Essays on Matching and Weighting for Causal Inference in Observational Studies

María de los Angeles Resa Juárez

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Essays on Matching and Weighting for Causal Inference in Observational Studies

María de los Angeles Resa Juárez

This thesis consists of three papers on matching and weighting methods for causal inference. The first paper conducts a Monte Carlo simulation study to evaluate the performance of multivariate matching methods that select a subset of treatment and control observations. The matching methods studied are the widely used nearest neighbor matching with propensity score calipers, and the more recently proposed methods, optimal matching of an optimally chosen subset and optimal cardinality matching. The main findings are: (i) covariate balance, as measured by differences in means, variance ratios, Kolmogorov-Smirnov distances, and cross-match test statistics, is better with cardinality matching since by construction it satisfies balance requirements; (ii) for given levels of covariate balance, the matched samples are larger with cardinality matching than with the other methods; (iii) in terms of covariate distances, optimal subset matching performs best; (iv) treatment effect estimates from cardinality matching have lower RMSEs, provided strong requirements for balance, specifically, fine balance, or strength-k balance, plus close mean balance. In standard practice, a matched sample is considered to be balanced if the absolute differences in means of the covariates across treatment groups are smaller than 0.1 standard deviations. However, the simulation results suggest that stronger forms of balance should be pursued in order to remove systematic biases due to observed covariates when a difference in means treatment effect estimator is used. In particular, if the true outcome model is additive then marginal distributions should be balanced, and

if the true outcome model is additive with interactions then low-dimensional joints should be balanced.

The second paper focuses on longitudinal studies, where marginal structural models (MSMs) are widely used to estimate the effect of time-dependent treatments in the presence of time-dependent confounders. Under a sequential ignorability assumption, MSMs yield unbiased treatment effect estimates by weighting each observation by the inverse of the probability of their observed treatment sequence given their history of observed covariates. However, these probabilities are typically estimated by fitting a propensity score model, and the resulting weights can fail to adjust for observed covariates due to model misspecification. Also, these weights tend to yield very unstable estimates if the predicted probabilities of treatment are very close to zero, which is often the case in practice. To address both of these problems, instead of modeling the probabilities of treatment, a design-based approach is taken and weights of minimum variance that adjust for the covariates across all possible treatment histories are directly found. For this, the role of weighting in longitudinal studies of treatment effects is analyzed, and a convex optimization problem that can be solved efficiently is defined. Unlike standard methods, this approach makes evident to the investigator the limitations imposed by the data when estimating causal effects without extrapolating. A simulation study shows that this approach outperforms standard methods, providing less biased and more precise estimates of time-varying treatment effects in a variety of settings. The proposed method is used on Chilean educational data to estimate the cumulative effect of attending a private subsidized school, as opposed to a public school, on students' university admission tests scores.

The third paper is centered on observational studies with multi-valued treatments. Generalizing methods for matching and stratifying to accommodate multi-valued treatments has proven to be a complex task. A natural way to address confounding in this case is by weighting the observations, typically by the inverse probability of treatment weights (IPTW). As in the MSMs case, these weights can be highly

variable and produce unstable estimates due to extreme weights. In addition, model misspecification, small sample sizes, and truncation of extreme weights can cause the weights to fail to adjust appropriately for observed confounders. The conditions the weights need to satisfy in order to provide close to unbiased treatment effect estimates with a reduced variability are determined and the convex optimization problem that can be solved in polynomial time to obtain them is defined. A simulation study with different settings is conducted to compare the proposed weighting scheme to IPTW, including generalized propensity score estimation methods that also consider explicitly the covariate balance problem in the probability estimation process. The applicability of the methods to continuous treatments is also tested. The results show that directly targeting balance with the weights, instead of focusing on estimating treatment assignment probabilities, provides the best results in terms of bias and root mean square error of the treatment effect estimator. The effects of the intensity level of the 2010 Chilean earthquake on posttraumatic stress disorder are estimated using the proposed methodology.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all I would like to express my deepest gratitude to Professor José Zubizarreta for being such a patient, supportive, and encouraging advisor. Sharing his experience and expertise in this field with me gave me the necessary tools to walk this road. Without his guidance and dedication the completition of this dissertation would have not been possible. I sincerely appreciate all his efforts to help me grow as a researcher.

I would also like to thank my committee members, Professors David Madigan, Michael Sobel, Arian Maleki, and Elizabeth Tipton for their time and attention dedicated to this work. I'm particularly grateful to Professor Madigan and Professor Sobel for introducing me to the exciting world of causal inference and inspiring me to follow this path, along with their constant support. On an unrelated matter, I was very touched by their thoughtfulness and care when they visited me at the hospital; I will never forget that.

On the same note, I need to thank Professor Richard Davis for saving my life in Copenhagen!

I am thankful to all my Professors at Columbia University, especially my first year instructors for providing me with the basics necessary to be able to call myself a statistician. Listening to Professor Philip Protter teach probability theory became one of my greatest joys that year. My gratitude also goes to Dood Kalicharan and Anthony Cruz for all their prompt help during these years.

I am deeply grateful to my fellow classmates and friends in the Ph.D. program, who made my experience here much more pleasant. I was very lucky to share my

Ph.D. years with this particular group of people. I want to give a special recognition to Jingjing, who pushed me and inspired me to work harder from the first cheat sheet to the very last line of this dissertation, and my officemates, for listening to me every time they did.

Finally, to my parents Argel Juárez and Gabriel Resa, thank you so much for believing in me and providing me your unconditional support on every step of the way. Thank you for being here or anywhere in the world every time I've needed you, and thank you for all the love, encouragement, and wisdom that you have shared with me.

A mis papás y mi abuelita, quien partió en la espera de este momento.

Chapter 1

Introduction

"Correlation does not imply causation," yet determining causality is the main goal of a considerable amount of studies. The gold standard to perform this type of inference are randomized studies, however, for many reasons, this is not always an available option and in these cases researchers can only rely on observational data. The problem with this data is that, since the treatment assignment is not random, there could be systematic differences between the people who received one treatment and the other, which may result in intrinsically different outcome values and consequently biased estimates of treatment effects. Causal inference is concerned with establishing assumptions, that combined with some methods, will allow researchers to correctly infer causality. The most common assumptions made include *no unmeasured confounders*, where it is assumed that there are no unobserved variables that predict the treatment assignment while at the same time affect the outcome variable, and *positivity*, where it is assumed that every subject for which inference is being made should have a positive probability of receiving any treatment. For the first assumption to be satisfied, it is important to collect enough covariates so that the treatment assignment appears to be random given the observed covariates.

One of the greatest advances in this field was the introduction of the propensity score (Rosenbaum and Rubin, 1983), which is a scalar that represents the probability

of a subject to get treated given his observed covariates. For its attractive ability to balance covariates, researchers have focused on developing methods based on the propensity score, like matching, weighting, and stratification, and on proposing different ways to estimate it. Despite its useful properties, implementing methods that rely on it imply engaging in an iterative process in which first the propensity score has to be estimated, then the particular method is applied, then balance is checked, and depending on the balance attained, the propensity score might need to be estimated again and the rest of the process repeated until a satisfactory level of balance is achieved. In light of balance being the main goal, researchers have started developing methods that directly balance the covariates, avoiding excessive iteration.

Given the rise of these balancing methods, a natural question that needs to be answered is, how should we balance covariates in order to obtain unbiased estimates of the treatment effects? The second chapter of this thesis provides some guidance with respect to this matter, using matching. The goal of matching is to find samples of treated and control units with similar or balanced observed covariate distributions (Stuart, 2010). When the distributions of the observed covariates are substantially different between treated and control samples before matching, as it tends to happen in observational studies, matching methods need to remove observations that do not have an appropriate counterpart on the other treatment group. Commonly used approaches to do this are to trim the sample before matching, discarding the observations outside the overlap region between the propensity score ranges of the two groups (see, e.g., Dehejia and Wahba (1999)) or to restrict the possible matches to be within a propensity score caliper. Neither of these approaches discards observations in an optimal way, but there are recently proposed methods that incorporate the selection of observations into their design and offer optimal solutions in some sense. One of these methods is optimal matching of an optimally chosen subset (Rosenbaum, 2012). This method pursues two specific goals: to minimize the total sum of covariate distances between the matched pairs, and to match as many pairs of treated and control

units as possible. Another method is optimal cardinality matching (Zubizarreta et al., 2014). This method finds the largest matched sample of treated and control units for which the covariate distributions are balanced as required by the researcher. Here we conduct a simulation study to evaluate the performance of the latter two methods and the widely used nearest neighbor matching with propensity score calipers. Our comparisons make emphasis on the covariate balance achieved by the methods, while considering also how covariates should be balanced. In particular, we are interested in determining when mean balance is sufficient to remove biases when estimating the treatment effect with a simple difference in means and when stronger forms of covariate balance should be pursued.

On the third chapter of this thesis we contemplate the setting where the observational study is longitudinal. Many applications in the health and social sciences are interested in evaluating the effect of a treatment or exposure that is assigned multiple times. The main particularity of this type of data is the possible presence of time-dependent confounders. Typical methods for covariate adjustments, such as matching or regression methods, fail to provide unbiased estimates of treatment effects when this is the case. For this reason, a popular alternative to these methods in the presence of time-dependent confounders is marginal structural models (MSMs). Under a sequential ignorability assumption, MSMs yield unbiased treatment effect estimates by weighting each observation by the inverse of the probability of their observed treatment sequence given their history of observed covariates. However, these probabilities are typically estimated by fitting a propensity score model, and the resulting weights can fail to adjust for observed covariates due to model misspecification. Also, these weights tend to yield very unstable estimates if the predicted probabilities of treatment are very close to zero, which is often the case in practice. To address both of these problems, instead of modeling the probabilities of treatment, we take a design-based approach and directly find the weights of minimum variance that adjust for the covariates across all possible treatment histories. We first analyze

the role of weighting in longitudinal studies of treatment effects and then we define the convex optimization problem that incorporates that information. We conduct a simulation study to compare the proposed methodology to standard methods used in this setting and present an application in which we analyze the effect that the total number of years a student spends in a private voucher school in Chile during secondary education has on the scores he obtains on the University Selection Test, as opposed to spending those years in a public school.

The fourth chapter of this thesis considers again the cross-sectional setting, in which the treatment is received only once, but focusing on the case where the treatment can take multiple values. While the binary case has been studied for many years and common practices are well known and widely used across different disciplines, the methods for multi-valued treatments have not reached that level of popularity. Many methods proposed are based on the generalized propensity score (Imbens, 2000), however, generalizing matching and stratification methods to this setting has proved to be a difficult task. One method that extends naturally to multi-valued treatments is weighting observations to address confounding, which has typically be done by using the inverse probability of treatment weights (IPTW). As with MSMs, the presence of extreme weights can induce large variability on the treatment effect estimator. Additionally, if the probabilities are not modeled correctly, which is more likely when there are more than two treatments, or if the weights are truncated to reduce variability of the estimates, the balancing properties of these weights will not hold. Considering these drawbacks of the common weighting estimator, we determine conditions that the weights need to satisfy to provide close to unbiased treatment effect estimates that also have low variability. Based on these conditions, we describe the convex optimization problem that needs to be solved to obtain the desired weights and conduct a simulation study to compare the proposed method with other IPTW methods for multi-valued treatments that have been recently proposed (Fong et al., 2017; McCaffrey et al., 2013, 2004). Finally, we present an application of the method where we

estimate the effects that different levels of intensity of an earthquake have on post-traumatic stress disorder, using data from Chile's 2009 national socioeconomic survey and the 2010 post-earthquake survey that was carried out after the February 2010 earthquake.

Chapter 2

Evaluation of Subset Matching Methods and Forms of Covariate Balance

2.1 Introduction

In observational studies, matching methods are often used to approximate the ideal randomized experiment that would have been conducted if controlled experimentation had been feasible (Cochran and Rubin, 1973). In these settings, the ultimate goal of matching is to free the comparisons of the outcomes in the treatment and control groups from biases due to differences in their observed covariates (Cochran et al., 1983). To achieve this goal, matching methods find samples of treated and control units with similar or balanced observed covariate distributions (Stuart, 2010). Ideally, these matched samples will include all the available treated and control observations. However, if the distributions of the observed covariates are substantially different between treated and control samples before matching, as it tends to happen in observational studies, then including all the observations will result in poor covariate balance.

To address this problem, matching methods typically rely on the propensity score to remove observations that do not have an appropriate counterpart on the other treatment group.[1] One way to do this is to trim the sample before matching, discarding the observations outside the overlap region between the propensity score ranges of the two groups (see, e.g., Dehejia and Wahba (1999)). Another commonly used approach is to restrict the possible matches to be within a prespecified propensity score value, called caliper. While these methods provide a broad solution to the problem, they are additions to methods for which discarding observations was not incorporated in the design of the problem.

Two recently proposed methods incorporate the selection of observations into their design and offer optimal solutions in some sense. The first method is optimal matching of an optimally chosen subset, or optimal subset matching for short (Rosenbaum, 2012). This method pursues two specific goals: to minimize the total sum of covariate distances between the matched pairs, and to match as many pairs of treated and control units as possible. Often these two goals are at odds with each other, and the trade-off between them is regulated by means of a tuning parameter discussed and interpreted by Rosenbaum (2012).

The second method is optimal cardinality matching or, simply, cardinality matching (Zubizarreta et al., 2014). This method finds the largest matched sample of treated and control units for which the covariate distributions are balanced as required by the researcher. With this method different forms of covariate balance can be achieved by design, including balance of means and higher order moments (Zubizarreta, 2012), balance of marginal distributions or fine balance (Rosenbaum et al., 2007), and balance of low dimensional joints or strength-k balance (Hsu et al., 2015).

[1] As argued in Stuart and Rubin (2007), dropping units results in a smaller sample size and, it may appear, in a larger variance of the estimator, however better covariate balance may actually improve the efficiency of the estimator (see section 18.2 of Snedecor and Cochran (1980)). Also, as argued in Rosenbaum (2005b), dropping units may help to increase unit homogeneity which in turn can reduce sensitivity to biases due to unobserved covariates.

While there have been a number of simulation studies comparing different matching methods, most of these studies are centered on methods that rely on the propensity score, and to our knowledge none of them have examined these two optimal matching methods. For this reason perhaps is that these methods are not used much for research in medicine and other related disciplines, where they can play an important role because of their optimality guarantees in terms of covariate distances and balance. Furthermore, while there have been studies evaluating algorithms, calipers, distances, and structures in matching, few have placed emphasis on how covariates should be balanced. In practice, a widely used rule of thumb is to consider a matched sample to be balanced if the absolute standardized differences in means of the covariates across treatment groups are smaller than 0.1. However, we are not aware of a systematic study of this rule. Also, there are reasons to think that the balance criteria should be data- and estimator-specific. In particular, we are interested in studying questions like, in which cases would balancing only the means of the covariates be sufficient to remove biases when estimating the treatment effect with a simple difference in means? When should stronger forms of covariate balance, such as balance of marginal or low-dimensional joint distributions, be pursued? Although we believe these questions should ultimately be addressed with formal statistical theory, in this chapter we conduct a Monte Carlo simulation study to provide initial answers. For this, in Section 2 we describe with more detail the matching methods that will be compared. In Section 3, we explain the specifics of the simulation study, and in Section 4 we show and discuss the simulation results. In Section 5 we explore additional considerations regarding the results. Finally, in Section 6 we conclude with a summary and remarks.

2.2 Matching methods

2.2.1 Greedy matching

Propensity score matching (Rosenbaum and Rubin, 1983) is, in all likelihood, the most frequently used matching method in medicine and related sciences. Accordingly, it has been studied extensively in the past (e.g., Austin (2009, 2011, 2014); Dehejia and Wahba (2002); Gu and Rosenbaum (1993)), and thus, we considered it in our simulation study as a benchmark. A commonly used algorithm for propensity score matching is greedy or first-best nearest neighbor matching (Rubin, 1973). In its most basic form, this algorithm first sorts the treated units in terms of the estimated propensity score (from highest to lowest, lowest to highest, or randomly), and then matches the first treated unit to the closest available control, making it no longer available for matching for the rest of the treated units. Closeness here may be defined as the propensity score distance, which is the absolute difference of the estimated propensity scores for two units (Rosenbaum and Rubin, 1985b). To avoid poor matches, a caliper may be added to the propensity score distance so that a control unit is matched to a treated unit only if it is within the caliper, and treated units for which there are no controls available within the caliper are discarded. In this manner, only a subset of the treated units are matched to controls. Simulation studies by Austin (2009, 2011, 2014) suggest that the best way to implement greedy nearest neighbor matching is by matching the treated units in random order and without replacement, using a linear propensity score distance, and imposing a caliper of 0.2 times the standard deviation of the linear propensity score. We followed these recommendations in our implementation of this method.

2.2.2 Optimal subset matching

Greedy matching does not, in general, minimize the total sum of distances between matched units (see chapter 10 of Rosenbaum (2002) for an example). In contrast,

optimal matching (Rosenbaum, 1989) finds the assignment of treated and control units that minimizes this global distance. In observational studies, the optimal matching problem can be cast as an assignment problem (Burkard et al., 2009), a special case of the minimum cost flow problem (Ahuja et al., 1993), that in turn can be written as a linear program (Bertsimas and Tsitsiklis, 1997). While it is possible to solve the assignment problem using the simplex algorithm, there exist specific algorithms such as the Hungarian algorithm (Kuhn, 1955) or the auction algorithm (Bertsekas, 1981) that can better exploit the assignment problem's particular structure. This is important because these algorithms can be solved "quickly," or more formally, in polynomial time; that is, in a number of arithmetic operations that is characterized by a polynomial function of certain parameters of the problem (as opposed to, say, an exponential function (Papadimitriou, 1994)).

Typically, optimal matching uses all the available treated units and does not have the flexibility to discard some treated units in the case where they are very hard to match. A recently proposed matching method that selects a subset of units is optimal subset matching (Rosenbaum, 2012). This is an elegant solution to the optimal subset matching problem, which consists of formulating the problem as an assignment problem on a modified matrix of distances between treated and control units; see Rosenbaum (2012) for details. As mentioned in the introduction, this method pursues two specific goals: to minimize the total sum of covariate distances between the matched units, and to match as many pairs of treated and control units as possible. Often these two goals are at odds with each other, but this trade-off is regulated by means of a prespecified covariate distance threshold $\tilde{\delta}$. For a given $\tilde{\delta}$, this method "prefers more treated subjects if their average increase in distance is less than $\tilde{\delta}$ and prefers fewer treated subjects if their average increase in distance is more than $\tilde{\delta}$, so $\tilde{\delta}$ is the distance at which there is indifference" (Rosenbaum, 2012).[2]

[2]This is achieved by solving the assignment problem with an augmented distance matrix, in which a certain number of columns, all with the value $\tilde{\delta}$, are added to the original treated-control distance

In our simulation study we used optimal subset matching with the prespecified covariate distance threshold $\tilde{\delta}$ set up at the 20% quantile of the distance matrix before adding the caliper. For comparability with greedy matching we used the linear propensity score distance. We implemented this method using the `R` functions in the supplementary materials of Rosenbaum (2012), which in turn uses the `R` function `pairmatch` in the `optmatch` package (Hansen and Klopfer, 2006).

