Politics* **2**(2) 281–293.

Sekhon, J. S. 2009. Opiates for the matches: Matching methods for causal inference. *Annu. Rev. Polit. Sci.* **12** 487–508.

Sekhon, J. S. 2011. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Stat. Softw.* **42**(7) 1–52.

Smith, J. A., P. E. Todd. 2001. Reconciling conflicting evidence on the performance of propensity score matching methods. *Amer. Econ. Rev.* **91**(2) 112–118.

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

Smith, J. A., P. E. Todd. 2005b. Rejoinder. *J. Econometrics* **125**(1-2) 365–375.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **25**(1) 1–21.

Vazirani, V. V. 2001. *Approximation Algorithms*. Springer-Verlag, New York, NY, USA.

Witkin, H. A., S. A. Mednick, F. Schulsinger, E. Bakkestrom, K. O. Christiansen, D. R. Goodenough, K. Hirschhorn, C. Lundsteen, D. R. Owen, J. Philip, D. B. Rubin, M. Stocking. 1976. Criminality in XYY and XXY men. *Science* **193**(4253) 547–555.

Yang, D., D. S. Small, J. H. Silber, P. R. Rosenbaum. 2012. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68**(2) 628–636.

Zubizarreta, J. R. 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Stat. Assoc.* **107**(500) 1360–1371.

Zubizarreta, J. R., M. Cerdá, P. R. Rosenbaum. 2013a. Effect of the 2010 chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiology* **24**(1) 79–87.

Zubizarreta, J. R., M. D. Neuman, J. H. Silber, P. R. Rosenbaum. 2012. Contrasting evidence within and between institutions that supply treatment in an observational study of alternative forms of anesthesia. *J. Amer. Stat. Assoc.* **107**(499) 901–915.

Zubizarreta, J. R., C. E. Reinke, R. R. Kelz, J. H. Silber, P. R. Rosenbaum. 2011. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Amer. Stat.* **65**(4) 229–238.

Zubizarreta, J. R., D. S. Small, N. K. Goyal, S. A. Lorch, P. R. Rosenbaum. 2013b. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Stat.* **7**(1) 25–50.

Appendix

Table A.1: Solutions for *data3c10k* ranked by $\mathcal{I}_{\chi^2}$ with 32 bins per covariate.

Imbalance		$\widetilde{\tau}_T^1$		KS	
Range	Observations	μ	σ	μ	σ
$\leq 1.0 \times 10^{-7}$	833	0.0122	0.0900	0.0274	0.0027
1.0×10^{-7} - 1.0	4,377	0.0679	0.0950	0.0282	0.0028
1.0 - 2.0	4,675	0.1478	0.1111	0.0294	0.0029
2.0 - 3.0	3,747	0.2291	0.1173	0.0312	0.0032
3.0 - 4.0	3,098	0.2948	0.1183	0.0328	0.0034
4.0 - 5.0	2,751	0.3596	0.1233	0.0344	0.0035
5.0 - 6.0	2,308	0.4085	0.1304	0.0356	0.0035
6.0 - 7.0	2,022	0.4666	0.1303	0.0370	0.0036
7.0 - 8.0	1,873	0.5173	0.1306	0.0381	0.0037
8.0 - 9.0	1,670	0.5584	0.1315	0.0394	0.0037
9.0 - 10.0	1,544	0.5881	0.1355	0.0402	0.0038
10.0 - 20.0	10,937	0.7889	0.1790	0.0449	0.0047
20.0 - 30.0	8,313	1.1213	0.1828	0.0528	0.0044
30.0 - 40.0	7,009	1.4045	0.1974	0.0597	0.0046
40.0 - 50.0	6,148	1.6617	0.1956	0.0659	0.0045
50.0 - 60.0	5,416	1.8779	0.2050	0.0713	0.0047
60.0 - 70.0	4,910	2.0778	0.2125	0.0762	0.0048
70.0 - 80.0	4,437	2.2490	0.2160	0.0808	0.0049
80.0 - 90.0	3,920	2.4258	0.2159	0.0854	0.0049
90.0 - 100.0	3,745	2.5803	0.2250	0.0892	0.0052
100.0 - 200.0	28,789	3.2693	0.4242	0.1068	0.0104
200.0 - 300.0	21,784	4.3260	0.3753	0.1332	0.0089
300.0 - 400.0	18,056	5.1472	0.3547	0.1531	0.0083
400.0 - 500.0	16,629	5.8218	0.3482	0.1692	0.0082
500.0 - 600.0	15,382	6.4058	0.3489	0.1824	0.0082
600.0 - 700.0	14,876	6.9137	0.3475	0.1939	0.0082
700.0 - 800.0	13,974	7.3915	0.3588	0.2044	0.0084
800.0 - 900.0	13,697	7.8138	0.3648	0.2136	0.0084
900.0 - 1000.0	13,336	8.2038	0.3614	0.2220	0.0085
1000.0 - 2000.0	66,747	9.2378	0.6579	0.2430	0.0137
2000.0 - 3000.0	2,638	11.5269	0.5772	0.2857	0.0121
3000.0 - 4000.0	34	12.7366	0.3777	0.3076	0.0099

Table A.2: Solutions for *data10c10k* ranked by $\mathcal{I}_{\chi^2}$ with 32 bins per covariate.