2.2.3 Cardinality matching

The third matching method we studied was cardinality matching (Zubizarreta et al., 2014). Cardinality matching is an optimal matching method that maximizes the cardinality or size of the matched sample subject to constraints on covariate balance. As described in Zubizarreta et al. (2014), the covariate balance constraints can be quite general. In their weakest form, they can require the means to be balanced (see Zubizarreta (2012) for details), but they can also require other forms of distributional balance such as fine balance (Rosenbaum et al., 2007) and strength-k balance (Hsu et al., 2015). To be precise, fine balance is a constraint on a nominal covariate that forces its marginal distributions to be identical, but does not require treated and control units to be matched within each of the categories of the nominal variable, as in exact matching (see Chapter 10 of Rosenbaum (2010) for details). Strength-k balancing provides a stronger form of balance by forcing low dimensional joint distributions of a nominal covariate to be identical; specifically, out of K nominal covariates, each of the $\binom{K}{l}$ possible interactions of covariates is finely balanced for all $l \leq k$, so the joint distributions of each of the $\sum_{l=1}^{k} \binom{K}{l}$ combinations of covariates is perfectly balanced. Clearly, strength-k balancing implies fine balance on each of the K nominal covariates. By imposing these constraints on covariate balance, cardinality

matrix. These columns can be thought of as additional controls for which the distance to every treated unit is $\tilde{\delta}$, and the pairs that involve any of these controls will not be used. The number of columns added is the number of treated subjects that will be allowed to be discarded.

matching directly balances the original covariates and does not require estimation of the propensity score or another summary of the covariates. Of course, stronger requirements on covariate balance tend to yield smaller matched samples.[3]

From a computational standpoint, cardinality matching solves a linear integer programming problem, and, while a polynomial time algorithm to solve the cardinality matching problem has not been found, many instances of this problem can be solved in time comparable to the user time of the two previous methods. In addition, if finding an exact solution for an instance of the cardinality matching problem is too demanding, then it is possible to find an approximate solution by solving a relaxation of the integer programming problem (Zubizarreta and Kilcioglu, 2016). This approximate solution may violate to some extent the covariate balancing constraints, but it will be found in polynomial time. In some settings, the balancing constraints can be formulated to require tighter balance than needed in view of the approximation.[4]

In our simulation study, we evaluated six types of covariate balance constraints with cardinality matching: (i) the widely used rule of absolute standardized differ-

[3] In cardinality matching, finding the largest balanced matched sample is followed by re-pairing the treated and control units that constitute the matched sample to minimize their total sum of covariate distances. It is known that the heterogeneity of the matched pair differences in outcomes affects the sensitivity of results to biases due to unobserved covariates Rosenbaum (2005b). In cardinality matching, if the covariates used in the re-pairing are predictive of the outcome, this will reduce heterogeneity within matched groups and therefore sensitivity to biases due to unobserved covariates (see Baiocchi (2011) for a general approach using the prognostic score, Hansen (2007)).

[4] At the present, exact and approximate solutions to the cardinality matching problem can be found with the package `designmatch` for `R` Zubizarreta (2012); Zubizarreta and Kilcioglu (2016). This package includes functions for the construction of matched samples that are balanced by design which can be used, among others, for matching in observational studies with treated and control units, as in this study, but also in settings with cases and controls (where the propensity score typically cannot be estimated) and under weaker identification assumptions with instrumental variables (e.g., Yang et al. (2014)) and discontinuity designs Keele et al. (2015).

ences in means, $|\hat{d}|$,[5] smaller than 0.1, $|\hat{d}| < 0.1$; (ii) $|\hat{d}| < 0.01$; (iii) $|\hat{d}| < 0.001$; (iv) fine balance by discretizing continuous covariates into 10 categories; (v) both fine balance and $|\hat{d}| < 0.01$; (vi) and both fine balance and $|\hat{d}| < 0.001$. In some instances, we considered strength-2 balancing using 10 categories to discretize the continuous variables to balance their marginal distributions and 5 categories to balance the two-dimensional joint distributions together with $|\hat{d}| < 0.001$. We studied strength-2 balancing with 10 categories to balance both marginal and two-dimensional joints in the additional considerations section.

2.3 Simulation study design

2.3.1 Data generating mechanisms

Our simulation study design combines elements of the designs in Gu and Rosenbaum (1993) and Austin (2009). Each simulated dataset consisted of $n = 250(1 + r)$ observations, where r is the number of controls available for each treated unit, so there were 250 treated units and $250r$ controls. The variables in each dataset consisted of three continuous outcomes and eight covariates, four of them continuous and four dichotomous. More specifically, the dichotomous covariates were two rare and two common Bernoulli random variables, all of them conditionally independent from each other and from the continuous covariates given the treatment assignment indicator Z. The continuous covariates followed a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}_t$ and mean vector $\boldsymbol{\mu}_t$ for the treated group, and $\mathbf{\Sigma}_c$ and $\boldsymbol{\mu}_c$ for the control group. The means of the eight covariates were selected in such a way that the true standardized differences in means, given by $d = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_t^2 + \sigma_c^2}{2}}}$ for the Normal variables and $d = \frac{p_t - p_c}{\sqrt{\frac{p_t(1-p_t) + p_c(1-p_c)}{2}}}$ for the Bernoulli random variables (Rosenbaum and Rubin,

[5] $\hat{d}$ refers to the standardized difference in means in the matched sample, while d refers to the true standardized difference in means in the data generating mechanism.

1985b), were either 0.2 or 0.5.

For the covariance matrices of the Normal random variables, we examined three different scenarios of increasing complexity:

- Scenario 1, same variances in the treated and control groups with independent covariates in both groups;
- Scenario 2, different variances between treated and control groups with independent covariates in both groups;
- Scenario 3, different variances between treated and control groups with independent covariates in the control group and correlated covariates in the treatment group.

We generated the outcome Y from $Y = f(\boldsymbol{X}) + Z + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 4)$, and so here the true treatment effect is one. We considered three different forms of f:

- Linear, $f(x) = 3.5x_1 + 4.5x_3 + 1.5x_5 + 2.5x_7$;
- Additive, $f(x) = 3.5x_1 + 4.5x_3 + 1.5x_5 + 2.5x_7 + 2.5sign(x_1)|x_1|^{1/2} + 5.5x_3^2$;
- Additive with interactions, $f(x) = 3.5x_1 + 4.5x_3 + 1.5x_5 + 2.5x_7 + 2.5sign(x_1)|x_1|^{1/2} + 5.5x_3^2 + 2.5x_3x_7 - 4.5\left|x_1x_3^3\right|$.

The reader may notice that in each of these models only half of the covariates were included. This mimics the fact that in practice the investigator does not know exactly which covariates affect the outcome. As a consequence, in order to avoid discarding a potentially relevant covariate, it is often preferable to match for more rather than fewer covariates (see section 6.2 of Stuart (2010)).

To assess the performance of the matching methods when there is different competition among the treated units for controls, we considered two values for the number of controls available per treated unit, $r = 1, 2$.[6] We generated a total of 1000 replications of each dataset and matched them with each method.

[6]We decided to fix the number of treated and control subjects, as in Gu and Rosenbaum (1993),

2.3.2 Performance measures

We compared the performance of the matching methods on the grounds of four criteria. First, we examined their ability to balance covariates. For this we used absolute standardized differences in means, variance ratios, Kolmogorov-Smirnov distances, and cross-match test statistics. More specifically, it has been said that absolute standardized differences in means smaller than 0.1 is evidence of covariate balance (Normand et al., 2001), but the fact that two covariates have distributions with similar first moments does not imply that differences in other moments do not introduce bias in the effect estimates (see sections 2.4 and 2.5; a related formal argument is proposition 4.2 of Zubizarreta (2015)). It is for this reason that we also calculated the variance ratio and the Kolmogorov-Smirnov distance between the empirical distributions of the continuous covariates across the treatment and control groups. Furthermore, to evaluate balance of the joint distributions we also calculated the standardized cross-match statistic (Rosenbaum, 2005a). Other interesting multivariate balance measures are discussed in Franklin et al. (2014); Hansen and Bowers (2008); Iacus et al. (2012); Imai et al. (2008), but the cross-match statistic yields an exact, distribution free test for comparing two high-dimensional distributions, with close connections to the procedures used to construct a matched sample and a nice interpretation in terms of the propensity score (Heller et al., 2010). To calculate this statistic we used the `R` package `crossmatch` (Heller et al., 2010) with the same propensity score distance used to match with nearest neighbor and optimal subset matching.

The second criterion by which we compared the matching methods was sample size. For a given level of covariate balance (that is, for a given level of bias reduction), the largest matched sample should be preferred since this directly translates into more efficient estimates (Haviland et al., 2007).

---

instead of randomly assigning to treatment each unit by means of another Bernoulli random variable, as in Austin (2009), to have a fixed number of maximum possible pairs and be able to better compare the number of units matched across simulations.

Next we used covariate distances. Though covariate distances play a secondary role in matching (they are often instrumental to achieve covariate balance), these distances can have an intrinsic importance depending on the statistical methods to be used after matching. If these methods explicitly use the matched-pair structure, then reducing covariate distances between matched pairs will increase efficiency and also reduce sensitivity to biases due to unobserved covariates (Rosenbaum, 2005b). To make cardinality matching comparable to the other matching methods in terms of distances, we re-paired the treated and control units initially selected by cardinality matching using optimal matching with a propensity score distance (as discussed in Zubizarreta et al. (2014), re-pairing is actually the second stage of cardinality matching).

Finally, we compared the matching methods in terms of simple treatment effect estimates. We estimated the treatment effects using the difference in outcome means between the treatment and control groups and recorded the estimated bias and RMSE of these estimates. When looking at these values, it is important to keep in mind that, even if a matching method managed to perfectly balance the covariate distributions and consequently completely eliminate the bias due to observed covariates, the RMSE of the estimator would still be $\sqrt{\frac{2Var(\varepsilon)}{m}} = 2\sqrt{\frac{2}{m}}$, where m is the number of pairs selected when matching. This also explicitly shows why, for a fixed level of covariate balance, a larger sample size is preferable.

2.4 Results

2.4.1 Covariate balance

In this section we compare the matching methods in terms of covariate balance. Figure 2.1 summarizes the distribution of absolute standardized differences in means across simulated datasets in scenario 2 with $r = 1$ for four representative covariates. Other scenarios, group sizes ratios, and covariates presented a similar pattern. In

the boxplots, we can see that the differences in means after matching with nearest neighbor and optimal subset matching varied to a fair extent from data set to data set, in many occasions being even over the 0.1 standard deviations threshold. On the other hand, with cardinality matching the absolute standardized differences in means were always smaller than 0.1, 0.01 or 0.001, as designed by the investigator. The evident performance difference takes place because with cardinality matching it is possible to finely-tune covariate balance adjustments, while with the other matching methods, covariate balance is attained indirectly, by matching on the propensity score and hoping that this will result in the desired covariate balance.

Figure 2.1: Boxplots of absolute standardized differences in means in scenario 2 when $r = 1$.

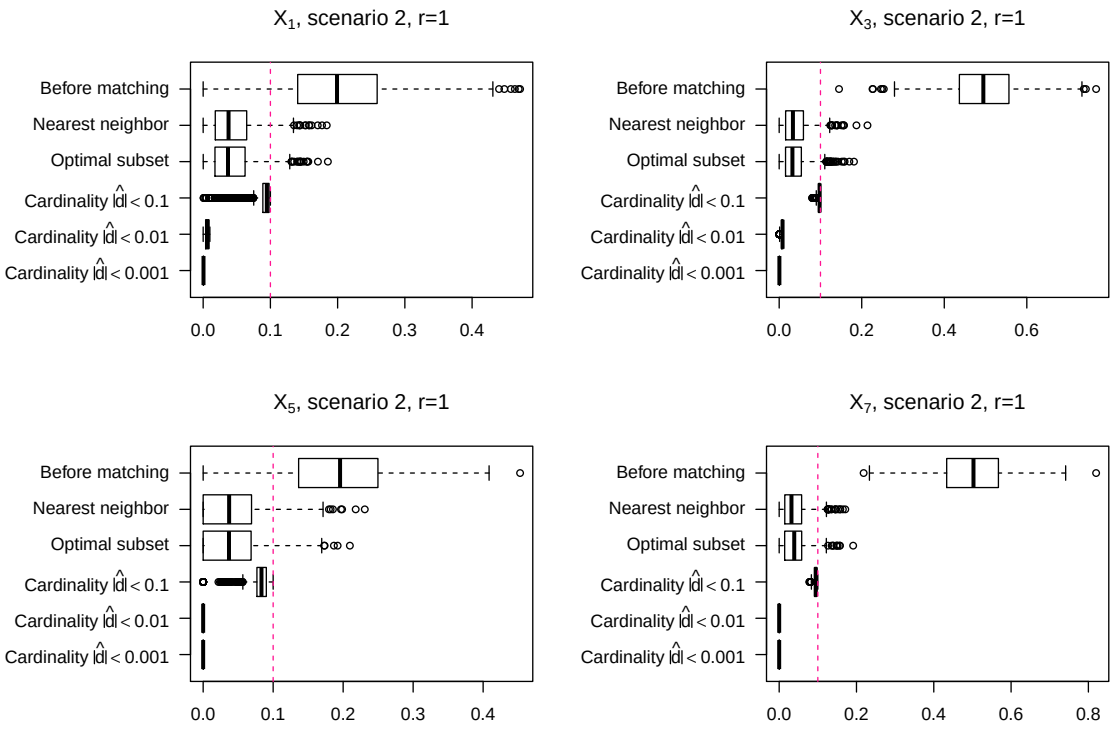


Table 2.1 shows summaries of balance for marginal distributions. In particular, it

shows the averages of the variance ratios and Kolmogorov-Smirnov distances for each type of continuous covariate in every scenario when $r = 1$. The rest of the continuous covariates presented a similar pattern, as well as the case of $r = 2$. We observe that cardinality matching with fine balance constraints presented the variance ratios closest to 1 and the smallest K-S distance values in all the scenarios. This suggests that these are the best methods to balance the complete distribution of the covariates. This advantage was even more evident when the distributions before matching differed in more than just the mean (scenarios 2 and 3). When the variance ratio was originally one (scenario 1), all the resulting matches ended up with a variance ratio close to one. However, when the variances were different before matching, the only methods that corrected this were the ones involving fine balance; the rest of them kept the same ratio as before matching.

In short, the best marginal balance was obtained with cardinality matching when we combined fine balance and tight mean balance constraints. This resulted in a matching method that gave the best results on all marginal balance measures, in practically every scenario and control sample size. It is important to note that only requiring $|\hat{d}| < 0.1$ was the worst method in all cases and measures. In some sense, this calls into question the rule of absolute standardized differences in means smaller than 0.1. Optimal subset was better than nearest neighbor in attaining marginal balance, but both of them were worse than the other methods that required fine balance.

Table 2.2 compares the values observed for the cross-match statistic in all the simulation settings studied. For this measure we can distinguish three performance groups in all settings: distance driven methods, mean balance only methods, and fine balance methods. Fine balance methods produced the best results in terms of multivariate covariate balance, despite the fact that they did not explicitly include multidimensional balance constraints (such as strength-k balancing). The next best group was the mean balance only group, with values that were at least closer to zero

Table 2.1: Balance measures: variance ratio and Kolmogorov-Smirnov distance, $r = 1$

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		$\frac{\sigma_t}{\sigma_c}$	K-S	$\frac{\sigma_t}{\sigma_c}$	K-S	$\frac{\sigma_t}{\sigma_c}$	K-S
X_1	Before matching	1.0021	0.1196	0.5066	0.1636	0.5050	0.1636
	Nearest neighbor	1.0065	0.0834	0.5128	0.1297	0.4999	0.1316
	Optimal subset	1.0039	0.0823	0.5125	0.1291	0.5002	0.1315
	Cardinality $\lvert\hat{d}\rvert < 0.1$	1.0054	0.0886	0.5095	0.1354	0.4967	0.1350
	Cardinality $\lvert\hat{d}\rvert < 0.01$	1.0056	0.0721	0.5110	0.1214	0.4937	0.1249
	Cardinality $\lvert\hat{d}\rvert < 0.001$	1.0052	0.0724	0.5103	0.1216	0.4930	0.1251
	Cardinality fine balance	1.0015	0.0574	0.9549	0.0654	0.9533	0.0655
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	1.0008	0.0546	0.9543	0.0614	0.9528	0.0620
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	1.0013	0.0542	0.9542	0.0613	0.9521	0.0615
X_3	Before matching	1.0065	0.2292	0.5020	0.2602	0.5035	0.2622
	Nearest neighbor	1.0175	0.0845	0.5450	0.1237	0.5136	0.1286
	Optimal subset	1.0142	0.0833	0.5434	0.1234	0.5138	0.1281
	Cardinality $\lvert\hat{d}\rvert < 0.1$	1.0063	0.0931	0.5200	0.1361	0.5033	0.1416
	Cardinality $\lvert\hat{d}\rvert < 0.01$	1.0110	0.0727	0.5263	0.1179	0.5036	0.1233
	Cardinality $\lvert\hat{d}\rvert < 0.001$	1.0113	0.0724	0.5261	0.1180	0.5029	0.1233
	Cardinality fine balance	0.9999	0.0595	0.9472	0.0684	0.9522	0.0670
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	0.9999	0.0553	0.9471	0.0638	0.9512	0.0619
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.9982	0.0550	0.9481	0.0635	0.9506	0.0618

Note: The values presented are the averages of the variance ratios and Kolmogorov-Smirnov distances observed with each method in each of the 1000 repetitions. Case $r = 2$ presents a similar pattern.

than before matching. The worst group was the distance driven methods group. This group had the largest deviations from zero, in some occasions even larger than before matching, even though these methods focused on the same distance used to compute the cross-match statistic. Within this group, nearest neighbor was the best.

The ideal way to obtain k-dimensional balance would be to directly add this kind of constraints in the optimization problem, that is, to perform strength-k balancing. However, with 8 covariates, trying to obtain strength-2 balance for all the observed covariates using fine categories for the continuous covariates can be very demanding and leave us with very few observations in this simulation setting. In a posterior section we discuss some results obtained when we perform strength-2 balancing on relevant covariates, in a situation in which we have more information on the relationship between the covariates and the outcome.

2.4.2 Sample sizes

Table 2.3 presents the average number of pairs matched with each method on every scenario and control sample size. Naturally, when there were more controls available per treated subject, every method was able to obtain a larger sample size. In our example, the number of matched pairs when $r = 2$ was between approximately 30% to 45% larger than the number of pairs matched when $r = 1$. The comparison among matching methods was similar for both values of r.

Cardinality matching with $|\hat{d}| < 0.1$ was always the matching method that produced the largest sample sizes, followed by the other mean balance methods, decreasing the sample size as the mean balance was tightened. For the other matching methods, the comparison depended on the type of imbalances in the covariate distributions before matching. When they differed only by their means (scenario 1), cardinality matching with fine balance methods kept more or a similar amount of observations than the distance driven methods. This changed when the observed covariates were more imbalanced before matching (scenario 2 and 3). In these cases,

Table 2.2: Cross-match test statistic

	Matching method	Scenario 1	Scenario 2	Scenario 3
$r = 1$	Before matching	−3.8466	−3.8693	−3.5923
	Nearest neighbor	2.4963	2.1126	2.2819
	Optimal subset	4.1896	4.2104	4.1977
	Cardinality $\lvert\hat{d}\rvert < 0.1$	−1.2758	−1.2974	−1.1795
	Cardinality $\lvert\hat{d}\rvert < 0.01$	−1.5434	−1.4843	−1.4886
	Cardinality $\lvert\hat{d}\rvert < 0.001$	−1.5388	−1.4903	−1.4661
	Cardinality fine balance	−0.1898	−0.0293	0.0150
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	−0.1922	0.0131	−0.0058
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	−0.1665	−0.0352	−0.0086
$r = 2$	Before matching	−4.3518	−4.1156	−3.9883
	Nearest neighbor	4.4745	3.9573	4.3405
	Optimal subset	5.4422	5.0051	5.2760
	Cardinality $\lvert\hat{d}\rvert < 0.1$	−2.0475	−1.1871	−1.6029
	Cardinality $\lvert\hat{d}\rvert < 0.01$	−2.3067	−1.7510	−2.0163
	Cardinality $\lvert\hat{d}\rvert < 0.001$	−2.2674	−1.7854	−2.0042
	Cardinality fine balance	−0.6880	−0.1225	−0.2265
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	−0.6204	−0.1633	−0.2100
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	−0.6092	−0.1872	−0.2291

Note: The values presented are the averages of the standardized cross-match statistics observed with each method in each of the 1,000 repetitions. The values are standardized by subtracting the expectation under the null hypothesis of equal distribution in SE units.

the latter methods had larger sample sizes (between 8% and 20% higher) than the fine balance methods. This happened because the fine balance methods had to discard more observations in order to correct important distributional imbalances which, as studied on the covariance balance section, the distance driven methods failed to accomplish. Again, in all scenarios, the sample size for cardinality matching with fine balance was smaller as tighter mean balance was required. Also, in every case, optimal subset matched more pairs than nearest neighbor, although the difference was small (less than 2%).

As we have seen, direct comparison of the sizes of the subsets selected by each method is rather unfair. A more valuable comparison would be to observe the sample sizes at a certain covariate balance level. This balance level cannot be set beforehand with nearest neighbor and optimal subset matching. Nevertheless, we can notice that cardinality matching with $|\hat{d}| < 0.01$ and $|\hat{d}| < 0.001$ produced a similar but better covariance balance than those methods and they provided sample sizes at least 10% larger than those observed with nearest neighbor or optimal subset matching. This suggests that at certain level of covariate balance, the sizes of the subsets that result from cardinality matching are larger than with the other methods.

Table 2.3: Sample size

	Matching method	Scenario 1	Scenario 2	Scenario 3
$r = 1$	Nearest neighbor	145.39	144.92	148.84
	Optimal subset	148.16	147.71	151.44
	Cardinality $\lvert\hat{d}\rvert < 0.1$	182.88	182.80	184.89
	Cardinality $\lvert\hat{d}\rvert < 0.01$	164.50	164.46	167.82
	Cardinality $\lvert\hat{d}\rvert < 0.001$	163.44	163.44	167.01
	Cardinality fine balance	149.72	126.40	128.25
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	148.79	125.40	127.31
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance*	147.30	124.36	126.29
$r = 2$	Nearest neighbor	192.49	200.13	199.95
	Optimal subset	195.80	203.42	203.12
	Cardinality $\lvert\hat{d}\rvert < 0.1$	241.83	243.95	244.12
	Cardinality $\lvert\hat{d}\rvert < 0.01$	222.51	226.83	228.41
	Cardinality $\lvert\hat{d}\rvert < 0.001$*	220.57	224.90	226.73
	Cardinality fine balance	208.11	185.24	186.89
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	207.12	184.13	185.83
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance *	205.71	182.93	184.69

Note: The values presented are the averages of the sample sizes observed with each method in each of the 1,000 repetitions.

* For some of the simulated datasets, these methods did not find a solution in the allotted time. The averages do not include those cases.

2.4.3 Distances

Table 2.4 shows the average propensity score distance between matched pairs obtained by each method. As expected, in terms of distances, the method that performed the best was optimal subset matching, followed by nearest neighbor matching. Optimal subset matching is designed precisely to minimize the global distance between matched pairs while matching as many pairs as possible. In contrast, cardinality matching maximizes the number of observations that satisfy some balance constraints. This explains the large difference between nearest neighbor matching

Table 2.4: Average propensity score distance

	Matching method	Scenario 1	Scenario 2	Scenario 3
$r = 1$	Nearest neighbor	0.0377	0.0405	0.0367
	Optimal subset	0.0241	0.0242	0.0223
	Cardinality $\lvert\hat{d}\rvert < 0.1$	0.3547	0.3606	0.3258
	Cardinality $\lvert\hat{d}\rvert < 0.01$	0.2444	0.2494	0.2315
	Cardinality $\lvert\hat{d}\rvert < 0.001$	0.2401	0.2450	0.2281
	Cardinality fine balance	0.1417	0.1250	0.1154
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	0.1361	0.1208	0.1129
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.1305	0.1171	0.1114
$r = 2$	Nearest neighbor	0.0238	0.0251	0.0227
	Optimal subset	0.0192	0.0196	0.0180
	Cardinality $\lvert\hat{d}\rvert < 0.1$	0.4341	0.3442	0.3612
	Cardinality $\lvert\hat{d}\rvert < 0.01$	0.3196	0.2570	0.2826
	Cardinality $\lvert\hat{d}\rvert < 0.001$	0.3117	0.2509	0.2764
	Cardinality fine balance	0.2064	0.1133	0.1268
	Cardinality $\lvert\hat{d}\rvert < 0.01$ + fine balance	0.2023	0.1099	0.8898
	Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.1963	0.1082	0.1241

Note: The values presented are the averages of the average propensity score distance observed with each method in each of the 1000 repetitions.

and cardinality matching with fine balance and $|\hat{d}| < 0.001$ constraints, which was the next best method. The average distance for the latter was between 3 to 8 times as large as nearest neighbor, depending on the scenario and the control sample size. The average distance increased as the distributional constraints to which the cardinality matching method was subject to were relaxed, resulting in the following order from best to worst: fine balance and $|\hat{d}| < 0.01$, fine balance, $|\hat{d}| < 0.001$, $|\hat{d}| < 0.01$, and, lastly, $|\hat{d}| < 0.1$. The results hold for all the scenarios and for both control sample sizes analyzed.

2.4.4 Treatment effect estimates

We now study which method provided the best treatment effect estimates for each form of relationship between the outcome and the observed covariates. Table 2.5 shows the bias and root mean square error (RMSE) for each method when $r = 1$. Results for $r = 2$ were slightly better for all the matching methods than when $r = 1$; however, the comparison among methods was similar.

When the outcome is a linear combination of the observed covariates, balancing the means of the covariates suffices to remove systematic biases in the treatment effect estimates. This means that the smaller the absolute standardized differences in means are, the less biased is the treatment effect estimate. For instance, if we compare Table 2.5 with Figure 2.1, we notice that the methods with lower and less variable absolute standardized differences in means were the ones that showed a closer and more precise estimate. In this case, cardinality matching with $|\hat{d}| < 0.001$ had the best results, followed by $|\hat{d}| < 0.01$. In some instances, these two showed a greater bias than those obtained by distance driven methods; nonetheless, their variability was lower and consequently the RMSE was, in some cases, half the RMSE obtained with the distance driven methods. In general, when the outcome is linear, it appears to be advisable to balance the means as closely as possible, as long as enough observations are matched. Note that in our setting we were able to tighten the balance constraints from $|\hat{d}| < 0.01$ to $|\hat{d}| < 0.001$ with practically no reduction in sample size.

When the outcome is not a linear combination of the observed covariates, which will most likely be the case in practice, it appears that a stronger form of covariate balance is needed. In this case, balancing only the means is not enough to remove systematic biases, unless the distributions of the covariates only differ by their means before matching (like in scenario 1) and the rest of the distributions remain balanced after matching. In our simulation we can see that even though the bias of cardinality matching with only $|\hat{d}| < 0.001$ was among the smallest, its RMSE was more than 2.5 times the RMSE of cardinality matching that in addition to $|\hat{d}| < 0.001$ required

Table 2.5: Treatment effect estimation performance, $r = 1$

Matching method	Scenario 1		Scenario 2		Scenario 3	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Linear						
Before matching	3.6210	3.6649	4.2477	4.2979	4.2741	4.3333
Nearest neighbor	0.0416	0.3601	−0.0058	0.4177	0.0506	0.4090
Optimal subset	0.0322	0.3414	0.0072	0.4053	0.0563	0.3887
Cardinality $\lvert\hat{d}\rvert < 0.1$	0.8790	0.9086	1.0269	1.0537	0.9627	1.0062
Cardinality $\lvert\hat{d}\rvert < 0.01$	0.0531	0.2299	0.0535	0.2244	0.0599	0.2302
Cardinality $\lvert\hat{d}\rvert < 0.001$	0.0063	0.2233	−0.0042	0.2164	0.0115	0.2178
Cardinality fine balance	0.0857	0.2762	0.1074	0.3203	0.1018	0.3241
Card. $\lvert\hat{d}\rvert < 0.01$ + fine balance	0.0248	0.2340	0.0219	0.2521	0.0285	0.2649
Card. $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0038	0.2163	0.0010	0.2220	0.0094	0.2343
Card. $\lvert\hat{d}\rvert < 0.01$ + strength-2*	0.2002	0.3887	0.1958	0.4347	0.1803	0.4469
Additive						
Before matching	5.4213	5.5432	1.3089	1.9695	1.3327	2.0644
Nearest neighbor	0.0777	1.0287	−4.4930	4.7318	−4.8763	5.0937
Optimal subset	0.0479	0.9863	−4.4849	4.7225	−4.8469	5.0582
Cardinality $\lvert\hat{d}\rvert < 0.1$	1.3111	1.5378	−3.0667	3.3076	−3.4446	3.6783
Cardinality $\lvert\hat{d}\rvert < 0.01$	0.0790	0.8214	−4.5232	4.6963	−4.8355	5.0088
Cardinality $\lvert\hat{d}\rvert < 0.001$	0.0051	0.8082	−4.6309	4.8011	−4.9353	5.1064
Cardinality fine balance	0.1312	0.4070	−0.1644	0.5608	−0.1507	0.5695
Card. $\lvert\hat{d}\rvert < 0.01$ + fine balance	0.0323	0.3334	−0.3241	0.5544	−0.2963	0.5501
Card. $\lvert\hat{d}\rvert < 0.001$ + fine balance	−0.0123	0.2981	−0.3635	0.5173	−0.3419	0.5217
Card. $\lvert\hat{d}\rvert < 0.01$ + strength-2*	0.2807	0.5872	−0.0950	0.7035	−0.1518	0.7392
Additive with interactions						
Before matching	3.9501	4.1854	15.8023	16.2399	14.7405	15.2330
Nearest neighbor	0.0966	1.4362	10.7838	11.7255	10.7544	11.6885
Optimal subset	0.0829	1.4655	10.7368	11.6525	10.7005	11.6010
Cardinality $\lvert\hat{d}\rvert < 0.1$	1.1222	1.6874	11.5650	12.2497	11.2694	11.9708
Cardinality $\lvert\hat{d}\rvert < 0.01$	0.1338	1.2818	10.5861	11.3906	10.5498	11.3333
Cardinality $\lvert\hat{d}\rvert < 0.001$	0.0822	1.2882	10.5860	11.4048	10.5906	11.4058
Cardinality fine balance	0.1022	1.0648	0.6323	2.3448	−0.8431	2.4767
Card. $\lvert\hat{d}\rvert < 0.01$ + fine balance	0.0246	1.0653	0.6041	2.3249	−0.8564	2.5347
Card. $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0219	0.9926	0.5875	2.1095	−0.8707	2.3721
Card. $\lvert\hat{d}\rvert < 0.01$ + strength-2*	0.2649	0.7434	0.7491	1.6680	0.5846	1.4790

Note: The bias and RMSE observed with each method with $r = 2$ present a similar pattern as shown in this table.

* Matches were obtained using the approximation algorithm in `designmatch` Zubizarreta and Kilcioglu (2016).

fine balance. This occurred because the magnitude of the differences of the complete matched distributions varied from dataset to dataset, which translated into differences in the nonlinear part of the outcome and therefore in the treatment effect estimate.

When the covariate distributions differed in more ways than in their means (scenarios 2 and 3), the only methods that corrected these differences were the ones involving fine balance and strength-2 balance. It is clear from Table 2.5 that these distributional differences can affect to a great extent the bias and RMSE of the treatment effect estimates. This depends on the importance and level of nonlinearity of the nonlinear part of the outcome, but in our simulation we observed RMSEs 5 to 10 times higher on the rest of the methods compared to cardinality matching with fine balance and strength-2.

In general, when the outcome is an additive function without interactions, the ideal way of matching is with fine balance plus tight mean balance constraints, whereas when the outcome is an additive function with interactions, it is best to match with strength-k balancing, again with tight mean balance. As we mentioned earlier, stronger requirements on covariate balance tend to yield smaller matched samples, and thus, matching with strength-k can be very demanding in terms of sample size. Here, the strength-2 balancing considered for all the possible pairs of the eight covariates resulted in matched samples that ranged in size between 35% and 40% of the original number of treated units available when $r = 1$ and between 57% and 62% when $r = 2$. However, if we had used 10 categories to balance both marginal and two-dimensional joint distributions, sample sizes could have been as small as 5 observations, which would be unacceptable. In practice, the decision to require stronger forms of covariate balance must be weighed against the resulting sample size.

In summary, if the outcome is known to be a linear function of the observed covariates, balancing the means would be enough. A good way to do this is to use cardinality matching to directly balance the means as closely as the data allows without sacrificing too much sample size. However, it is quite unlikely that this will be the

case, so it is important to always consider the possibility that there are nonlinearities. Not doing this can result in extremely biased estimations. Thus, if the treatment effect is estimated with no further covariate adjustments, it is recommended to use cardinality matching with both fine balance and tight mean balance constraints if the outcome model is believed to be additive, and with strength-k balancing and tight mean balance constraints if the model is conjectured to be additive with interactions.

2.5 Additional considerations

2.5.1 Incorporating prior knowledge about the relationship between the observed covariates and the outcome

In this section we analyze the effect estimates when the investigator has different levels of knowledge about the true functional form that relates the outcome and the covariates, and this knowledge is included in the matching process by imposing different forms of covariate balance. Here we focus on cardinality matching since this method gives the investigator the flexibility to directly obtain different forms of covariate balance. From the previous simulation results section, for each functional form we selected the form of covariate balance that provided the best treatment effect estimates and compared these estimates with the ones obtained when different levels of knowledge were incorporated into the matching process. These levels of knowledge will be referred to as "no," "weak," "strong," and "complete knowledge" about the true function that relates the covariates to the outcome.

Specifically, no knowledge refers to the situation where the investigator does not know which covariates affect the outcome nor in which way (this is, linearly, nonlinearly without interactions, or nonlinearly allowing for interactions between the covariates). Weak knowledge refers to the case where the investigator knows which covariates are relevant in terms of the outcome, but does not know if they relate to the outcome linearly, nonlinearly without interactions, or nonlinearly with interactions.

Strong knowledge is the case where one knows not only which covariates are relevant, but also, to some extent, the role they play in the outcome model. For example, this would be the case if one knew which covariates were relevant in a non-linear way and which covariates interacted between each other, but without specifically knowing the non-linear functional form or how they interacted. Finally, complete knowledge refers to the utopic situation where the investigator knows the true functional form of the model that relates the covariates and the outcome. This case is only considered as a benchmark to evaluate the rest of the matches. The forms of covariate balance considered for each of these situations are described in Table 2.6.

Table 2.7 and Table 2.8 summarize the main results in terms of treatment effect estimation and sample sizes. The results show that, when the outcome is linear, having more previous knowledge does not necessarily translate into better treatment effect estimation. However, having more information allows the distributional balance constraints to be relaxed as needed, obtaining, as a consequence, larger sample sizes and therefore more precise estimates. When the outcome is additive, again, the RMSEs of the treatment effect estimate are close between each other when there is no complete knowledge. Nonetheless, in our case, these RMSEs were between 50% larger to more than twice the size obtained when there was complete knowledge. This is due to the fact that continuous covariates were approximately balanced by balancing their categorized versions instead of being perfectly balanced. When there are interactions, it is possible to appreciate the great advantage of reducing the number of covariates we match on, especially when only strength-2 balancing of a coarsely discretized version of the continuous covariates is feasible, which was the case in this simulation setting. Being able to perform strength-2 balancing with finer categories on the variables that interact, significantly improved the estimation of the treatment effect compared to the performance when only fine balance or a coarse version of strength-2 was obtained.