Imbalance		$\widetilde{\tau}_T^1$		KS	
Range	Observations	μ	σ	μ	σ
$\leq$ 2.0	0	N/A	N/A	N/A	N/A
2.0 - 3.0	1	0.2168	0.0000	0.0260	0.0000
3.0 - 4.0	25	0.2409	0.1056	0.0251	0.0014
4.0 - 5.0	116	0.2809	0.1113	0.0251	0.0016
5.0 - 6.0	229	0.3567	0.1065	0.0255	0.0014
6.0 - 7.0	332	0.4024	0.1198	0.0259	0.0013
7.0 - 8.0	327	0.4467	0.1189	0.0262	0.0016
8.0 - 9.0	377	0.4914	0.1200	0.0267	0.0016
9.0 - 10.0	350	0.5159	0.1225	0.0271	0.0015
10.0 - 20.0	3,305	0.7416	0.1719	0.0295	0.0021
20.0 - 30.0	3,105	1.0607	0.1679	0.0328	0.0021
30.0 - 40.0	2,737	1.3523	0.1748	0.0359	0.0021
40.0 - 50.0	2,677	1.6002	0.1855	0.0384	0.0022
50.0 - 60.0	2,608	1.8155	0.1970	0.0409	0.0022
60.0 - 70.0	2,649	2.0576	0.1899	0.0434	0.0023
70.0 - 80.0	2,499	2.2616	0.1956	0.0456	0.0024
80.0 - 90.0	2,527	2.4404	0.2036	0.0477	0.0024
90.0 - 100.0	2,221	2.6453	0.2113	0.0499	0.0024
100.0 - 200.0	22,575	3.5447	0.5077	0.0600	0.0058
200.0 - 300.0	20,452	4.9360	0.4429	0.0760	0.0051
300.0 - 400.0	19,317	6.0623	0.4187	0.0889	0.0048
400.0 - 500.0	19,179	7.0259	0.3946	0.1000	0.0046
500.0 - 600.0	19,216	7.8660	0.3913	0.1096	0.0046
600.0 - 700.0	19,778	8.6073	0.3793	0.1178	0.0045
700.0 - 800.0	20,144	9.2818	0.3880	0.1254	0.0047
800.0 - 900.0	21,140	9.8486	0.3882	0.1317	0.0048
900.0 - 1000.0	21,901	10.3721	0.3997	0.1373	0.0048
1000.0 - 2000.0	31,255	11.1042	0.5728	0.1448	0.0064
2000.0 - 3000.0	11	13.5719	0.6053	0.1663	0.0081

Table A.3: Solutions for *data10c10k* ranked by $\mathcal{I}_{\text{DOM}}$.

Imbalance		$\widetilde{\tau}_T^1$		KS	
Range	Observations	μ	σ	μ	σ
$\leq$ 0.001	0	N/A	N/A	N/A	N/A
0.001 - 0.01	12,004	0.0596	0.0857	0.4101	0.0258
0.01 - 0.02	66,859	0.0789	0.0913	0.4167	0.0276
0.02 - 0.03	94,364	0.1115	0.0916	0.4201	0.0272
0.03 - 0.04	94,269	0.1548	0.0920	0.4200	0.0264
0.04 - 0.05	83,005	0.2015	0.0938	0.4199	0.0265
0.05 - 0.10	286,406	0.3434	0.1323	0.4236	0.0266
0.10 - 0.20	374,035	0.7421	0.2066	0.4419	0.0276
0.20 - 0.30	290,608	1.2774	0.2244	0.4721	0.0291
0.30 - 0.40	255,131	1.7747	0.2439	0.5027	0.0289
0.40 - 0.50	238,708	2.2529	0.2560	0.5347	0.0306
0.50 - 0.60	244,812	2.7030	0.2688	0.5667	0.0301
0.60 - 0.70	241,576	3.1296	0.2770	0.5999	0.0315
0.70 - 0.80	226,956	3.5528	0.2829	0.6350	0.0313
0.80 - 0.90	229,046	3.9600	0.2831	0.6688	0.0312
0.90 - 1.00	235,354	4.3380	0.2934	0.7032	0.0313
1.00 - 1.10	4,678	4.7390	0.2917	0.7380	0.0318
1.10 - 1.20	4,804	5.1176	0.2926	0.7725	0.0330
1.20 - 1.30	4,834	5.4931	0.2999	0.8078	0.0326
1.30 - 1.40	5,008	5.8682	0.2997	0.8410	0.0334
1.40 - 1.50	5,338	6.2376	0.3079	0.8757	0.0330
1.50 - 1.60	5,558	6.5948	0.3171	0.9116	0.0325
1.60 - 1.70	5,508	6.9412	0.3106	0.9457	0.0324
1.70 - 1.80	5,708	7.2902	0.3236	0.9810	0.0324
1.80 - 1.90	5,986	7.6334	0.3333	1.0144	0.0331
1.90 - 2.00	6,440	7.9841	0.3249	1.0483	0.0332
2.00 - 2.10	6,826	8.3060	0.3324	1.0822	0.0338
2.10 - 2.20	7,278	8.6380	0.3338	1.1183	0.0340
2.20 - 2.30	7,628	8.9539	0.3276	1.1515	0.0348
2.30 - 2.40	8,028	9.2669	0.3400	1.1856	0.0349
2.40 - 2.50	8,625	9.5876	0.3449	1.2214	0.0353
2.50 - 2.60	9,537	9.8854	0.3455	1.2562	0.0348
2.60 - 2.70	10,327	10.1846	0.3498	1.2913	0.0346
2.70 - 2.80	11,252	10.4831	0.3508	1.3263	0.0351
2.80 - 2.90	11,548	10.7891	0.3525	1.3605	0.0350
2.90 - 3.00	10,196	11.0766	0.3548	1.3950	0.0351
3.00 - 3.50	17,388	11.6992	0.5155	1.4686	0.0563
3.50 - 4.00	1,650	13.0601	0.4422	1.6321	0.0480
4.00 - 4.50	1	14.0314	0.0000	1.7500	0.0000a

Table A.4: Balance quality for matching and BOSS solutions for *data25c1k*.