In short, if we want to minimize our estimates' dependence on model assumptions, the most appropriate choice is to match with the strongest constraints, that is, to

Table 2.6: Forms of covariate balance considered for different levels of knowledge about the true outcome model

	Outcome model		
	Linear	Additive	Additive with interactions
Knowledge			
No	$\lvert\hat{d}\rvert < 0.001$ for $X_1, X_2, ..., X_8$; $\lvert\hat{d}\rvert < 0.001$ + f.b. for $X_1, X_2, ..., X_8$	$\lvert\hat{d}\rvert < 0.001$ + f.b. for $X_1, X_2, ..., X_8$	$\lvert\hat{d}\rvert < 0.001$ + f.b. for $X_1, X_2, ..., X_8$; $\lvert\hat{d}\rvert < 0.001$ + s.-2 for $X_1, X_2, ..., X_8$*
Weak	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7; $\lvert\hat{d}\rvert < 0.001$ + f.b. for X_1, X_3, X_5, X_7; $\lvert\hat{d}\rvert < 0.001$ + s.-2 for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001$ + f.b. for X_1, X_3, X_5, X_7; $\lvert\hat{d}\rvert < 0.001$ + s.-2 for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001$ + f.b. for X_1, X_3, X_5, X_7; $\lvert\hat{d}\rvert < 0.001$ + s.-2 for X_1, X_3, X_5, X_7
Strong	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001$ for X_5, X_7 + $\lvert\hat{d}\rvert < 0.001$ and f.b. for X_1, X_3	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7 + f.b. for X_1, X_3 + s.-2 for X_1, X_3 and X_3, X_7
Complete	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7, $sign(X_1)\lvert X_1\rvert^{1/2}, X_3^2$	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7, $sign(X_1)\lvert X_1\rvert^{1/2}, X_3^2, X_3X_7, \lvert X_1X_3^3\rvert$

Note: $\lvert\hat{d}\rvert < 0.001$ denotes absolute differences in means smaller than 0.001 standard deviations; f.b. stands for fine balance, and s.-2 for strength-2 balancing. In this way, for example, when there is no knowledge about the form of the true outcome model, and this model is linear (top left corner of the table), we imposed two forms of covariate balance: absolute differences in means smaller than 0.001 standard deviations for all the covariates (this is, $\lvert\hat{d}\rvert < 0.001$ for $X_1, X_2, ..., X_8$), and both absolute differences in means smaller than 0.001 standard deviations and fine balance for all the covariates ($\lvert\hat{d}\rvert < 0.001$ + f.b. for $X_1, X_2, ..., X_8$).

* Using the approximation algorithm in `designmatch` Zubizarreta and Kilcioglu (2016) with 5 categories to balance two-dimensional joints.

Table 2.7: Treatment effect estimation performance at different levels of previous knowledge, $r = 1$

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
Linear	Before matching	3.6210	3.6649	4.2477	4.2979	4.2741	4.3333
	All: Cardinality $\lvert\hat{d}\rvert < 0.001$	0.0063	0.2233	−0.0042	0.2164	0.0115	0.2178
	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0038	0.2163	0.0010	0.2220	0.0094	0.2343
	W/S/C: Cardinality $\lvert\hat{d}\rvert < 0.001$	0.0061	0.2082	−0.0132	0.2047	0.0051	0.2053
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0052	0.2107	−0.0088	0.2178	−0.0015	0.2232
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	0.0072	0.2360	−0.0065	0.2383	−0.0010	0.2450
Additive	Before matching	5.4213	5.5432	1.3089	1.9695	1.3327	2.0644
	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	−0.0123	0.2981	−0.3635	0.5173	−0.3419	0.5217
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	−0.0021	0.2909	−0.3376	0.5060	−0.3423	0.5108
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	0.0061	0.3125	−0.3046	0.5066	−0.3029	0.4993
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	−0.0016	0.2967	−0.3461	0.5124	−0.3417	0.5130
	Complete: $\lvert\hat{d}\rvert < 0.001$	0.0046	0.2063	−0.0156	0.2099	0.0022	0.2122
Additive with	Before matching	3.9501	4.1854	15.8023	16.2399	14.7405	15.233
interactions	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0219	0.9926	0.5875	2.1095	−0.8707	2.3721
	All: Cardinality $\lvert\hat{d}\rvert < 0.01$ + strength-2*	0.2649	0.7434	0.7491	1.6680	0.5846	1.4790
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0208	0.9303	0.5592	1.9611	−0.6769	2.1902
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	−0.0001	0.3863	0.3396	0.8675	0.2883	0.9219
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	0.0169	0.9345	0.5498	1.9155	−0.6471	2.1735
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	0.0006	0.3867	0.3586	0.8644	0.3445	0.9356
	Complete: $\lvert\hat{d}\rvert < 0.001$	0.0040	0.2099	−0.0049	0.2130	0.0051	0.2124

Note: The bias and RMSE observed with each method and level of knowledge with $r = 2$ present a similar pattern as the shown in this table.

* Using the approximation algorithm in `designmatch` Zubizarreta and Kilcioglu (2016) with 5 categories to balance two-dimensional joints.

Table 2.8: Sample sizes when matching with different levels of previous knowledge

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		$r=1$	$r=2$	$r=1$	$r=2$	$r=1$	$r=2$
Linear	All: Cardinality $\lvert\hat{d}\rvert < 0.001$	163.44	220.57	163.44	224.90	167.01	226.73
	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance all	147.30	205.71	124.36	182.93	126.29	184.69
	W/S/C: Cardinality $\lvert\hat{d}\rvert < 0.001$	186.22	244.32	186.24	244.73	186.98	244.59
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	144.38	199.46	136.01	199.14	133.86	197.40
Additive	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	147.30	205.71	124.36	182.93	126.29	184.69
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	144.38	199.46	136.01	199.14	133.86	197.40
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Complete: Cardinality $\lvert\hat{d}\rvert < 0.001$	185.75	243.41	178.56	242.80	179.16	242.68
Additive with interactions	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	147.30	205.71	124.36	182.93	126.29	184.69
	All: Cardinality $\lvert\hat{d}\rvert < 0.01$ + strength-2*	100.20	154.00	89.69	147.60	86.68	142.80
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	144.38	199.46	136.01	199.14	133.86	197.40
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.35	235.09	164.35	230.51	164.52	230.91
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	151.84	208.50	142.63	207.06	140.51	205.10
	Complete: Cardinality $\lvert\hat{d}\rvert < 0.001$	184.46	237.99	176.05	237.37	177.31	238.03

Note: Some values are repeated because some matchings are the same for different functional forms.

* Using the approximation algorithm in `designmatch` Zubizarreta and Kilcioglu (2016) with 5 categories to balance two-dimensional joints.

require tight mean balance, fine balance, and strength-2 balance. Doing this will result in smaller sample sizes, but the consequences of this will depend on the original number of observations available. However, if relevant covariate distributions are not properly balanced, the matching process will fail to remove an important part of the original bias. If the available data does not allow the researcher to obtain this degree of balance, some relaxation of constraints will be needed. It would be ideal if the researcher had more information about the relationship between the outcome and the covariates in order to use it to determine which constraints to relax. For instance, this knowledge could be obtained by splitting the sample into a small planification sample to learn features of the outcome model (for instance, by using LASSO regression with the original covariates and transformations of them) and a larger analysis sample to conduct the actual matching and outcome analyses (see Heller et al. (2009); Zhang et al. (2011) for related ideas).

2.5.2 Larger number of covariates

We explored the performance of cardinality matching with a larger number of covariates using approximation algorithms. With 50 covariates, we were not able to find an exact solution within a one-hour time window, whereas by using the approximation algorithm in `designmatch` (Zubizarreta and Kilcioglu, 2016) we typically found a solution in a few minutes. This approximate solution occasionally violated some of the covariate balancing constraints, but the resulting balance was systematically better than the attained with the distance-based methods. In general, the outcome results were qualitatively similar than with 8 covariates. We also explored the case with 100 covariates, but we were not able to find a solution because of memory constraints. In general, for matching problems with large number of covariates, a practical way to proceed is to use the approximate algorithm in `designmatch` and then either tighten the balancing requirements if the balancing constraints are violated by too much, or use this approximate solution as a "warm start" to find an exact solution in shorter

amount of time. Broadly, the use of approximation algorithms has not been explored much in matching in observational studies and it is an interesting area of research.

2.5.3 Heterogenous effects

When the effects are homogeneous, as in our simulation study, the average treatment effect on the treated units is the same as the average treatment effect on the *matched* treated units, but this is not necessarily the case when the effects are heterogeneous. When the effects are heterogeneous, if the estimand is the average treatment effect on the matched treated units, then all our results carry over, as the only source of bias is that of imbalances in the functions of the covariates that intervene in the true outcome model. If the estimand is the average treatment effect on the treated units, and all the treated units cannot be matched due to limited overlap in covariate distributions between the matched groups, then, in addition to the previous bias due to imbalances, there will be another source of bias due to incomplete matching (Rosenbaum and Rubin, 1985a). In settings where there is limited overlap in covariate distributions, we find more meaningful to target the average treatment effect on the matched treated units, as any estimate of the average treatment effect on the treated units will rely to some extent on extrapolation. In view of this limitation imposed by the data, one way to make progress without making further modeling assumptions is by describing both the matched and unmatched samples (Hill, 2008). This gives a basic understanding of the population to which the results of the matched analysis can be generalized in principle (see Traskin and Small (2011) and Fogarty et al. (2016) for related methods).

2.5.4 Correct specification of the propensity score model

Throughout we have estimated the propensity score using logistic regression, including all the covariates as linear terms in the propensity score model (Austin, 2009, 2011, 2014). We also explored the changes that would be observed if the propensity

score was estimated in a more flexible way, for example, by including higher order and interaction terms of the covariates that intervene in the true outcome and propensity score models. When the true outcome model was additive, it made a substantial difference to include these terms in the propensity score model, as both bias and RMSE were considerably reduced in scenarios 2 and 3 for the distance-based methods. Something similar happened when the true outcome model was additive with interactions. In some of these instances, when the propensity score model was specified in accordance with the true outcome and propensity score models, bias was the lowest with the distance-based methods, however the estimates were quite variable and cardinality matching with fine balance still achieved lower RMSEs. In practice, a good alternative for distance driven matching methods may be to estimate the propensity score using a more flexible approach than logistic regression, for example, by using ensemble methods as in Lee et al. (2010).

2.5.5 Limited overlap in covariate distributions

Assessing limited overlap or lack of common support in covariate distributions is a common practice in observational studies. Its goal is to avoid extrapolating or fabricating results from models that assume specific functional forms (see Section 18.2 of Rosenbaum (2010) and Chapter 14 of Imbens and Rubin (2015) for related discussions). This assessment is typically done, first, by trimming the sample on the propensity score, and second, by checking balance (see, for instance, Crump et al. (2009) and Imbens (2015)). In contrast, cardinality matching directly trims the sample by selecting the largest matched sample that satisfies the investigator's requirements for covariate balance. Certainly, if there is no overlap in a given covariate, then this matched sample will be empty. In a sense, since cardinality matching maximizes the size of the matched sample that is balanced, the sample it finds constitutes the portion of the data set of "maximal" overlap of the covariates distributions given the requirements for covariate balance.

2.5.6 Sensitivity to hidden biases

In standard practice, the construction of a pair-matched sample is done in a single step, by selecting pairs of treated and control units that are similar in terms of a summary of the covariates and hoping that these paired groups will be balanced on aggregate, in terms of each of the covariates. By contrast, in cardinality matching the tasks of balancing and pairing are separated into two steps. First, the method finds the largest pair-matched sample that is balanced, and then, with the balanced sample in hand, it re-pairs the treated and control units minimizing the total sum of distances in some of the covariates. It is known that reducing heterogeneity in pair differences in outcomes results in reduced sensitivity to hidden bias (Rosenbaum, 2005b). Therefore, if the covariates used in re-pairing are strong predictors of the outcome, this second stage will result in reduced sensitivity. This is something that can be exploited in practice either by relying on substantive knowledge of which covariates are predictive of the outcome or by learning them from the data itself (see Baiocchi (2011) for an interesting related method).

2.5.7 Exploring the trade-off between covariate balance and sample size

With any matching method that selects a subset of treated and control observations there is a tension between covariate balance and sample size; this is, a bias-variance trade-off between removing biases due to imbalances in observed covariates and using a larger matched sample to thereby reduce variance. In this study, we have called into question the extensively used rule of thumb of balancing covariates so that their differences in means are not greater than 0.1 standard deviations, and gave broad recommendations to balance covariates under general outcome models. Of course, the applicability of these recommendations will depend on the available data and the resulting sample size after matching. To select a particular balance-size matched

design, one way to proceed in the spirit of King et al. (2017) is to plot the covariate balance-sample size pairs in a two-dimensional plot to explore this trade-off and select a design in the plot. Since this selection does not require the outcomes, it does not affect the objectivity of the study nor the validity of the statistical tests (Rubin, 2008). As we mentioned in the introduction, how to optimally balance covariates is an open question to which this simulation study has given some answers, and which we believe should ultimately be addressed with formal statistical theory.

2.6 Summary and remarks

We presented a Monte Carlo simulation study of three multivariate matching methods that select a subset of treatment and control observations: the widely used nearest neighbor matching with propensity score calipers, and the more recently proposed, optimal subset matching and cardinality matching. We evaluated the performance of these methods according to four different criteria: covariate balance, sample size, covariate distances between matched pairs, and treatment effect estimates. The main findings are the following.

In terms of covariate balance, cardinality matching had the best performance among the three methods. As shown in Figure 2.1, cardinality matching gives the investigator a precise degree of control over covariate balance adjustments. For example, the investigator can require the absolute standardized differences in means to be smaller than 0.1, 0.01, 0.001, and so on, and at the same time directly balance marginal and k-way joint distributions via fine balance and strength-k balancing. In principle, if the investigator assigns more importance to some covariates than others, he or she can balance these covariates more tightly by imposing stronger mean balance or distributional balance constraints on them. Unlike most matching methods, with cardinality matching the propensity score is not needed to balance the covariates because it directly balances the original covariates; however, with cardinality

matching the propensity score may as well be balanced as an additional covariate.

In terms of the size of the matched samples, the results show that, for a given level of covariate balance (e.g., for absolute standardized differences in means smaller than 0.01), cardinality matching systematically selects a larger subset of the observations. Certainly, for stronger forms of covariate balance, the size of the matched sample will be smaller.

In terms of covariate distances, optimal subset matching exhibits the best performance among the three methods. While often considered an instrumental objective in order to balance covariates, reducing covariate distances between matched pairs can be an objective per se because it results into reduced heterogeneity between matched pairs. This, in turn, translates into reduced sensitivity to biases due to hidden covariates under certain models of analysis (see chapter 4 of Rosenbaum (2002)).

The previous findings should not be surprising because cardinality matching is designed to explicitly optimize sample size and directly constrain covariate balance, and optimal subset matching is designed to minimize the covariate distances between matched pairs given a threshold distance. On the other hand, nearest neighbor matching is a greedy method with no optimality guarantees regarding any of the three previous comparison criteria (covariate balance, sample size, and covariate distances).

In terms of treatment effect estimates, our simulation study results suggest that, with a simple differences in means estimator, better covariate balance translates into better estimates. In general, the estimates obtained with cardinality matching have lower RMSEs, except for the case in which it only requires the means to have absolute standardized differences smaller than 0.1. In particular, when the outcome is known to be exactly a linear combination of the observed covariates, tight mean balance appears to be enough to remove systematic biases. However, in practice the investigator does not really know this, and the covariates may affect the outcome in a nonlinear way, so it is preferable to match with fine balance for all the covariates in addition to a tight mean balance constraint. The inclusion of strength-2 balance constraints

for all covariates could significantly improve the estimation when there are important interaction terms affecting the outcome. However, the number of these type of restrictions grows quickly with each additional covariate, and this could be very demanding for some datasets. Thus, it is advised that if the researcher conjectures a possible interaction term between some specific covariates, strength-k balancing for those covariates should be performed if feasible. Note that even when strength-k balancing is not used, fine balance with mean balance constraints provides the best results when estimating a treatment effect under all the scenarios and outcomes examined.

A last but important point observed in this simulation study is the relatively poor performance in every aspect evaluated of the matching that only requires the standardized differences in means of all covariates to be below 0.1. This suggests that the common rule of thumb of balancing covariates so that their absolute standardized differences in means are not greater than 0.1 is typically not enough, and that stronger forms of balance should be pursued in practice when using a simple difference in means effect estimator.

Chapter 3

Stable Balancing Weights for Marginal Structural Models

3.1 Introduction

Many of the most important questions in the health and social sciences relate to the effect of exposures or treatments that are applied not once but in multiple occasions through time. For example, what is the effect of AZT treatment on CD4 counts of HIV patients? What is the impact of antirheumatic drugs on disability and death of the elderly? What are the consecuences of experiencing poverty during childhood on later life outcomes? In these settings, typical methods of adjustment for covariates, such as standard matching or regression methods, fail to yield unbiased estimates of treatment effects, since they would be conditioning on post treatment covariates (Rosenbaum, 1984). A popular alternative to these methods is marginal structural models (MSMs) (Robins et al., 2000).

Under a sequential ignorability assumption (Robins and Hernan, 2008), MSMs yield unbiased treatment effect estimates by weighting each observation by the inverse of the probability of receiving their observed treatment history given their history of observed covariates. However, these probabilities are typically estimated by fitting a

model, and the resulting weights can fail to adjust for observed covariates due to model misspecification. Also, the resulting weights tend to yield very unstable estimates if the predicted probabilities of treatment are very close to zero, which is often the case in practice. To address both of these problems, instead of modeling the probabilities of treatment, we take a design-based approach and directly find the weights of minimum variance that adjust for the covariates across all possible treatment histories. As we show, the proposed approach outperforms standard methods both in terms of bias and variance across a variety of settings.

This chapter is organized in seven sections including this introduction. In Section 2 we provide the setup and notation, and review marginal structural models. In Section 3 we analyze the role of weighting in longitudinal studies. In Section 4 we present the proposed methodology, and in Section 5 we test it, comparing it to standard and recent approaches. In Section 6 we apply the proposed methodology to Chilean educational data. Finally, in Section 7 we conclude with a summary and some remarks.

3.2 Overview of marginal structural models

3.2.1 Setup and notation

We assume that we have a random sample of n individuals that are followed over a total of $T+1$ time periods. In each period $t = 1, 2, ..., T$, subject $i = 1, ..., n$ receives a time-dependent binary treatment Z_{it} and, before the treatment is received, a set of time-dependent covariates X_{it} (usually a vector) is recorded. When the time-dependent covariates predict future treatment assignments and outcome, while also being affected by previous treatment assignments, they are referred to as time-dependent confounders. Treatments and covariates take values from the sets $\mathcal{Z}_t$ and $\mathcal{X}_t$ respectively. The outcome Y_i is observed at the end of follow up at time $T+1$. Let $\bar{Z}_{it} = \{Z_{i1}, ..., Z_{it}\}$ and $\bar{X}_{it} = \{X_{i1}, ..., X_{it}\}$ be the treatment and covariates histories up to time t, so $\bar{Z}_{it}$ takes values from $\bar{\mathcal{Z}}_t = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_t$ and $\bar{X}_{it}$ from $\bar{\mathcal{X}}_t = \mathcal{X}_1 \times \cdots \times \mathcal{X}_t$.

Following the potential outcomes framework formalized by Rubin (1978) and extended to the longitudinal case by Robins (1986), we denote by $Y_i(\bar{z}_T)$ the potential outcome that subject i would have observed if he had received, possibly contrary to fact, treatment history $\bar{z}_T$, for any $\bar{z}_T \in \bar{\mathcal{Z}}_T$. In an analogous way, $X_{it}(\bar{z}_{t-1})$ represents the potential value of the covariates at time t for subject i under treatment history $\bar{z}_{t-1} \in \bar{\mathcal{Z}}_{t-1}$. This notation also implies that we are assuming no interference, this is, the treatment assigned to subject j does not affect the potential values of covariates or outcome of unit i (Rubin, 1980). Additionally, to be able to connect observed quantities and potential values, we assume consistency, which means that when subject i actually receives treatment sequence $\bar{Z}_{iT}$, the observed outcome and covariates are the potential outcome and potential values of covariates under the observed treatment sequence, this is, $Y_i = Y_i(\bar{Z}_{iT})$ and $X_{it} = X_{it}(\bar{Z}_{i(t-1)})$, for each time $t = 1, ..., T$.

3.2.2 Marginal structural models

MSMs are a class of models for some aspect, commonly the mean, of the marginal distribution of potential outcomes (Robins et al., 2000). There are many potential outcome variables $Y_i(\bar{z}_T)$, so a parametric model is used to describe the relationship between the treatment sequences $\bar{z}_T$ and the potential outcomes

$$\mathrm{E}[Y_i(\bar{z}_T)] = g(\bar{z}_T; \boldsymbol{\beta}) \tag{3.1}$$

where g is some known function and $\boldsymbol{\beta}$ is the parameter that needs to be estimated. This model is different from the associational model

$$\mathrm{E}[Y_i \mid \bar{Z}_{iT} = \bar{z}_T] = g(\bar{z}_T; \boldsymbol{\gamma}) \tag{3.2}$$

and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ will be the same only when there are no observed or unobserved confounders. However, Robins et al. (2000) show that under the assumption of no unmeasured confounding (also called sequential ignorability),

$$Y_i(\bar{z}_T) \amalg \bar{Z}_{it} \mid \bar{X}_{it}, \bar{Z}_{i(t-1)} \quad \forall \bar{z}_T \in \bar{\mathcal{Z}}_T, \forall t \in 1, ..., T \tag{3.3}$$

and the positivity assumption,

$$f(\bar{z}_{t-1}, \bar{x}_t) > 0 \Rightarrow f(\bar{z}_t \mid \bar{z}_{t-1}, \bar{x}_t) > 0 \tag{3.4}$$

it is possible to obtain an asymptotically unbiased estimator of the causal parameter $\boldsymbol{\beta}$ by weighting the observed outcomes with the subject-specific weights

$$SW_i = \prod_{t=1}^{T} \frac{f(Z_{it} \mid \bar{Z}_{i(t-1)})}{f(Z_{it} \mid \bar{Z}_{i(t-1)}, \bar{X}_{it})} \tag{3.5}$$

and getting the weighted regression estimator of $\boldsymbol{\gamma}$. This estimator is called inverse-probability of treatment weighted (IPTW) estimator and is called this way because the denominator can be interpreted as the probability for subject i of receiving his observed treatment sequence $\bar{Z}_{iT}$. In fact, if instead of SW_i we use the weights

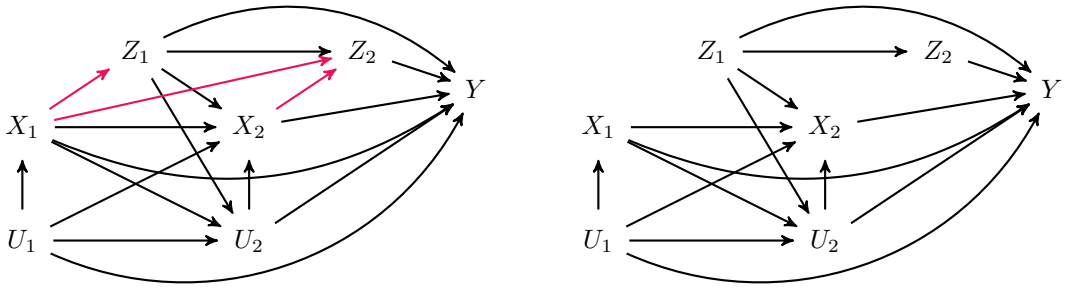
$$W_i = \frac{1}{\prod_{t=1}^{T} f(Z_{it} \mid \bar{Z}_{i(t-1)}, \bar{X}_{it})}, \tag{3.6}$$

the IPTW estimator will still be consistent and asymptotically normal (Robins et al., 2000). However, these weights are quite unstable and an estimator based on them may have a very large variance. For this reason, it is common to follow the suggestion of using the stabilized weights, SW_i. Unfortunately, even though these weights have an improved variance performance compared to the W_i's, as we discuss below, they are still quite unstable themselves.

3.3 On the role of weighting in longitudinal studies

When the probabilities $f(Z_{it} \mid \bar{Z}_{i(t-1)}, \bar{X}_{it})$ are known or are modeled correctly, the stabilized weights SW_i will, on average, remove the confounding. This is, in the weighted sample, the treatment assignment mechanism will be unrelated to the time-dependent confounders. But what are we specifically trying to achieve with the weights used to estimate a MSM? Ideally we would want to create a pseudo-population in which the treatment assignment is not confounded, not only on average across

repeated samples, but in any given sample. So, what does this mean in the case where we have time-dependent confounders that are also affected by previous treatment assignments? In terms of a causal graph like the ones presented in Robins et al. (2000) we would want to go from the structure in Figure 3.1(a) to the structure in Figure 3.1(b), that is, to create a pseudo-population in which the red lines in Figure 3.1(a) are removed.



(a) No unobserved confounders (b) No unobserved or observed confounders

Figure 3.1: Causal graphs

A common way to think about an unconfounded assignment mechanism is to compare it with a randomized experiment, where, at each time t, the treatment assignment Z_{it} does not depend on any of the previous covariate values X_{it}. Another way to look at this is from the covariate balance standpoint. In the setting where we have a binary point exposure, if the distribution of any covariate related to the outcome is the same under both treatment groups, then we could say that the treatment assignment mechanism is not related to these covariates, and conclude that there is no confounding. In the longitudinal case, the balance requirements needed to conclude unconfoundedness are not as straightforward and we need to see what would the goal be in terms of covariate balance in this case.

When using marginal structural models, what we need is a population in which $\mathrm{E}[Y_i(\bar{z}_T)] = \mathrm{E}[Y_i \mid \bar{Z}_{iT} = \bar{z}_T]$, however, to be able to get there we need to first

analyze the role of the observed confounders in those expectations. Suppose we can model the conditional expectation of Y_i given covariates and treatment assignments as $\mathrm{E}[Y_i \mid \bar{X}_{iT}, \bar{Z}_{iT}] = h(\bar{X}_{iT}, \bar{Z}_{iT}; \alpha)$, where h is a known function parametrized by α, so that

$$\mathrm{E}[Y_i \mid \bar{Z}_{iT} = \bar{z}_T] = \mathrm{E}\Big[\mathrm{E}\big[Y_i \mid \bar{X}_{iT}, \bar{Z}_{iT}\big] \;\Big|\; \bar{Z}_{iT} = \bar{z}_T\Big] = \mathrm{E}\big[h(\bar{X}_{iT}, \bar{Z}_{iT}; \alpha) \mid \bar{Z}_{iT} = \bar{z}_T\big]. \tag{3.7}$$

For now, lets say h is linear in X_i's (or that the part of h that involves X_i's can be reasonably well approximated by a linear function), this is, $h(\bar{X}_{iT}, \bar{Z}_{iT}; \alpha) = \alpha_0 + \zeta(\bar{Z}_{iT}; \alpha_Z) + \sum_t \alpha_{Xt} X_{it}$, with ζ a not necessarily known function parametrized by α_Z, then,

$$\mathrm{E}[Y_i \mid \bar{Z}_{iT} = \bar{z}_T] = \alpha_0 + \zeta(\bar{z}_T; \alpha_Z) + \sum_t \alpha_{Xt} \mathrm{E}[X_{it} \mid \bar{Z}_{iT} = \bar{z}_T]. \tag{3.8}$$

This way we can see that the key quantities are the conditional expectations $\mathrm{E}[X_{it} \mid \bar{Z}_{iT} = \bar{z}_T]$, $t = 1, ..., T$, and the conditions needed to estimate $\mathrm{E}[X_{it}(\bar{z}_{t-1})]$, $t = 1, ..., T$, from them.

Now, in a population where there is no confounding, we have that $\mathrm{E}[X_{it}(\bar{z}_{t-1})] = \mathrm{E}[X_{it}(\bar{z}_{t-1}) \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}] = \mathrm{E}[X_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}]$. So we need to create a pseudo-population where $\mathrm{E}[X_{it} \mid \bar{Z}_{iT} = \bar{z}_T] = \mathrm{E}[X_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}, \underline{Z}_{it} = \underline{z}_t] = \mathrm{E}[X_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}]$ $\forall \underline{z}_t$ which would give us the desired

$$\mathrm{E}[X_t \mid \bar{Z}_T = \bar{z}_T] = \mathrm{E}[X_t(\bar{z}_{t-1})] = \mathrm{E}[X_t(\bar{z}_T)], \tag{3.9}$$

where the last equality comes from the fact that an event in the future cannot have a causal effect on the past.

When there is confounding of the effect of the treatment on the time varying covariates, we need to proceed in a similar way as we did with the outcome. Let's say that $\mathrm{E}[X_{it} \mid \bar{X}_{i(t-1)}, \bar{Z}_{iT}] = h_{Xt}(\bar{X}_{i(t-1)}, \bar{Z}_{iT}; \eta_t) = \eta_{t0} + \zeta_{Xt}(\bar{Z}_{iT}; \eta_{tZ}) + \sum_{j=1}^{t-1} \eta_{tXj} X_{ij}$, with ζ_{Xt} a not necessarily known function parametrized by η_{tZ}, then,

$$\begin{aligned}\mathrm{E}\big[X_{it} \mid \bar{Z}_{iT} = \bar{z}_T\big] &= \mathrm{E}\Big[\mathrm{E}\big[X_{it} \mid \bar{X}_{i(t-1)}, \bar{Z}_{iT}\big] \;\Big|\; \bar{Z}_{iT} = \bar{z}_T\Big] \\ &= \eta_{t0} + \zeta_{Xt}(\bar{z}_T; \eta_{tZ}) + \sum_{j=1}^{t-1} \eta_{tXj}\, \mathrm{E}\big[X_{ij} \mid \bar{Z}_{iT} = \bar{z}_T\big].\end{aligned} \tag{3.10}$$

This will yield the quantities of interest if $\mathrm{E}[X_{ij} \mid \bar{Z}_{iT} = \bar{z}_T] = \mathrm{E}[X_{ij}(\bar{z}_{j-1})]$, $j = 1, ..., t-1$, and $\zeta_{Xt}(\bar{z}_{t-1}, \underline{z}_{t-1}; \eta_{tZ}) = \zeta_{Xt}(\bar{z}_{t-1}; \eta_{tZ})\ \forall \underline{z}_{t-1}$.

The first condition will be attained from the "balance" obtained from the previous time points; the second one needs to be targeted directly. This can be done by forcing the pseudo-population to satisfy

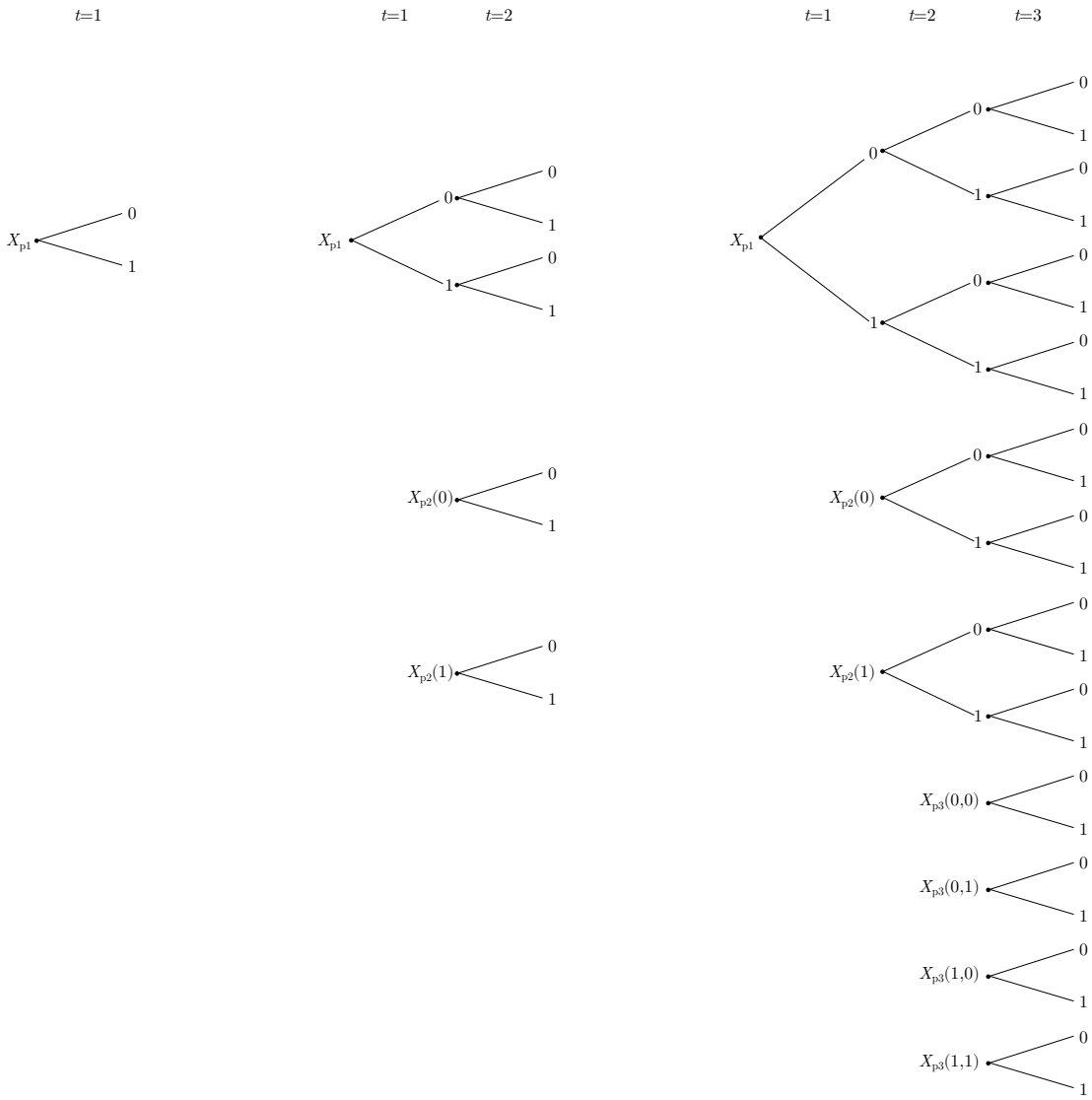
$$\begin{aligned}\mathrm{E}\left[X_{it} - \sum_{j=1}^{t-1} \eta_{tXj} X_j \;\middle|\; \bar{Z}_{iT} = \bar{z}_T\right] &= \mathrm{E}\left[X_{it} - \sum_{j=1}^{t-1} \eta_{tXj} X_{ij} \;\middle|\; \bar{Z}_{i(t-1)} = \bar{z}_{t-1}, \underline{Z}_{i(t-1)} = \underline{z}_{t-1}\right] \\ &= \mathrm{E}\left[X_{it} - \sum_{j=1}^{t-1} \eta_{tXj} X_{ij} \;\middle|\; \bar{Z}_{i(t-1)} = \bar{z}_{t-1}\right] \qquad \forall \underline{z}_t,\end{aligned} \tag{3.11}$$

for which the parameters η_{tXj} need to be estimated. Since the variables $V_{it} = X_{it} - \sum_{j=1}^{t-1} \eta_{tXj} X_{ij}$ are not confounded by previous covariates, like in the last case, we have that $\mathrm{E}[V_{it}(\bar{z}_T)] = \mathrm{E}[V_{it}(\bar{z}_{t-1})] = \mathrm{E}[V_{it}(\bar{z}_{t-1}) \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}] = \mathrm{E}[V_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}] = \mathrm{E}[V_{it} \mid \bar{Z}_{iT} = \bar{z}_T]$.

In general, we can say that we need weights that, at each time t, for every past treatment sequence $\bar{z}_{t-1}$, will balance the observed covariates through all possible "future" treatment sequences $\underline{z}_t$. Figure 3.2 provides a visualization of the number of conditions that need to be satisfied. As the reader may notice, this number grows exponentially with T, the total number of periods.

However, this number may be reduced if one chooses a simpler marginal structural model $\mathrm{E}[Y_i(\bar{z}_T)] = g(\bar{z}_T; \beta)$. For example, we can use the model $\mathrm{E}[Y_i(\bar{z}_T)] = \beta_0 + \beta_1 \sum_t z_t$, with which we would be implying that the effect of the treatment on the outcome depends only on the cumulative amount of treatment and not on the specific

Figure 3.2: Balancing conditions for different study lengths T



order in which the treatment was received. This means that we would only be comparing the $T+1$ possible groups that have been assigned different cumulative amounts of treatment, this is, $\mathrm{E}[Y_i(\bar{z}_T)]$ could be expressed as $\mathrm{E}\left[Y_i\left(\sum_t z_t\right)\right] = \mathrm{E}[Y_i(s_T)]$ where $s_T \in \{0, 1, ..., T\}$.