	Mahalanobis Metric				Propensity Score				BOSS with $\mathcal{I}_{\text{Diff}}$ ($n_i = 20$)						
	KS		p-value		KS		p-value		KS		p-value		Objective		CPU
$\lvert\mathcal{P}\rvert$	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Best	LB	Time (s)
1	0.020	0.020	1.000	1.000	0.020	0.020	1.000	1.000	0.080	0.080	0.906	0.906	0	0	0.04
2	0.045	0.050	1.000	1.000	0.130	0.140	0.374	0.281	0.060	0.060	0.994	0.994	0	0	0.03
3	0.073	0.080	0.935	0.906	0.150	0.170	0.244	0.111	0.063	0.080	0.967	0.906	0	0	0.06
4	0.075	0.100	0.907	0.699	0.182	0.230	0.148	0.010	0.062	0.080	0.968	0.906	0	0	0.12
5	0.088	0.090	0.831	0.813	0.168	0.210	0.173	0.024	0.070	0.090	0.936	0.813	2	2.00	0.35
6	0.097	0.120	0.729	0.468	0.158	0.200	0.246	0.037	0.065	0.100	0.942	0.699	16	14.79	416.17
7	0.133	0.180	0.391	0.078	0.156	0.220	0.257	0.016	0.073	0.100	0.910	0.699	36	36.00	32.29
8	0.131	0.170	0.406	0.111	0.168	0.260	0.217	0.002	0.091	0.140	0.773	0.281	74	74.00	15.59
9	0.127	0.190	0.484	0.054	0.168	0.220	0.182	0.016	0.096	0.140	0.731	0.281	102	99.86	393.28
10	0.143	0.180	0.300	0.078	0.166	0.260	0.230	0.002	0.111	0.170	0.592	0.111	134	134.00	353.87
11	0.147	0.230	0.310	0.010	0.161	0.230	0.215	0.010	0.118	0.170	0.510	0.111	168	166.22	422.02
12	0.155	0.230	0.264	0.010	0.163	0.240	0.213	0.006	0.115	0.170	0.549	0.111	212	210.28	417.50
13	0.153	0.260	0.300	0.002	0.151	0.210	0.245	0.024	0.121	0.190	0.510	0.054	248	248.00	286.72
14	0.164	0.270	0.263	0.001	0.159	0.220	0.228	0.016	0.129	0.190	0.429	0.054	294	289.08	418.92
15	0.165	0.300	0.231	0.000	0.170	0.230	0.150	0.010	0.137	0.200	0.385	0.037	324	320.70	376.54
16	0.164	0.290	0.245	0.000	0.158	0.200	0.216	0.037	0.149	0.230	0.281	0.010	370	364.98	371.70
17	0.164	0.270	0.221	0.001	0.158	0.210	0.236	0.024	0.142	0.210	0.327	0.024	396	389.83	380.27
18	0.174	0.270	0.155	0.001	0.159	0.220	0.241	0.016	0.151	0.250	0.291	0.004	430	423.30	379.77
19	0.173	0.260	0.153	0.002	0.161	0.230	0.201	0.010	0.143	0.250	0.357	0.004	470	460.16	437.41
20	0.172	0.240	0.156	0.006	0.160	0.220	0.223	0.016	0.137	0.210	0.382	0.024	492	474.62	438.29
21	0.166	0.230	0.228	0.010	0.152	0.210	0.274	0.024	0.147	0.220	0.320	0.016	516	508.72	401.62
22	0.167	0.230	0.216	0.010	0.151	0.220	0.282	0.016	0.143	0.210	0.368	0.024	548	533.04	417.12
23	0.173	0.270	0.203	0.001	0.154	0.210	0.263	0.024	0.145	0.260	0.374	0.002	590	568.27	420.70
24	0.175	0.260	0.185	0.002	0.162	0.250	0.247	0.004	0.148	0.270	0.348	0.001	632	612.84	421.06
25	0.186	0.270	0.147	0.001	0.156	0.200	0.267	0.037	0.151	0.230	0.310	0.010	674	654.36	390.72

Table A.5: Balance quality for matching and BOSS solutions for *data25c5k*.