Following the same reasoning as before, we can see that, without making any further assumption on X_{it}'s, the balancing conditions that need to be satisfied at each time t for every sequence $\bar{z}_{t-1}$ are $\mathrm{E}\left[X_{it} \,\middle|\, \bar{Z}_{i(t-1)} = \bar{z}_{t-1}, \sum_{j=t}^{T} Z_(ij) = s_T - \sum_{j=1}^{t-1} z_j\right] = \mathrm{E}[X_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}]$ $\forall s_T \in \{0, 1, ..., T\}$, which still grows exponentially with T. If in addition we assume that, as with the outcome, the observed covariates are affected by the treatment only through the cumulative amount of treatment received, then the balancing conditions that the pseudo-population should satisfy would be that, for each time t and cumulative amount of treatment received before time t, $s_{t-1} \in \{0, 1, ..., t-1\}$, $\mathrm{E}\left[X_{it} \,\middle|\, \sum_{j=1}^{t-1} Z_{ij} = s_{t-1}, \sum_{j=t}^{T} Z_{ij} = s_T - s_{t-1}\right] = \mathrm{E}\left[X_{it} \,\middle|\, \sum_{j=1}^{t-1} Z_{ij} = s_{t-1}\right]$, $\forall s_T \in \{0, 1, ..., T\}$. The same applies for the V_{it} variables when the effect of the treatment on covariates is confounded by other covariates.

Certainly, if we add lag structure assumptions, the number of conditions that need to be satisfied will decrease considerably. On the other hand, if the model for the conditional mean of the outcome is not linear in the observed covariates, or want to use a higher order approximation of the function, then we would have to balance higher order moments or obtain fine balance for a discretized version of the continuous covariates.

3.4 Stable balancing weights in longitudinal studies

As stated before, inverse probability of treatment weights can be highly unstable, which results in a large variance estimator and wide confidence intervals. In addi-

tion, we can also fail to adjust for observed covariates due to model misspecification. To address these problems, instead of modeling the conditional probability of each subject receiving their own treatment history given its observed covariates, we take a design-based approach and directly find the weights of minimum variance that balance covariates as determined by the researcher, following the analysis developed in Section 3.3. To obtain the weights, we solve the convex optimization problem

$$
\begin{aligned}
\underset{\boldsymbol{w}}{\text{minimize}} \quad & \|\boldsymbol{w} - \overline{\boldsymbol{w}}\|_2^2 \\
\text{subject to} \quad & \left| \frac{\sum_{i \in \mathcal{I}_{\bar{z}_{t-1}} \cap \mathcal{I}_{\underline{z}_t}} w_i X_{it_p}}{\sum_{i \in \mathcal{I}_{\bar{z}_{t-1}} \cap \mathcal{I}_{\underline{z}_t}} w_i} - \frac{\sum_{i \in \mathcal{I}_{\bar{z}_{t-1}}} X_{it_p}}{\left|\mathcal{I}_{\bar{z}_{t-1}}\right|} \right| \leq \delta_{\bar{z}_{t-1_p}} \quad \begin{array}{c} \forall \bar{z}_{t-1} \in \bar{\mathcal{Z}}_{t-1} \\ \forall \underline{z}_t \in \underline{\mathcal{Z}}_t \\ t=1...,T \\ p=1,...,P \end{array} \\
& \sum_{i \in \mathcal{I}_{\bar{z}_T}} w_i = \tfrac{1}{n} \left|\mathcal{I}_{\bar{z}_T}\right| \quad \forall \bar{z}_T \in \bar{\mathcal{Z}}_T \\
& w_i \geq 0 \qquad i = 1, ..., n
\end{aligned}
\tag{3.12}
$$

where $\boldsymbol{w}$ is the n-dimensional vector of weights w_i's, $\overline{\boldsymbol{w}}$ is a vector with all entries equal to the mean of the weights, X_{it_p} is covariate p at time t for subject i, $\mathcal{I}_{\bar{z}_t} = \{i \in \{1, ..., n\} : \bar{Z}_{it} = \bar{z}_t\}$ and $\mathcal{I}_{\underline{z}_t} = \{i \in \{1, ..., n\} : \underline{Z}_{it} = \underline{z}_t\}$ are the corresponding index sets, and $\delta_{\bar{z}_{t-1_p}}$ is a scalar determined by the researcher.

With the objective function in (3.12) we are minimizing the variance of the weights, thus obtaining the most "stable" weights for a given level of covariate balance. The inequality constraints correspond to the balance conditions $\mathrm{E}[X_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}, \underline{Z}_{it} = \underline{z}_t] = \mathrm{E}[X_{it} \mid \bar{Z}_{i(t-1)} = \bar{z}_{t-1}]$ from Section 3.3. With this, we are forcing the expectation on the left hand side in the pseudo-population to be close to the expectation on the right hand side, which comes from the original population. Notice that if we want to balance higher order moments of the covariates, or obtain fine balance, we only need to augment the covariate matrix $\boldsymbol{X}$ with the corresponding transformations of the covariates. When there are variables that confound the relationship between the time-dependent treatment and the time varying confounders, then we would replace X_{it}'s for V_{it}'s, for $t = 2, ..., T$, after estimating the parameters of the models $\mathrm{E}[X_{it} \mid \bar{X}_{i(t-1)}, \bar{Z}_{i(t-1)}] = h_{Xt}(\bar{X}_{i(t-1)}, \bar{Z}_{iT}; \eta_t)$. Furthermore, the

lag-structure and the "simplifications" of the marginal structural model would be incorporated in the definition of the index sets. For example, when we assume that both outcome and time-varying covariates are affected by the treatment only through the total amount of treatment received, the index sets would be defined as $\mathcal{I}_{s_t} = \left\{ i \in \{1, ..., n\} : \sum_{j=1}^{t} Z_{ij} = s_t \right\}$, for $s_t \in \{0, 1, ..., s_T\}$, instead of $\mathcal{I}_{\bar{z}_t}$, and $\mathcal{I}_{s_T} = \left\{ i \in \{1, ..., n\} : \sum_{j=1}^{T} Z_{ij} = s_T \right\}$, $s_T \in \{0, 1, ..., T\}$, instead of $\mathcal{I}_{\underline{z}_T}$. This way we can incorporate our structural assumptions in a convex quadratic programming problem, which means that, if the problem has a solution, it can be found quickly.

It is important to note that when the problem is not feasible, we are still obtaining relevant information about the data. In a certain way, this would be telling us that the data cannot be balanced as tightly as we might need it to be. Moreover, even though we could increase the δ's to obtain a solution, we would know that we should be cautious about the conclusions we draw from our analysis. It is always crucial to check the balance obtained with any covariate adjustment method that does not balance covariates directly, and this is generally performed for the one-time treatment case. However, in the longitudinal case, this step is usually not performed. When using marginal structural models, researchers are not used to evaluate the covariate balance reached after weighting. In this regard, Jackson (2016) discusses the lack of tools available for this kind of assessments and develops diagnostics to describe residual confounding after weighting, along with a companion `R` package.

3.5 Simulation study

In this section we conduct two simulation studies to evaluate the performance of the proposed method. Here, we compare the estimates obtained using the proposed weighting method with the ones obtained using the inverse probability of treatment weights in terms of bias, root mean square error (RMSE), and interval coverage and

length. The probability of treatment is estimated using both, maximum likelihood estimation and the robust estimation in Imai and Ratkovic (2015).

3.5.1 Data generating mechanism

The first simulation setup considered followed one of the scenarios studied in Imai and Ratkovic (2015), where the treatment assignment model was correctly specified. However, instead of using $T = 3$ time periods, we set $T = 2$ so that an acceptable level of covariate balance (Chapter 2) could be achieved in most samples. Figure 3.3 shows graphically the data structure in this simulation. Here, X_t refers to a 4-dimensional continuous time-varying covariate, Z_t is a binary treatment at each time point, and Y is a continuous outcome at the end of the $T = 2$ time periods. Specifically, let $X'_{itk} \sim \mathcal{N}(0, 1)$, $i = 1, ..., n$, $t = 1, 2$, $k = 1, 2, 3, 4$,

$X_{it} = \left(C_{it}X'_{it1}, C_{it}X'_{it2}, |C_{it}X'_{it3}|, |C_{it}X'_{it4}|\right)$, where $C_{i1} = 1$ and $C_{i2} = 2 + \frac{2Z_{i1}-1}{3}$,

$\text{logit}\left(P(Z_{it} = 1)\right) = -Z_{i(t-1)} + X_{it1} - \frac{1}{2}X_{it2} + \frac{1}{4}X_{it3} + \frac{1}{10}X_{it4} + \left(-\frac{1}{2}\right)^t$, where $Z_{i0} = 0$, and

$Y_i = 250 - 10\sum_{t=1}^{2} Z_{it} + \sum_{t=1}^{2}(27.4X_{it1} + 13.7X_{it2} + 13.7X_{it3} + 13.7X_{it4}) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 25)$.

The marginal structural model used was $E[Y_i(\bar{z}_2)] = \beta_0 + \beta_1 z_1 + \beta_2 z_2$, that is, treatment at different times could have different effects on the outcome. It is important to notice that in this scenario, the relationship of the treatment and time-varying confounders is not confounded by other covariates. This means that we can use directly the covariates X_t's (and the desired transformations) as input in the optimization problem without first obtaining their corresponding V_t variables.

The second design was constructed based on the first one and modified accordingly to follow the data structure shown in Figure 3.4. Here we can see the following three main differences with respect to the previous setting:

- There is an unobserved variable U_i that influences both the observed covariates

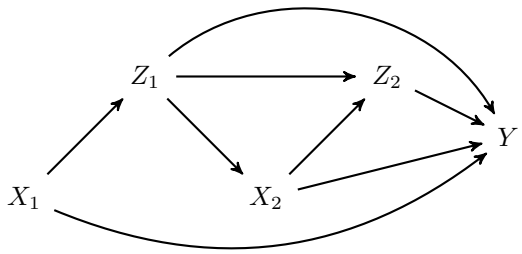


Figure 3.3: Data structure for simulation 1

and the outcome. In health related examples, a variable like this is sometimes thought of as a "general health status" of individual i.

- The observed covariates at time t are influenced by their value at time $t-1$.
- $T = 3$ time periods.

Specifically, again, let $X'_{itk} \sim \mathcal{N}(0,1)$, $i = 1, ..., n$, $t = 1, 2, 3$, $k = 1, 2, 3, 4$, and $U_i \sim \mathcal{U}(1,5)$, $i = 1, ..., n$,

$X_{i1} = \left(\frac{1}{U_i}X'_{i11}, \frac{1}{U_i}X'_{i12}, \left|\frac{1}{U_i}X'_{i13}\right|, \left|\frac{1}{U_i}X'_{i14}\right|\right)$, with the following definitions for $t = 2, 3$,

$X_{itp} = X_{i(t-1)p} + \frac{1}{U_i}X'_{itp} + Z_{i(t-1)}$ for $p = 1, 2$, and

$X_{itp} = X_{i(t-1)p} + \min\left\{X_{i(t-1)p}, \max\left\{-X_{i(t-1)p}, \frac{1}{U_i}X'_{itp}\right\}\right\} + Z_{i(t-1)}$ for $p = 3, 4$.

Treatment assignment is the same as in the previous setting and the outcome is given by

$Y_i = 250 - 10\sum_{t=1}^{2} Z_{it} - 58.5Z_{i3} + 27.4X_{i31} + 13.7X_{i32} + 13.7X_{i33} + 13.7X_{i34} + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 25)$.

The marginal structural model used in this case was $E[Y_i(\bar{z}_3)] = \beta_0 + \beta_1 \sum_{t=1}^{3} z_t$, that is, the effect of the treatment on the outcome is a function of the total amount of treatment received.

In both simulations, three different sample sizes were considered, $n = 500, 1000$, and 2500. Finally, each data set was generated 2500 times.

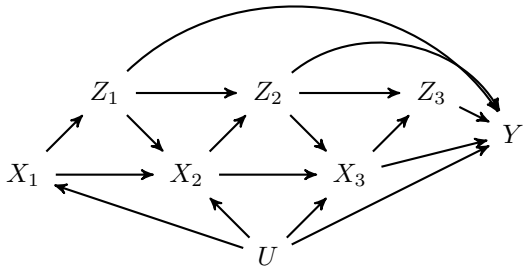


Figure 3.4: Data structure for simulation 2

3.5.2 Estimation

For each repetition in the simulations, four sets of weights were obtained:

- LSBW: stable balancing weights with all standardized differences in means smaller than 0.01.
  - Scenario 1: no further assumptions were made, so the original covariates X_{it_p} were balanced through all possible corresponding future treatment sequences $\underline{Z}_{it}$ given the past treatment history $\bar{Z}_{i(t-1)}$.
  - Scenario 2: since there is dependance between X_{it} and $X_{i(t-1)}$, we need to model the part in $\mathrm{E}[X_{it} \mid \bar{X}_{i(t-1)}, \bar{Z}_{iT}]$ that refers to this dependence structure . The model used was $\mathrm{E}[X_{it} \mid \bar{X}_{i(t-1)}, \bar{Z}_{iT}] = \xi_0^0 + \xi_1^0 X_{i(t-1)}$ when $Z_{i(t-1)} = 0$, and $\mathrm{E}[X_{it} \mid \bar{X}_{i(t-1)}, \bar{Z}_{iT}] = \xi_0^1 + \xi_1^1 X_{i(t-1)}$ when $Z_{i(t-1)} = 1$. With this model we are assuming that X_{it} is affected by $\bar{X}_{i(t-1)}$ only through $X_{i(t-1)}$ and we are allowing this relationship to depend on the previous treatment assignment $Z_{i(t-1)}$. Notice that since we do not need to specify a functional form for the model $\eta_{t0} + \zeta_{Xt}(\bar{Z}_{iT}; \eta_{tZ})$ or any part of the h_{Xt} that do not include X_{it}'s, these parts are incorporated in the coefficients ξ_0^0 and ξ_0^1. After estimating the relevant parameters of the model, we balanced the variables $V_{it_p} = X_{it_p} - \hat{\xi}_1^{Z_{i(t-1)}} X_{i(t-1)_p}$ through all possible cumulative amounts of treatment from time t to time T, given $Z_{i(t-1)}$. This is, we

assumed that only the immediate previous treatment assignment, $Z_{i(t-1)}$, has a direct causal effect on X_{it}.

- GLM: stabilized version of inverse probability weights using a logistic regression model fitted separately for each time period for the denominator, and the sample proportion of each treatment sequence for the numerator. The logistic regression models were fitted using the `glm` function in `R`.

- CBPS: the corresponding robust estimation weights were obtained with the `R` function `CBMSM` from the `CBPS` package (Fong et al., 2016) for the same logistic regression model as in GLM. The `twostep` argument was set to TRUE, for a two-step estimator, and `msm.variance` was set to "approx", for the low-rank approximation of the variance. Setting `twostep = TRUE` and `msm.variance = "full"` was also explored, but the performance of the previous function arguments was better and therefore those are the results presented in Subsection 3.5.3.

- True: stabilized inverse probability weights using the true probability of treatment.

The parameters of the marginal structural model were estimated via weighted LS regression using these four sets of weights. The parameters were also estimated without weights (Unwt) to give an idea of how confounded the original data was. 95% confidence intervals using the robust sandwich variance estimator were recorded to assess their coverage and average length.

3.5.3 Results

Figure 3.5 summarizes the distribution of the estimated parameters of the MSM, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, across simulated datasets in the first simulation design. Here we can observe that for inverse probability weights, using both the true probability and

the GLM estimated probability, the distribution is indeed centered around the true parameter value. However, the estimator is quite variable and in many instances, the estimated value is very far from the true value, even farther than every point in the unweighted estimator distribution. In these boxplots, we can also notice that the estimates of the LSBW are quite stable, since, in addition to being approximately unbiased, they also present a low variability, with practically no outliers, and all of them still close to the true value of the parameters. Table 3.1, which complements the boxplots, shows that its root mean square error (RMSE) is the lowest of all the methods for every sample size. Regarding the performance of CBPS, we can see that the estimates are less variable than with standard IPTW, with fewer outliers, but more biased than without weighting. Due to this large bias, its RMSE is the largest among all methods. We can also observe from Figure 3.5 and Table 3.1 that every weighting method improves as the sample size increases, being the standard IPTW methods the most benefited from the increase in sample size.

Table 3.2 shows how the previous performance analysis translates into confidence intervals. We can see that, even using the robust sandwich variance estimator, the actual coverage of the intervals obtained using the standard IPTW is lower than the nominal coverage of 95%. From these two methods, the coverage of the intervals obtained with the estimated probability was higher than when the true probability was used, while their average length was only slightly bigger. We can also say that the intervals obtained using LSBW are quite conservative, since its actual coverage is higher than 99% for all the three parameters and with the 3 sample sizes. Moreover, these confidence intervals are narrower than the standard IPTW ones. The intervals obtained using CBPS practically never covered the true value of the parameters.