	Mahalanobis Metric				Propensity Score				BOSS with $\mathcal{I}_{\text{Diff}}$ ($n_i = 20$)						
	KS		p-value		KS		p-value		KS		p-value		Objective		CPU
$\lvert\mathcal{P}\rvert$	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Best	LB	Time (s)
1	0.010	0.010	1.000	1.000	0.010	0.010	1.000	1.000	0.070	0.070	0.967	0.967	0	0	0.10
2	0.035	0.040	1.000	1.000	0.115	0.130	0.533	0.367	0.050	0.050	1.000	1.000	0	0	0.12
3	0.043	0.050	1.000	1.000	0.167	0.180	0.133	0.078	0.060	0.070	0.987	0.967	0	0	0.16
4	0.052	0.060	0.998	0.994	0.172	0.240	0.202	0.006	0.060	0.060	0.994	0.994	0	0	0.53
5	0.084	0.120	0.836	0.468	0.136	0.170	0.348	0.111	0.060	0.070	0.985	0.967	0	0	0.73
6	0.092	0.110	0.768	0.581	0.145	0.180	0.279	0.078	0.062	0.090	0.966	0.813	0	0	1.91
7	0.083	0.130	0.838	0.367	0.157	0.180	0.180	0.078	0.071	0.090	0.931	0.813	0	0	95.26
8	0.094	0.130	0.754	0.367	0.185	0.210	0.096	0.024	0.065	0.100	0.952	0.699	4	0	424.26
9	0.091	0.150	0.782	0.211	0.172	0.230	0.166	0.010	0.078	0.120	0.875	0.468	8	0	411.39
10	0.114	0.150	0.530	0.211	0.163	0.210	0.179	0.024	0.072	0.110	0.909	0.581	16	0	404.13
11	0.129	0.170	0.412	0.111	0.155	0.210	0.244	0.024	0.067	0.100	0.952	0.699	24	9.48	412.71
12	0.122	0.170	0.487	0.111	0.185	0.240	0.143	0.006	0.068	0.100	0.939	0.699	46	26.17	391.33
13	0.119	0.180	0.499	0.078	0.185	0.240	0.113	0.006	0.078	0.130	0.880	0.367	68	49.86	412.87
14	0.127	0.160	0.428	0.155	0.169	0.210	0.150	0.024	0.083	0.170	0.848	0.111	92	79.04	411.03
15	0.129	0.180	0.436	0.078	0.161	0.240	0.213	0.006	0.085	0.150	0.813	0.211	126	108.58	408.30
16	0.141	0.190	0.307	0.054	0.166	0.300	0.244	0.000	0.094	0.200	0.765	0.037	156	139.61	402.25
17	0.144	0.190	0.293	0.054	0.174	0.230	0.157	0.010	0.089	0.150	0.790	0.211	190	167.25	398.37
18	0.143	0.250	0.344	0.004	0.162	0.210	0.198	0.024	0.094	0.160	0.755	0.155	228	203.53	407.44
19	0.141	0.200	0.355	0.037	0.158	0.210	0.218	0.024	0.096	0.160	0.713	0.155	254	232.84	401.71
20	0.141	0.220	0.369	0.016	0.158	0.240	0.228	0.006	0.101	0.170	0.682	0.111	270	247.69	383.66
21	0.144	0.240	0.361	0.006	0.162	0.270	0.242	0.001	0.098	0.150	0.704	0.211	290	265.58	376.95
22	0.149	0.220	0.306	0.016	0.155	0.210	0.255	0.024	0.105	0.160	0.642	0.155	314	281.18	410.41
23	0.148	0.220	0.305	0.016	0.162	0.230	0.210	0.010	0.104	0.170	0.639	0.111	362	325.44	404.68
24	0.143	0.200	0.336	0.037	0.162	0.240	0.237	0.006	0.105	0.180	0.636	0.078	382	350.86	400.27
25	0.160	0.230	0.258	0.010	0.156	0.240	0.250	0.006	0.113	0.170	0.559	0.111	420	385.84	404.11

Table A.6: Balance quality for matching and BOSS solutions for *data25c10k*.

	Mahalanobis Metric				Propensity Score				BOSS with $\mathcal{I}_{\text{Diff}}$ ($n_i = 20$)						
	KS		p-value		KS		p-value		KS		p-value		Objective		CPU
$\lvert\mathcal{P}\rvert$	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Best	LB	Time (s)
1	0.010	0.010	1.000	1.000	0.020	0.020	1.000	1.000	0.060	0.060	0.994	0.994	0	0	0.20
2	0.030	0.030	1.000	1.000	0.115	0.130	0.533	0.367	0.080	0.090	0.890	0.813	0	0	0.21
3	0.047	0.050	1.000	1.000	0.127	0.160	0.439	0.155	0.073	0.090	0.906	0.813	0	0	0.28
4	0.065	0.070	0.980	0.967	0.145	0.200	0.313	0.037	0.083	0.100	0.853	0.699	0	0	0.83
5	0.062	0.070	0.984	0.967	0.158	0.180	0.190	0.078	0.058	0.070	0.991	0.967	0	0	1.23
6	0.070	0.080	0.956	0.906	0.140	0.170	0.327	0.111	0.077	0.090	0.906	0.813	0	0	2.37
7	0.084	0.110	0.809	0.581	0.154	0.220	0.269	0.016	0.069	0.090	0.941	0.813	0	0	286.21
8	0.079	0.100	0.881	0.699	0.156	0.230	0.235	0.010	0.066	0.090	0.947	0.813	2	0	384.32
9	0.096	0.140	0.732	0.281	0.156	0.200	0.218	0.037	0.071	0.090	0.923	0.813	6	0	390.37
10	0.112	0.160	0.588	0.155	0.154	0.200	0.271	0.037	0.067	0.090	0.947	0.813	10	0	420.46
11	0.121	0.160	0.500	0.155	0.175	0.230	0.138	0.010	0.075	0.120	0.899	0.468	14	0	427.72
12	0.124	0.170	0.465	0.111	0.166	0.240	0.180	0.006	0.068	0.110	0.936	0.581	20	0	383.45
13	0.099	0.140	0.687	0.281	0.172	0.230	0.142	0.010	0.075	0.120	0.873	0.468	24	0	372.30
14	0.110	0.170	0.599	0.111	0.176	0.240	0.178	0.006	0.075	0.100	0.912	0.699	40	9.86	392.78
15	0.127	0.180	0.452	0.078	0.162	0.200	0.190	0.037	0.068	0.100	0.942	0.699	60	30.11	390.87
16	0.132	0.220	0.459	0.016	0.176	0.250	0.149	0.004	0.073	0.100	0.911	0.699	88	52.11	380.05
17	0.132	0.230	0.432	0.010	0.180	0.240	0.119	0.006	0.079	0.130	0.871	0.367	108	71.86	367.48
18	0.138	0.210	0.403	0.024	0.159	0.250	0.246	0.004	0.083	0.140	0.829	0.281	128	99.39	370.97
19	0.144	0.220	0.355	0.016	0.173	0.270	0.152	0.001	0.092	0.170	0.756	0.111	158	112.63	356.19
20	0.140	0.230	0.413	0.010	0.166	0.230	0.193	0.010	0.087	0.150	0.790	0.211	170	120.00	364.33
21	0.147	0.220	0.344	0.016	0.165	0.220	0.195	0.016	0.088	0.170	0.790	0.111	188	145.67	360.22
22	0.142	0.270	0.394	0.001	0.172	0.250	0.165	0.004	0.087	0.170	0.797	0.111	208	165.52	351.22
23	0.140	0.260	0.384	0.002	0.164	0.220	0.186	0.016	0.095	0.200	0.732	0.037	236	190.02	370.30
24	0.138	0.220	0.419	0.016	0.160	0.250	0.217	0.004	0.092	0.180	0.762	0.078	264	214.79	351.65
25	0.140	0.230	0.393	0.010	0.155	0.230	0.252	0.010	0.098	0.190	0.708	0.054	298	247.57	356.18

Table A.7: Estimated treatment effects for matching and BOSS solutions for *data25c1k*.

	Initial Estimate			Mahalanobis Metric			Propensity Score			BOSS with $\mathcal{I}_{\text{Diff}}$		
$\lvert\mathcal{P}\rvert$	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
1	1.550	3.886	3.886	0.132	0.125	0.125	0.128	0.122	0.122	0.153	0.093	0.093
2	3.290	2.981	4.986	0.052	0.184	0.081	0.262	1.542	3.028	0.296	0.263	0.041
3	2.624	2.314	4.586	0.233	0.106	0.061	−0.265	1.942	4.033	0.298	0.297	−0.042
4	3.917	3.607	8.545	0.128	0.203	0.296	0.588	2.547	7.613	0.162	0.149	0.020
5	4.646	5.184	11.951	0.411	−1.127	0.279	0.258	−11.386	4.481	0.136	0.861	0.564
6	3.964	8.183	10.988	0.673	0.404	1.584	−0.289	0.537	4.651	0.443	−0.891	−1.277
7	4.347	8.566	11.371	0.377	−1.321	2.200	−0.483	−10.233	1.933	0.678	1.571	1.035
8	7.061	10.237	13.041	1.952	5.297	1.781	−0.225	2.709	3.370	1.162	−6.137	2.130
9	7.257	13.202	16.007	2.544	3.569	2.009	0.720	−1.260	2.018	2.052	9.580	4.592
10	8.913	9.296	14.487	3.106	4.231	3.360	0.434	8.222	14.447	2.387	−0.252	4.708
11	9.546	7.352	12.543	3.200	4.949	3.033	0.808	7.130	8.483	3.218	2.726	0.382
12	9.332	8.104	13.295	3.410	7.309	2.785	−0.248	7.678	6.565	2.470	4.032	4.560
13	7.319	9.446	14.637	2.806	2.119	−0.946	0.168	−8.406	5.766	1.995	5.642	1.501
14	8.010	14.444	19.635	3.327	4.789	0.582	0.226	0.301	4.750	3.273	3.021	10.802
15	10.111	24.437	29.628	5.002	9.853	3.803	2.254	−0.786	7.436	3.256	11.242	18.529
16	11.262	25.589	30.780	5.930	11.577	6.879	1.243	−2.185	−1.078	5.406	21.183	20.352
17	11.919	21.150	31.073	5.392	10.727	7.917	1.647	−12.493	−0.098	5.131	8.399	17.243
18	14.506	23.736	33.660	7.787	12.810	15.029	2.369	−13.394	−0.250	6.983	11.221	16.241
19	14.977	28.966	34.186	6.499	13.716	11.992	1.556	−2.068	−1.739	7.456	6.139	20.481
20	14.915	29.494	34.714	8.204	13.348	9.287	2.748	−6.721	−2.108	5.260	11.465	14.557
21	14.549	29.127	34.348	7.339	11.544	12.420	0.625	−16.817	−7.906	5.000	17.261	14.475
22	14.168	35.606	40.827	6.648	12.125	15.270	1.385	7.781	7.953	6.071	8.995	22.773
23	14.908	36.347	41.567	7.213	7.501	15.246	3.090	−0.961	12.830	7.257	10.659	26.676
24	16.041	42.129	42.906	8.539	14.518	14.396	1.186	−20.264	−5.870	8.992	24.211	31.871
25	15.109	54.267	55.044	9.568	30.617	23.537	2.912	−0.420	−8.759	8.493	22.584	39.753

Table A.8: Estimated treatment effects for matching and BOSS solutions for *data25c5k*.