The results obtained for the second setting are rather similar to the first one. Figure 3.6 shows the corresponding boxplots that summarize the distribution of the estimates of the two parameters of the marginal structural model used in the second setting, $\hat{\beta}_0$ and $\hat{\beta}_1$. Again, the estimates obtained using inverse probability weights

Figure 3.5: Boxplots of the estimated parameters in the first setting

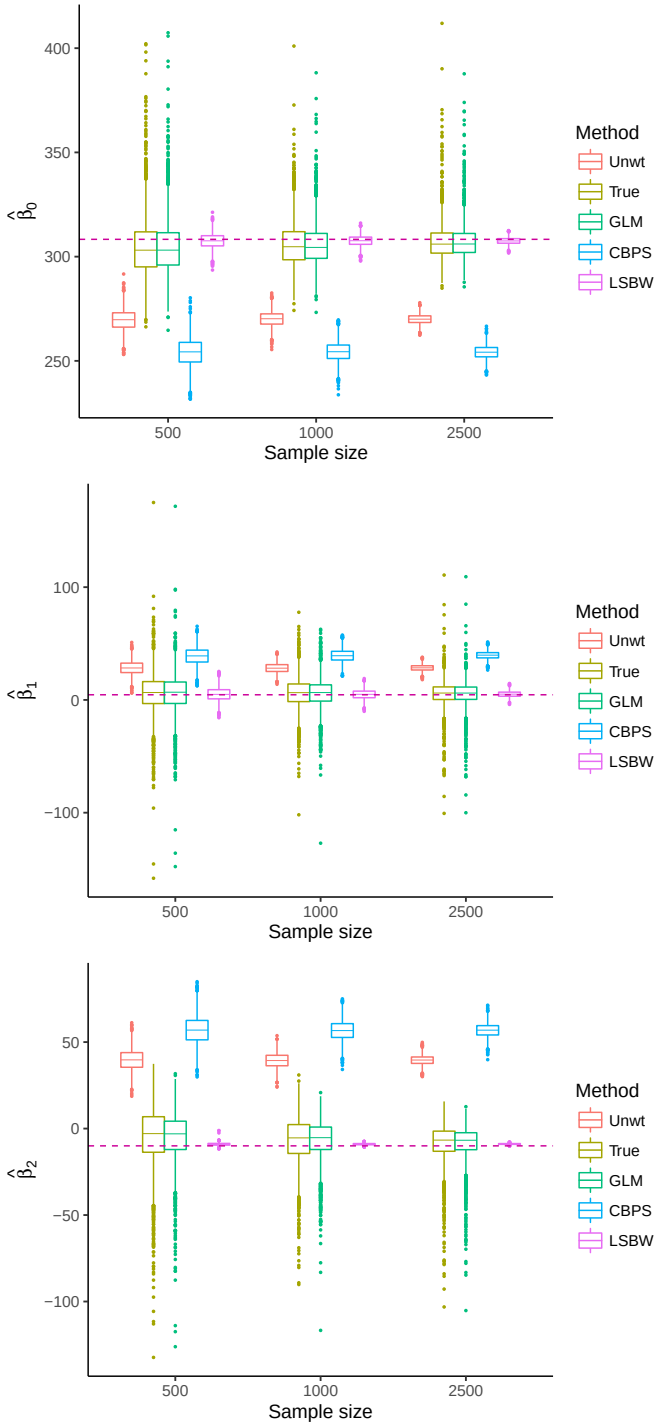


Table 3.1: Bias and RMSE of the estimated parameters of the MSM in the first setting

	Method	$\hat{\beta}_0$		$\hat{\beta}_1$		$\hat{\beta}_2$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$n = 500$	Unwt	-38.54	38.91	23.79	24.63	49.78	50.19
	True	-3.55	15.41	1.19	18.55	4.93	18.31
	GLM	-3.46	14.12	1.14	18.05	4.70	15.68
	CBPS	-54.19	54.66	34.31	35.19	67.00	67.54
	LSBW	-0.73	3.72	0.43	6.06	0.92	1.19
$n = 1000$	Unwt	-38.19	38.37	23.69	24.10	49.34	49.54
	True	-2.28	11.36	1.35	13.98	2.98	13.98
	GLM	-2.25	10.70	1.33	13.72	3.00	12.34
	CBPS	-54.02	54.25	34.72	35.20	66.63	66.91
	LSBW	-0.68	2.66	0.31	4.29	1.00	1.12
$n = 2500$	Unwt	-38.31	38.38	23.89	24.05	49.53	49.61
	True	-1.02	9.16	0.83	11.87	1.55	11.23
	GLM	-0.97	8.53	0.76	11.49	1.49	10.10
	CBPS	-54.13	54.23	35.10	35.28	66.79	66.91
	LSBW	-0.78	1.79	0.54	2.73	1.09	1.14

Note: In the case where $n = 500$, in four of the 2500 repetitions the optimization problem of the LSBW was infeasible. These cases were not included in the bias and RMSE computations.

Table 3.2: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the first setting

	Method	β_0		β_1		β_2	
		Coverage	Length	Coverage	Length	Coverage	Length
$n = 500$	Unwt	0.00	20.66	4.24	24.81	0.00	24.43
	True	81.60	42.47	89.12	51.11	79.88	50.16
	GLM	84.60	42.87	90.96	51.61	86.44	50.78
	CBPS	0.00	23.07	0.48	27.20	0.00	26.48
	LSBW	99.92	30.56	99.60	36.55	100.00	35.99
$n = 1000$	Unwt	0.00	14.62	0.00	17.58	0.00	17.30
	True	83.64	34.36	90.80	41.77	81.64	40.99
	GLM	87.36	34.61	92.92	41.95	88.16	41.24
	CBPS	0.00	16.31	0.00	19.31	0.00	18.76
	LSBW	99.92	21.02	99.52	25.14	100.00	24.74
$n = 2500$	Unwt	0.00	9.25	0.00	11.12	0.00	10.94
	True	86.04	25.85	91.40	31.81	84.60	31.13
	GLM	88.88	26.00	93.20	31.94	89.52	31.27
	CBPS	0.00	10.28	0.00	12.22	0.00	11.84
	LSBW	99.92	13.09	99.64	15.70	100.00	15.43

with the true probability and the probability estimated using GLM, were centered around the true values of the parameters, but with substantial variability and many outliers. On the other hand, LSBW estimates were also centered around the true values of the parameters while achieving reduced variability with very few outliers. The RMSEs shown in Table 3.3 quantify this comparison. Here we can see that the RMSEs from the former methods were, in this case, more than 2.5 times those of LSBW. It is interesting to note that this difference became more pronounced as the sample size increased. CBPS estimates presented less outliers than the standard IPTW methods, however, in this setting they were in general more variable than the latter and more biased than the unweighted estimates.

As an alternative way to study the accuracy and precision of the methods, Table 3.4 shows the coverage of the intervals and their average length. Like in the first setting, we can see that the actual coverage of the standard IPTW was lower than the 95% nominal coverage, with the GLM method being better than when the true probability was used. LSBW achieved a coverage greater than 95%, while presenting the shortest intervals of all the weighting methods. The intervals obtained using the CBPS weights were slightly narrower than the other IPTW ones, but only covered the true parameter values in very few occasions.

In general, from the two simulations we can see that, in these settings, the proposed LSBW managed to reduce the variability of the MSM parameter estimates compared to the other methods, while staying approximately unbiased.

Figure 3.6: Boxplots of the estimated parameters in the second setting

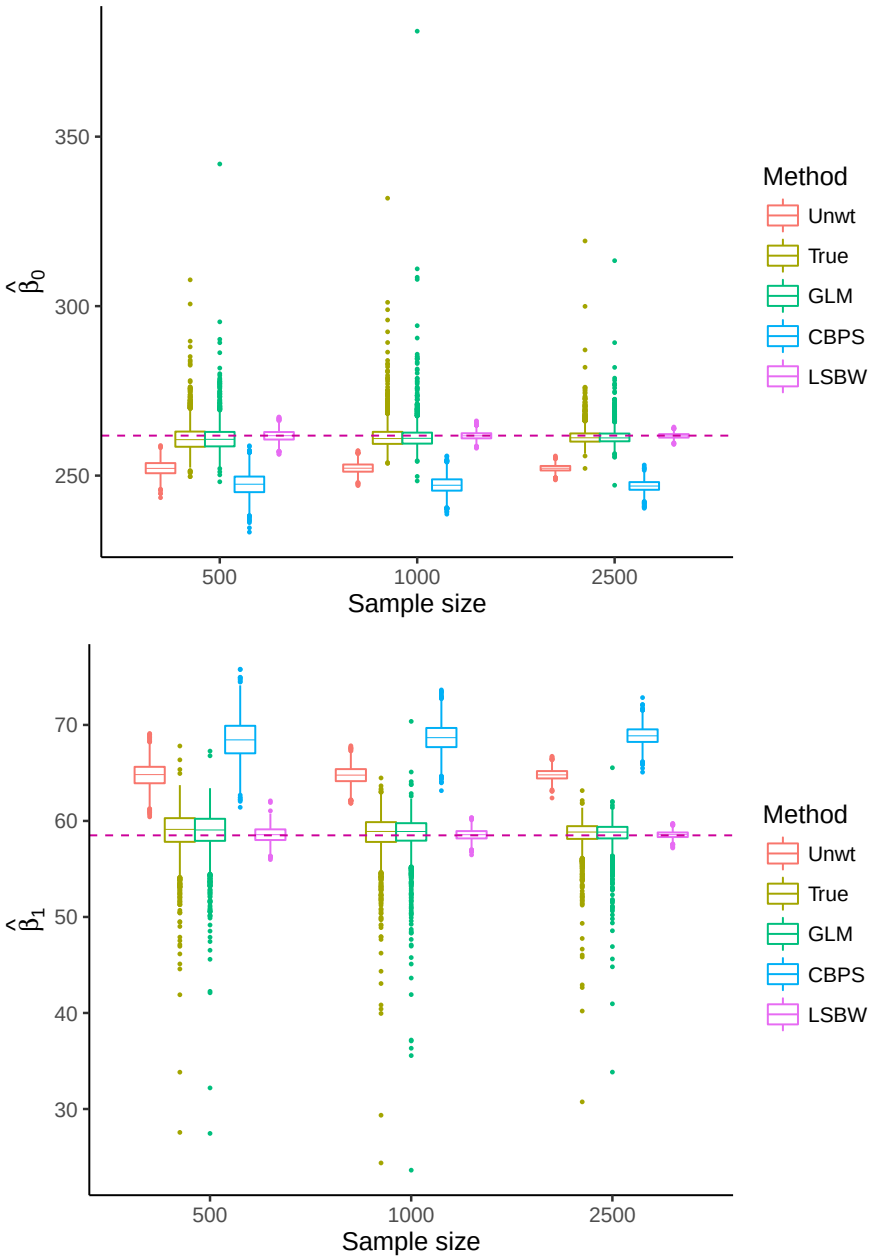


Table 3.3: Bias and RMSE of the estimated parameters of the MSM in the second setting

	Method	$\hat{\beta}_0$		$\hat{\beta}_1$	
		Bias	RMSE	Bias	RMSE
$n = 500$	Unwt	-9.62	9.86	6.30	6.43
	True	-0.69	4.19	0.37	2.37
	GLM	-0.67	4.23	0.34	2.26
	CBPS	-14.36	14.78	9.98	10.20
	LSBW	-0.04	1.64	0.06	0.81
$n = 1000$	Unwt	-9.59	9.71	6.29	6.36
	True	-0.26	3.95	0.14	2.19
	GLM	-0.26	4.44	0.14	2.16
	CBPS	-14.61	14.83	10.19	10.30
	LSBW	-0.03	1.16	0.06	0.58
$n = 2500$	Unwt	-9.63	9.68	6.31	6.33
	True	-0.27	2.79	0.12	1.58
	GLM	-0.26	2.61	0.12	1.47
	CBPS	-14.89	14.99	10.38	10.43
	LSBW	-0.06	0.73	0.07	0.36

Table 3.4: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the second setting

	Method	β_0		β_1	
		Coverage	Length	Coverage	Length
$n = 500$	Unwt	0.60	8.50	0.20	5.02
	True	88.48	12.86	88.64	7.15
	GLM	90.16	13.00	91.92	7.23
	CBPS	0.92	12.25	0.04	7.35
	LSBW	99.48	9.74	99.92	5.49
$n = 1000$	Unwt	0.00	6.04	0.00	3.56
	True	88.08	10.40	88.92	5.78
	GLM	90.80	10.43	91.68	5.85
	CBPS	0.00	8.86	0.00	5.31
	LSBW	99.56	6.78	99.88	3.82
$n = 2500$	Unwt	0.00	3.83	0.00	2.26
	True	89.40	7.36	89.48	4.14
	GLM	90.84	7.41	91.68	4.16
	CBPS	0.00	5.76	0.00	3.45
	LSBW	99.60	4.27	99.84	2.40

3.6 Case study: education voucher system in Chile

To show how the proposed stable balancing weights work on a real application, we implement the method on a controversial matter on Chilean education. The education reform in Chile in the 80's introduced a voucher system where public and private schools would receive a payment that depends on the number of students they have and their daily attendance. Since the establishment of this system, the number of students in private schools has been increasing, especially in the last decade. In this period, the percentage of students in public schools decreased from 51% in 2014 to 37% in 2015, while the percentage of students in subsidized private schools increased

from 42% in 2014 to 55% in 2015 (MINEDUC, 2016). The rest of the students attend private non-subsidized schools that are entirely funded by student tuition, this is, they don't receive any public funding.

Following these changes, there has been an open debate on whether subsidized private schools, also referred to as voucher schools, provide a better education than their public counterpart. This question has been analyzed and discussed from different angles in several studies (Anand et al., 2009; Hsieh and Urquiola, 2006; Lara et al., 2011; McEwan, 2001; Mizala and Romaguera, 2001; Sapelli and Vial, 2002, 2005; Zubizarreta and Keele, 2017), arriving at different conclusions. For instance, in Zubizarreta and Keele (2017) the outcome used was the 2006 language and math SIMCE scores from tenth grade students and use student level data from 2004, when students were in eighth grade, and school level measurements from 2003 to do a multilevel matching. They conclude that the hypothesis that attending a private subsidized school has no effect on test scores cannot be rejected. Lara et al. (2011) compare 2006 test scores from tenth grade students that moved at the end of the eighth grade from a public school to a private voucher school with those of the students who stayed in the public school system. They use propensity score methods and the changes-in-changes approach (Athey and Imbens, 2006) using the test scores of the same students in 2004, when they were in eighth grade. This is, they study the effect of being two years in private voucher school after being in a public school. From both approaches, they conclude that the effect is positive but very small and can be non statistically significant depending on the method used. Alternatively, other studies (Anand et al., 2009; Mizala and Romaguera, 2001) that use cross sectional data from students in tenth grade had found that attending a private voucher school has a positive effect of 15% to 20% of a standard deviation on test scores.

Unlike the previous studies, in our case study we explicitly consider the fact that education is cumulative, so we analyze the effect that the total number of years a student spends in a private voucher school during secondary education (from eleventh

year to twelfth year) has on the scores they obtain on the University Selection Test (PSU, Prueba de Selección Universitaria in Spanish).

3.6.1 Data

SIMCE standardized tests are administered to students in fourth year every year and is alternated between students in eighth year and tenth year each year. In 2013, the students that took the eighth grade exam in 2011, also took the tenth grade exam, providing another set of test scores for the same students. Using student's unique identifiers, SIMCE data from 2011 to 2014 were merged to the PSU data from 2015. This way, we have panel data for the students who presented the PSU on 2015 that includes their SIMCE test scores when they were in eighth (2011) and tenth (2013) grades, as well as other background characteristics collected in those years. Yearly data from MINEDUC, including student's GPA and school attendance, was also included in the panel.

Only students that remained in urban Santiago City area from 2011 to 2015 were considered in the study. Students that at any point of this period switched to a private non-subsidized school or for which school dependency was not available were excluded, as well as those without observed math and language PSU scores. The exclusion criteria was satisfied by 26,389 observations.

The baseline covariates that were balanced were father's education (5 categories), mother's education (5 categories), household income (7 categories), and gender. The time varying covariates to be balanced were, from 2011 and 2013, school's socioeconomic group (5 categories), language score, and math score, and for every year from 2011 to 2014, student's GPA and attendance.

3.6.2 Results

The outcomes in this study were the 2015 PSU language score and math score. Both were analyzed separately but using the same marginal structural model, $\mathrm{E}[Y(\bar{z}_T)] =$

$\beta_0 + \beta_1 \sum_{t=1}^{4} z_t$, where $z_t = 1$ indicates going to a private voucher school on year t and $t = 1$ represents the year 2012. The observed covariates were balanced in accordance with this model. SIMCE language and math scores from 2013 were assumed to depend on previous treatment assignments in a cumulative way and linearly on the 2011 corresponding score. Imposing mean balance for all the covariates represented a total of 205 balance conditions. The general balance achieved was absolute standardized differences in means lower than 0.05. Figure 3.7 shows a summary of the improvement of balance for all covariates for all times and Tables 3.5 and 3.6 show in detail the balance of the demographic covariates before weighting and after weighting, respectively. Table 3.5 shows that in many instances, the balance was good in the unweighted data. However, in other cases, in particular for the students who spent two years in a private voucher school, the absolute standardized differences in means were higher, but balanced after weighting. In Figure 3.7, we can observe that for every covariate, time, and total of years in a voucher school, a tight balance was obtained, even for the most unbalanced cases.

Using these weights, the effect of each additional year in a private voucher school on PSU language and math scores was estimated (Table 3.7). According to these estimates, each additional year a student from urban Santiago City area spends in a private voucher educational institution in middle school has a significant positive average effect of 3.85 points (3.5% of a standard deviation) on their PSU language score and a significant positive average effect of 4.83 points (4.4% of a standard deviation) on their PSU math score. These results represent a small effect that may not be deemed relevant when we consider the effect of one year at a time. However, if we compare four years of continuous private voucher education to four years of continuous public education, the cumulative effect of attending a voucher school is an average increase of 15.40 points in the language score and 19.32 in the math score. This represents 14.0% and 17.6% of a standard deviation, which is close to what some consider a meaningful effect.