	Initial Estimate			Mahalanobis Metric			Propensity Score			BOSS with $\mathcal{I}_{\text{Diff}}$		
$\lvert\mathcal{P}\rvert$	A	B	C	A	B	C	A	B	C	A	B	C
1	1.406	3.968	3.968	0.124	0.119	0.119	−0.101	−0.106	−0.106	−0.040	−0.086	−0.086
2	3.023	2.747	5.306	0.104	0.070	0.069	0.078	1.291	3.608	−0.087	−0.293	0.277
3	2.389	2.113	4.901	−0.101	−0.049	−0.022	0.487	1.754	4.102	0.143	−0.014	−0.407
4	3.702	3.427	8.586	0.003	−0.578	−0.347	−0.177	3.626	8.501	0.301	0.140	0.240
5	4.542	6.341	12.068	0.559	2.721	0.405	0.502	−6.259	1.615	0.031	−0.080	−1.672
6	3.758	9.438	11.207	0.284	1.025	0.056	−0.545	−8.142	0.090	−0.021	−0.621	−0.835
7	4.191	9.870	11.639	0.258	1.236	0.605	0.377	2.112	3.637	0.054	−0.356	−3.309
8	6.936	11.559	13.328	1.014	3.491	1.784	0.120	−12.073	9.700	0.169	−0.337	0.470
9	7.155	14.682	16.451	0.496	−3.430	0.296	−1.154	−0.708	8.571	0.240	−0.185	1.358
10	8.733	11.426	14.410	1.343	−1.266	−0.529	0.388	7.473	8.254	0.123	1.471	−0.714
11	9.381	9.667	12.651	2.115	1.295	−0.055	−0.890	−1.282	2.527	0.874	−0.206	1.120
12	9.217	10.242	13.225	2.479	3.703	3.881	−0.103	4.814	6.206	0.027	−1.600	−0.995
13	7.280	11.533	14.517	1.814	4.388	0.373	0.977	−3.939	10.037	0.608	2.054	0.487
14	7.916	16.272	19.255	2.255	3.775	−1.029	−0.325	16.819	12.079	1.024	1.160	2.373
15	9.941	25.424	28.408	3.206	−0.227	−2.062	−1.420	−9.275	−7.876	1.530	7.320	6.999
16	11.107	26.590	29.574	3.775	2.108	4.401	−0.087	9.011	9.433	1.564	3.777	8.416
17	11.575	22.510	30.154	3.197	1.119	2.977	−0.979	−20.321	−13.234	2.824	1.896	5.565
18	14.067	25.003	32.647	4.631	4.293	3.453	−0.224	−4.423	−5.838	3.324	2.798	5.351
19	14.415	31.028	33.386	3.597	5.983	3.439	−0.943	−8.425	−3.355	3.489	−0.590	4.088
20	14.291	32.483	34.841	4.825	10.194	7.709	1.477	−26.917	−32.711	4.356	5.274	8.246
21	13.938	32.130	34.488	4.038	11.131	7.485	−2.331	2.018	5.543	3.657	9.047	11.662
22	13.733	38.615	40.973	5.339	7.628	7.355	1.250	2.289	3.157	3.727	5.985	11.303
23	14.521	39.403	41.761	6.121	11.337	11.137	−0.171	7.008	8.110	3.474	5.902	10.135
24	15.835	44.228	43.572	6.131	10.742	8.668	1.226	11.904	5.617	6.034	18.713	11.363
25	14.887	55.301	54.645	7.091	19.956	17.639	0.628	−19.830	−9.886	6.015	20.063	21.462

Table A.9: Estimated treatment effects for matching and BOSS solutions for *data25c10k*.

	Initial Estimate			Mahalanobis Metric			Propensity Score			BOSS with $\mathcal{I}_{\text{Diff}}$		
$\lvert\mathcal{P}\rvert$	A	B	C	A	B	C	A	B	C	A	B	C
1	1.248	3.939	3.939	−0.133	−0.133	−0.133	−0.117	−0.117	−0.117	−0.173	0.218	0.218
2	2.855	2.757	5.245	−0.224	−0.184	−0.214	−0.250	1.062	2.449	−0.088	−0.036	0.558
3	2.314	2.215	4.866	−0.192	−0.229	−0.267	−0.723	1.854	3.262	−0.140	0.052	0.753
4	3.615	3.517	8.551	0.047	0.015	−0.428	−0.471	0.805	8.087	−0.302	0.052	0.679
5	4.384	5.939	11.881	−0.096	0.774	0.626	−1.172	−14.864	−1.223	−0.300	−0.755	0.045
6	3.740	8.954	11.018	0.264	0.006	0.302	0.061	−9.008	−0.150	−0.292	−0.487	−0.448
7	4.178	9.392	11.457	0.218	−0.182	0.874	0.772	−15.024	5.093	0.077	−1.922	0.082
8	6.859	11.042	13.106	0.507	−1.423	0.132	−0.290	−0.540	−1.378	−0.121	−0.243	−0.785
9	7.106	14.239	16.303	1.149	0.576	−0.263	−0.160	1.316	3.857	0.140	0.059	−0.793
10	8.702	10.841	14.546	0.804	−3.278	0.390	0.519	12.247	5.437	−0.029	−0.952	−2.646
11	9.363	9.110	12.816	1.498	3.348	0.474	−0.206	8.145	6.042	0.128	0.025	−0.067
12	9.192	9.710	13.415	2.673	3.096	1.588	−0.561	−4.444	3.669	0.151	2.081	−0.181
13	7.226	11.020	14.726	1.805	1.778	1.813	−1.800	0.151	5.576	0.370	−0.175	−0.442
14	7.882	16.127	19.832	2.623	1.746	2.075	0.332	2.433	5.214	0.275	0.954	1.814
15	9.884	26.035	29.741	3.498	6.418	6.261	−0.843	2.177	19.126	0.577	1.474	3.074
16	10.987	27.138	30.844	3.378	−0.515	4.467	−2.001	−12.751	−14.263	1.439	−0.831	2.689
17	11.548	22.972	31.442	3.290	1.539	4.380	0.163	−5.000	2.838	1.647	0.631	3.722
18	14.064	25.488	33.958	4.486	3.646	6.230	0.103	1.474	−0.771	2.073	3.798	4.549
19	14.484	31.467	34.835	5.590	8.374	9.288	3.197	4.514	−6.714	3.498	6.703	8.527
20	14.454	28.902	32.270	4.779	9.157	10.164	1.293	−9.882	−11.849	2.992	0.908	7.782
21	14.053	28.501	31.868	5.532	16.397	16.240	1.798	4.924	2.617	2.369	0.628	5.118
22	13.872	35.035	38.403	5.853	14.142	13.681	−0.081	3.158	−1.903	1.921	4.978	10.023
23	14.679	35.842	39.210	5.341	5.400	4.498	0.504	−2.039	10.905	2.917	5.815	7.692
24	15.889	41.082	40.797	6.985	9.838	9.055	−1.618	3.602	−4.376	3.259	10.886	12.378
25	14.939	52.023	51.739	7.085	14.913	16.074	0.444	−6.257	−14.753	3.765	5.896	6.618

Table A.10: Covariate balance values for *nswcps* and *nswpsid* solutions.

		nswcps			*nswpsid*		
Objective	Covariate	DOM	KS	p-value	DOM	KS	p-value
$\mathcal{I}_{\text{DOM}}$	age	0.0000	0.2492	0.0000	1.5960	0.1818	0.0001
	education	0.0000	0.0606	0.6465	1.0202	0.2694	0.0000
	Black	0.0000	0.0000	1.0000	0.4108	0.4108	0.0000
	Hispanic	0.0000	0.0000	1.0000	0.0505	0.0505	0.8432
	married	0.0000	0.0000	1.0000	0.5253	0.5253	0.0000
	nodegree	0.0000	0.0000	1.0000	0.2694	0.2694	0.0000
	RE75	0.0009	0.0976	0.1178	0.0071	0.1683	0.0004
$\mathcal{I}_{\text{SDOM}}$	age	0.0000	0.2424	0.0000	2.9966	0.1818	0.0001
	education	0.0000	0.0774	0.3353	0.0000	0.0572	0.7154
	Black	0.0000	0.0000	1.0000	0.0404	0.0404	0.9686
	Hispanic	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
	married	0.0000	0.0000	1.0000	0.3266	0.3266	0.0000
	nodegree	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
	RE75	0.0009	0.1919	0.0000	5.3520×10^3	0.4848	0.0000
$\mathcal{I}_{\text{KS}}$	age	0.0236	0.0101	1.0000	5.7710	0.1953	0.0000
	education	0.0101	0.0034	1.0000	0.0640	0.0606	0.6465
	Black	0.0000	0.0000	1.0000	0.0471	0.0471	0.8962
	Hispanic	0.0000	0.0000	1.0000	0.0303	0.0303	0.9992
	married	0.0000	0.0000	1.0000	0.2828	0.2828	0.0000
	nodegree	0.0000	0.0000	1.0000	0.0101	0.0101	1.0000
	RE75	6.8930×10^1	0.0101	1.0000	4.2697×10^3	0.3266	0.0000
$\mathcal{I}_{\text{KS:max}}$	age	0.0168	0.0067	1.0000	6.0539	0.1886	0.0001
	education	0.0101	0.0067	1.0000	0.4983	0.1886	0.0001
	Black	0.0067	0.0067	1.0000	0.1886	0.1886	0.0001
	Hispanic	0.0067	0.0067	1.0000	0.0640	0.0640	0.5777
	married	0.0067	0.0067	1.0000	0.1886	0.1886	0.0001
	nodegree	0.0067	0.0067	1.0000	0.1886	0.1886	0.0001
	RE75	6.5656×10^1	0.0067	1.0000	2.5374×10^3	0.1886	0.0001
$\mathcal{I}_{\text{Diff}}$ (20)	age	0.0404	0.0539	0.7819	5.2155	0.2020	0.0000
	education	0.0000	0.0000	1.0000	0.1380	0.0404	0.9686
	Black	0.0000	0.0000	1.0000	0.0875	0.0875	0.2051
	Hispanic	0.0000	0.0000	1.0000	0.0067	0.0067	1.0000
	married	0.0000	0.0000	1.0000	0.3300	0.3300	0.0000
	nodegree	0.0000	0.0000	1.0000	0.0034	0.0034	1.0000
	RE75	2.7071×10^1	0.0370	0.9870	4.3406×10^3	0.4074	0.0000
$\mathcal{I}_{\text{Diff}^2}$ (20)	age	0.0168	0.0471	0.8962	7.7104	0.2896	0.0000
	education	0.0000	0.0000	1.0000	0.1481	0.1347	0.0091
	Black	0.0034	0.0034	1.0000	0.0875	0.0875	0.2051
	Hispanic	0.0000	0.0000	1.0000	0.0572	0.0572	0.7154
	married	0.0000	0.0000	1.0000	0.1549	0.1549	0.0016
	nodegree	0.0000	0.0000	1.0000	0.0774	0.0774	0.3353
	RE75	6.3631×10^1	0.0337	0.9960	3.8527×10^3	0.3805	0.0000

Table A.11: Covariate balance values for *nswre74cps* and *nswre74psid* solutions.

		nswre74cps			*nswre74psid*		
Objective	Covariate	DOM	KS	p-value	DOM	KS	p-value
$\mathcal{I}_{\text{DOM}}$	age	0.0000	0.2432	0.0000	3.4541	0.1838	0.0039
	education	0.0000	0.0811	0.5773	0.8054	0.2162	0.0004
	Black	0.0000	0.0000	1.0000	0.4703	0.4703	0.0000
	Hispanic	0.0000	0.0000	1.0000	0.0162	0.0162	1.0000
	married	0.0000	0.0000	1.0000	0.4811	0.4811	0.0000
	nodegree	0.0000	0.0000	1.0000	0.2162	0.2162	0.0004
	RE74	0.0050	0.1027	0.2834	0.0051	0.2162	0.0004
	RE75	0.0037	0.0973	0.3453	0.0037	0.0919	0.4155
$\mathcal{I}_{\text{SDOM}}$	age	0.0000	0.2595	0.0000	0.0000	0.1297	0.0889
	education	0.0000	0.0703	0.7510	0.0000	0.0649	0.8312
	Black	0.0000	0.0000	1.0000	0.0270	0.0270	1.0000
	Hispanic	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
	married	0.0000	0.0000	1.0000	0.1622	0.1622	0.0154
	nodegree	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
	RE74	0.0051	0.0811	0.5773	4.9880×10^3	0.6108	0.0000
	RE75	0.0037	0.1027	0.2834	4.8732×10^3	0.5730	0.0000
$\mathcal{I}_{\text{KS}}$	age	0.8432	0.0595	0.8992	5.0432	0.1784	0.0056
	education	0.0000	0.0000	1.0000	0.0270	0.0216	1.0000
	Black	0.0000	0.0000	1.0000	0.0811	0.0811	0.5773
	Hispanic	0.0000	0.0000	1.0000	0.0054	0.0054	1.0000
	married	0.0000	0.0000	1.0000	0.1351	0.1351	0.0682
	nodegree	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
	RE74	1.0720×10^2	0.0108	1.0000	5.2585×10^3	0.4432	0.0000
	RE75	3.5793×10^1	0.0108	1.0000	4.6300×10^3	0.3784	0.0000
$\mathcal{I}_{\text{KS:max}}$	age	0.2000	0.0216	1.0000	6.3351	0.2054	0.0008
	education	0.0595	0.0216	1.0000	0.6919	0.1784	0.0056
	Black	0.0162	0.0162	1.0000	0.2054	0.2054	0.0008
	Hispanic	0.0162	0.0162	1.0000	0.0108	0.0108	1.0000
	married	0.0162	0.0162	1.0000	0.2054	0.2054	0.0008
	nodegree	0.0216	0.0216	1.0000	0.1784	0.1784	0.0056
	RE74	6.4619×10^1	0.0216	1.0000	2.8268×10^3	0.2054	0.0008
	RE75	1.7347×10^2	0.0216	1.0000	3.1029×10^3	0.2054	0.0008
$\mathcal{I}_{\text{Diff}}$ (20)	age	0.2541	0.0811	0.5773	4.2811	0.1568	0.0212
	education	0.0000	0.0000	1.0000	0.1676	0.0541	0.9498
	Black	0.0000	0.0000	1.0000	0.0919	0.0919	0.4155
	Hispanic	0.0000	0.0000	1.0000	0.0108	0.0108	1.0000
	married	0.0000	0.0000	1.0000	0.2541	0.2541	0.0000
	nodegree	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
	RE74	4.2474×10^1	0.2378	0.0001	3.4996×10^3	0.4811	0.0000
	RE75	5.1548×10^1	0.0973	0.3453	3.0987×10^3	0.4162	0.0000
$\mathcal{I}_{\text{Diff}^2}$ (20)	age	0.2703	0.0541	0.9498	5.4108	0.2270	0.0001
	education	0.0000	0.0108	1.0000	0.1946	0.1135	0.1843
	Black	0.0054	0.0054	1.0000	0.0811	0.0811	0.5773
	Hispanic	0.0000	0.0000	1.0000	0.0324	0.0324	1.0000
	married	0.0054	0.0054	1.0000	0.1243	0.1243	0.1146
	nodegree	0.0000	0.0000	1.0000	0.0378	0.0378	0.9994
	RE74	1.8373×10^2	0.2432	0.0000	3.4044×10^3	0.4541	0.0000
	RE75	2.8913×10^2	0.1027	0.2834	3.4247×10^3	0.4000	0.0000