Figure 3.7: Boxplots of absolute standardized differences in means before and after weighting

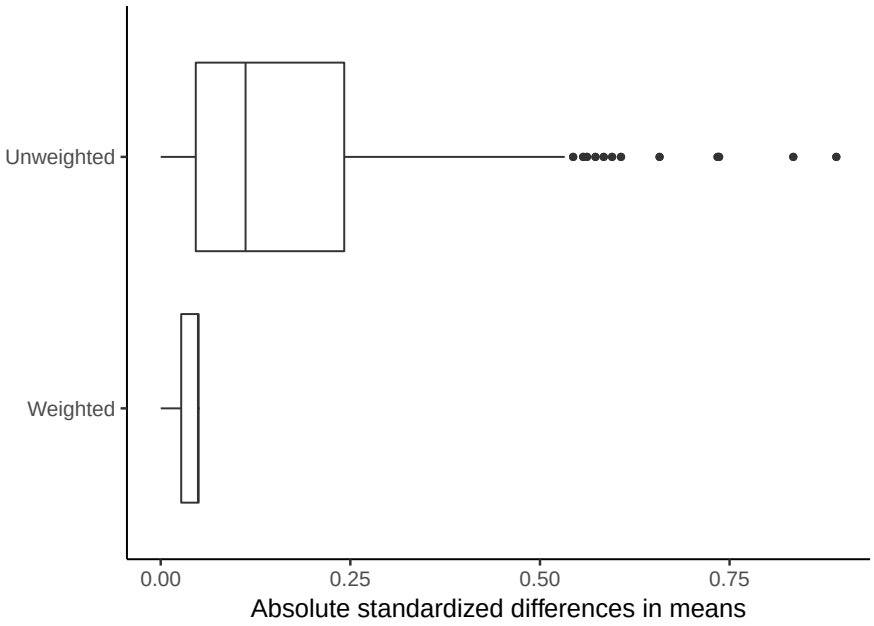


Table 3.5: Absolute standardized differences in means of baseline demographic covariates for each possible number of total years spent in private voucher school without weighting

	Years spent in a private voucher school				
Covariate	0	1	2	3	4
Father's education					
Primary	0.1118	0.0415	0.3962	0.2692	0.1163
Secondary	0.0374	0.082	0.0168	0.0431	0.0107
Technical	0.0975	0.0732	0.2288	0.1539	0.0755
College	0.0496	0.1199	0.2928	0.2395	0.0985
Missing	0.0509	0.0204	0.0240	0.0194	0.0059
Mother's education					
Primary	0.1839	0.012	0.4332	0.302	0.1354
Secondary	0.0868	0.0882	0.0153	0.004	0.0078
Technical	0.1067	0.0993	0.2887	0.1803	0.0900
College	0.0613	0.0838	0.2101	0.1788	0.076
Missing	0.0521	0.0345	0.0420	0.0357	0.0017
Household income category (in 1000 pesos)					
[0,100)	0.0994	0.0043	0.1596	0.1588	0.0658
[100-200]	0.1637	0.0792	0.3717	0.2454	0.1191
(200-400]	0.0070	0.0557	0.0650	0.0755	0.0272
(400-600]	0.1021	0.0039	0.1915	0.1039	0.0554
(600-1400]	0.1533	0.1331	0.2684	0.2361	0.1120
>1400	0.0636	0.1279	0.1639	0.1610	0.0720
Missing	0.0603	0.0365	0.0411	0.0363	0.0004
Female	0.1540	0.0439	0.0236	0.0204	0.0156

Table 3.6: Absolute standardized differences in means of baseline demographic covariates for each possible number of total years spent in private voucher school after weighting

	Years spent in a private voucher school				
Covariate	0	1	2	3	4
Father's education					
Primary	0.0143	0.0406	0.0472	0.0218	0.0190
Secondary	0.0382	0.0451	0.0412	0.0179	0.0099
Technical	0.0259	0.0500	0.0351	0.0441	0.0207
College	0.0500	0.0284	0.0393	0.0500	0.0421
Missing	0.0468	0.0134	0.0045	0.0470	0.0212
Mother's education					
Primary	0.0470	0.0500	0.0092	0.0500	0.0457
Secondary	0.0500	0.0475	0.0338	0.0320	0.0109
Technical	0.0257	0.0089	0.0500	0.0433	0.0274
College	0.0056	0.0398	0.0500	0.0321	0.0325
Missing	0.0328	0.0361	0.0197	0.0384	0.0126
Household income category (in 1000 pesos)					
[0,100)	0.0344	0.0154	0.0500	0.0227	0.0283
[100-200]	0.0414	0.0048	0.0457	0.0361	0.0458
(200-400]	0.0094	0.0418	0.0500	0.0268	0.0020
(400-600]	0.0500	0.0366	0.0500	0.0202	0.0208
(600-1400]	0.0500	0.0406	0.0262	0.0500	0.0458
>1400	0.0064	0.0500	0.0325	0.0500	0.0342
Missing	0.0500	0.0461	0.0500	0.0327	0.0147
Female	0.0500	0.0500	0.0500	0.0500	0.0112

Table 3.7: Estimated effect of each additional year in a private voucher school on PSU test scores

Outcome	$\hat{\beta}_1$	95% C.I.	p
Language	3.85	(1.68,6.02)	5.17e-04
Math	4.83	(2.64,7.02)	1.55e-05

It is important to acknowledge the fact that the assumption of no unobserved confounders may not be completely satisfied. For instance, voucher private schools are allowed to charge additional tuition (copayment) to students. This additional payment that is different for each school could lead schools with higher fees to be more similar to private non-subsidized schools, and schools with lower fees more similar to public schools. These could attract different types of students among private voucher schools but could also affect the quality of the education in each school. It would be interesting to analyze not only the effect of the type of school attended, but also the effect of the total amount of tuition paid by students. However that information was not available to use in this study.

3.7 Summary and concluding remarks

Marginal structural models are a widely used tool to estimate the causal effect of a time-dependent treatment on some outcome variable in the presence of time-dependent confounders. The inverse-probability of treatment weights, when known or modeled correctly, appropriately adjust for these type of confounders without introducing post-treatment bias. However, a main concern with these weights is that they tend to be somewhat unstable, which leads to MSM parameter estimators with large variance.

In this work we introduced a new method to obtain weights for estimation of MSM parameters that are stable and at the same time hold balancing properties. We

studied the role of weighting in the longitudinal setting and what the weighted sample needs to satisfy in order to, in some sense, unveil the distribution of the potential outcomes. This gives some guidance into the type of covariate balance a researcher needs to target to obtain unbiased estimates of the MSM parameters and eventually determine how valid his conclusions are.

A set of simulations were conducted to compare the performance of the proposed method with standard and more recently proposed estimation methods. In these settings, the stable balancing weights outperformed the other methods considered, both in bias and RMSE. This shows estimation of MSM parameters can improve when the specific goal is considered into the process of obtaining the weights, which, in this case, was reduced variability and balance as specified by the researcher. The simulation shows that some of the other methods are indeed unbiased, which is an attractive property for an estimator. However, since in most cases we can only run an experiment or collect data once, we cannot rely on a property that only tells us that in average the estimator will be close to the true value. We need the estimated value to be close for every possible realization. Furthermore, the proposed method can help the researcher decide if the available data is adequate to estimate the MSM that he is using or if further assumptions or a simpler model are needed.

In addition to the simulations, an application on Chilean educational data was presented to show how the proposed method works with real data. The causal question of interest was to determine the effect that each additional year a student spends in a private voucher school has on the student's University Selection Test scores. The estimated effect for each additional year was positive and significant, but small and possibly not relevant in educational policy. However, when we considered the cumulative effect that four years of continuous voucher school education have on PSU test scores, the effect size was close to what some consider a meaningful effect.

Chapter 4

Optimal Weighting for Observational Studies with Multi-Valued Treatments

4.1 Introduction

To estimate causal effects in observational studies when the treatment is binary, researchers typically rely on the propensity score to balance observed covariates, either by matching, weighting, or stratifying. While it has the attractive property of being the coarsest balancing score (Rosenbaum and Rubin, 1983), the propensity score usually has to be estimated and will not yield the desired results if the model is misspecified. Additionally, the propensity score balances covariates on average across repeated samples, not on any given sample. To address these limitations, other methods for binary treatments that directly balance observed covariates have been proposed (Imai and Ratkovic, 2014; Zubizarreta, 2012, 2015).

Generalizing the propensity score to multiple treatments is more complicated and there is no standard way of using it in matching and stratification. Lopez and Gutman (2017) provide a review on estimation methods for causal effects with multiple

treatments and describe methods that go from using the binary propensity score to compare pairs of treatments (Lechner, 2001, 2002) to their proposed method of vector matching. Vector matching is an algorithm in which subjects with similar values on the complete vector of generalized propensity scores are matched using available software. An alternative approach to matching is to use inverse probability of treatment weights (IPTW) (Imbens, 2000). However, the presence of extreme weights that result from conditional treatment assignment probabilities that are very close to zero can generate unstable estimates with large variance. Additionally, as in the binary case, if the probabilities are not modeled correctly, which is more likely when there are more than two treatments, or if the weights are truncated to reduce variability of the estimates, the balancing properties of these weights will not hold.

To address the latter problems that may arise when using inverse probability weights, different methods to estimate the treatment probabilities have been proposed. Imai and Ratkovic (2014) introduced the covariate balancing propensity score (CBPS) in which balance conditions are incorporated in the estimation process within the generalized method of moments or empirical likelihood framework (Fong et al., 2017). Another approach that has been proposed is the use of generalized boosted models (GBM) to estimate the treatment assignment probabilities (McCaffrey et al., 2013, 2004). GBM is a flexible nonparametric model that iteratively fits multiple regression trees and has the advantage that it can incorporate the covariate balance into that process by means of a tuning parameter. Neither of these methods, however, considers in the estimation process the problem of the variability and extremeness of the weights.

In this chapter we derive robust weights for observational studies with multi-valued treatments. In particular, we generalize the stable balancing weights of Zubizarreta (2015) from the binary treatment case to the case in which there are multiple treatment categories. Stable balancing weights (SBW) address the high variability problem while balancing at the same time the observed covariates directly without the need

to estimate treatment assignment probabilities.

The remainder of the chapter is organized as follows. In Section 2 we establish the setup, notation, and assumptions used throughout the chapter. Section 3 describes the proposed method and explains the rationale behind it. Section 4 presents a simulation in which the proposed methodology is compared to other currently available methods and in Section 5 we obtain the proposed weights to estimate the effect that the 2010 earthquake in Chile had at different levels of ground movement intensity on posttraumatic stress disorder. Finally, Section 6 provides concluding remarks and a summary of the work.

4.2 Setup, notation, and assumptions

Let Y_i be the observed outcome in a sample with n independent observations indexed by $i = 1, ..., n$, $\boldsymbol{X}_i$ be the vector containing its P observed pretreatment covariates, and Z_i the treatment assigned to that subject. Using the potential outcomes framework (Holland, 1986; Neyman, 1990; Rubin, 1974, 1978), we define the set of potential outcomes of subject i as $\mathcal{Y}_i = \{Y_i(z), z \in \mathcal{Z}\}$, where $\mathcal{Z}$ is the set of all possible treatment values, and $Y_i(z)$ refers to the outcome that subject i would have observed if he had received treatment z. We will assume that the stable unit treatment value assumption or SUTVA is satisfied (Rubin, 1980, 1986). Only the potential outcome of the actual treatment received will be observed, $Y_i = Y_i(Z_i)$, and the other potential outcomes are counterfactual values. For categorical and ordinal treatments, $|\mathcal{Z}| = K$, with K the total number of treatment levels.

To be able to identify causal effects from observed data, we will assume positivity, and weak ignorability (Imbens, 2000), this is,

$$p(Z_i = z \mid \boldsymbol{X}_i) > 0 \quad \text{and} \quad Y_i(z) \amalg \mathbb{1}_{\{Zi=z\}} \mid \boldsymbol{X}_i \quad \forall z \in \mathcal{Z}. \tag{4.1}$$

The positivity assumption is requiring that every subject in the target population has a positive probability of receiving any given treatment. Without this assumption,

there could be "types" of subjects for which there is no potential outcome information under some treatments and we would not be able to estimate treatment effects without relying on extrapolation. Additionally, in practice, observed subjects with a very low probability of receiving the treatment they received would be very influential in the analysis and the variance of the estimates.

The ignorability assumption means that the treatment assignment is practically random given observed covariates. This is, there are no other variables that are both predictors of treatment assignment and outcome: no unmeasured confounders. Unfortunately, this is not a testable assumption, so subject matter knowledge is needed to ensure that $\boldsymbol{X}_i$ includes the necessary covariates to satisfy it.

4.3 Stable balancing weights for multiple treatments

When there are multiple treatments, we would like to estimate the mean of the potential outcomes distribution under each possible treatment $\mathrm{E}[Y_i(z)]$. This is equivalent to say that we want to estimate the parameters of the marginal structural model $\mathrm{E}[Y_i(z)] = g(z;\boldsymbol{\beta})$ (Robins et al., 2000), where, for categorical and ordinal treatments, g represents the means model

$$g(z_T;\boldsymbol{\beta}) = \sum_{k=1}^{K} \beta_k \mathbb{1}_{\{z=k\}}. \tag{4.2}$$

In randomized studies, by design we have that $Y_i(z)$ II Z_i for all z, and therefore $\mathrm{E}[Y_i \mid Z_i = z] = \mathrm{E}[Y_i(z) \mid Z_i = z] = \mathrm{E}[Y_i(z)]$. This means that we can unbiasedly estimate the mean of each potential outcome by the sample mean of each group. In other words, we can estimate the parameters of the causal model (4.2) by estimating the parameters of the associational model

$$\mathrm{E}[Y_i \mid Z_i = z] = \sum_{k=1}^{K} \gamma_k \mathbb{1}_{\{z=k\}} \tag{4.3}$$

using OLS, for example. When the data in hand does not come from a randomized experiment but we collect enough variables such that the assumption of strong or weak ignorability is plausible, we can still estimate the causal parameters if we use weights to create a pseudo-population in which treatment assignment is independent of the observed covariates. In practice, this can be accomplished by balancing the empirical distributions of the observed covariates across all treatments.

As mentioned before, the typical ways of weighting generally yields highly variable estimates, so, in addition to balancing covariates in order to unbiasedly estimate the causal parameters, it is also desirable that those weights provide the least variable estimator given the desired level of balance.

The weighted least squares estimator for each parameter in model (4.2) is given by

$$\hat{\beta}_k = \frac{\sum_{i=1}^{n} w_i Y_i \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}}, \qquad k = 1, ..., K, \tag{4.4}$$

and its corresponding variance is

$$\text{Var}(\hat{\beta}_k) = \sigma^2 \frac{\sum_{i=1}^{n} w_i^2 \mathbb{1}_{\{Z_i=k\}}}{\left(\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}\right)^2}, \qquad k = 1, ..., K. \tag{4.5}$$

Since the $\hat{\beta}_k$'s are independent, the variance of any linear combination of them is just a positive linear combination of the variances of the $\hat{\beta}_k$'s. From this we can see that, to minimize the variance of the estimators, we need to minimize the sum of squares of the weights, given the total sum of weights for each treatment.

Now, how should the pseudo-population look like in order to estimate the causal parameter of interest without bias? Let $\beta_k^S = \frac{1}{n}\sum_{i=1}^{n} Y_i(k)$ be the (unobserved) average potential outcome in the sample under treatment k. If we have the conditional mean model

$$\text{E}[Y_i(z) \mid \boldsymbol{X}_i] = \sum_{k=1}^{K} h_k(\boldsymbol{X}_i)\mathbb{1}_{\{z=k\}}, \tag{4.6}$$

then, in terms of $\boldsymbol{X}_i$, the parameters in model (4.2) are given by $\beta_k = \mathrm{E}[h_k(\boldsymbol{X}_i)]$. Hence,

$$\begin{aligned}
\left|\mathrm{E}\left[\hat{\beta}_k - \beta_k^S \mid \boldsymbol{X}, Z\right]\right| &= \left|\mathrm{E}\left[\frac{\sum_{i=1}^n w_i Y_i \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^n w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^n Y_i(k) \;\middle|\; \boldsymbol{X}, Z\right]\right| \\
&= \left|\frac{\sum_{i=1}^n w_i \,\mathrm{E}[Y_i(k) \mid \boldsymbol{X}_i, \mathbb{1}_{\{Z_i=k\}}]\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^n w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^n \mathrm{E}[Y_i(k) \mid \boldsymbol{X}_i, \mathbb{1}_{\{Z_i=k\}}]\right| \\
&= \left|\frac{\sum_{i=1}^n w_i h_k(\boldsymbol{X}_i)\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^n w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^n h_k(\boldsymbol{X}_i)\right|.
\end{aligned} \tag{4.7}$$

This means that the closer we force $\frac{\sum_{i=1}^n w_i h_k(\boldsymbol{X}_i)\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^n w_i \mathbb{1}_{\{Z_i=k\}}}$ to be to $\frac{1}{n}\sum_{i=1}^n h_k(\boldsymbol{X}_i)$, the smaller the bias will be. If we assume $h_k(\boldsymbol{x})$ is linear in the covariates, then this simplifies to forcing $\frac{\sum_{i=1}^n w_i X_{ip} \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^n w_i \mathbb{1}_{\{Z_i=k\}}}$ to be close to $\frac{1}{n}\sum_{i=1}^n X_{ip}$ for all $p = 1, ..., P$, where X_{ip} is the p-th observed covariate for subject i.

Explicitly, if $h_k(\boldsymbol{x}_i) = \alpha_0^k + \sum_{p=1}^P \alpha_p^k x_{ip}$ and we impose a maximum distance between the previous quantities of $\delta_p > 0$, that is, $\left|\frac{\sum_{i=1}^n w_i X_{ip} \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^n w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^n X_{ip}\right| < \delta_p$ for all $p = 1, ..., P$, we can easily see that the absolute bias of $\hat{\beta}_k$ with respect to β_k^s will be bounded by $\sum_{p=1}^P \delta_p |\alpha_p|$:

$$
\begin{aligned}
\left|\mathrm{E}\Big[\hat{\beta}_k - \beta_k^S \;\Big|\; Z = z\Big]\right| &\leq \mathrm{E}\left[\left|\mathrm{E}\Big[\hat{\beta}_k - \beta_k^S \;\Big|\; \boldsymbol{X}, Z\Big]\right| \;\middle|\; Z = z\right] \\
&= \mathrm{E}\left[\left|\frac{\sum_{i=1}^{n} w_i \left(\alpha_0^k + \sum_{p=1}^{P} \alpha_p^k X_{ip}\right) \mathbb{1}_{\{Z_i = k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i = k\}}}\right.\right. \\
&\qquad \left.\left. - \frac{1}{n}\sum_{i=1}^{n}\left(\alpha_0^k + \sum_{p=1}^{P} \alpha_p^k X_{ip}\right)\right| \;\middle|\; Z = z\right] \qquad (4.8) \\
&= \mathrm{E}\left[\left|\sum_{p=1}^{P} \alpha_p^k \left(\frac{\sum_{i=1}^{n} w_i X_{ip} \mathbb{1}_{\{Z_i = k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i = k\}}} - \frac{1}{n}\sum_{i=1}^{n} X_{ip}\right)\right| \;\middle|\; Z = z\right] \\
&\leq \sum_{p=1}^{P} \delta_p |\alpha_p^k|,
\end{aligned}
$$

where conditioning on $Z = z$ means conditioning on the observed treatment assignments in the sample.

It is important to notice that this is a bound for the absolute bias with respect to the corresponding sample average potential outcome. The bias with respect to the target population will depend on how well the sample represents that population. To avoid additional sampling bias, we need either to sample $\boldsymbol{X}_i$'s from the marginal distribution of $\boldsymbol{X}_i$ in the target population or to know the marginal treatment assignment distribution in the target population.

When h_k's are not linear, we could balance auxiliary covariates to approximately balance the whole distribution of the observed covariates. See Zubizarreta (2015) and Chapter 2 for a discussion on how to balance non-linear functions.

Having established the conditions we would like the weights to satisfy, we solve the following convex optimization problem to find the weights that minimize the variance of the estimator given imbalance bounds and total sum of weights determined by the

researcher

$$
\begin{aligned}
\underset{\boldsymbol{w}}{\text{minimize}} \quad & \|\boldsymbol{w}\|_2^2 \\
\text{subject to} \quad & \left| \frac{\sum_{i=1}^{n} w_i X_{ip} \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n} \sum_{i=1}^{n} X_{ip} \right| \leq \delta_p \qquad \begin{matrix} k=1...,K \\ p=1,...,P \end{matrix} \\
& \sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Z_i=k\}} \\
& w_i \geq 0 \qquad i = 1, ..., n,
\end{aligned}
\tag{4.9}
$$

where $\boldsymbol{w}$ is the n-dimensional vector of weights w_i's and δ_p is a scalar determined by the researcher that represents the desired level of covariate balance.

Alternatively, instead of including the bias term as restrictions, we can incorporate them in the objective function. Doing this would imply sacrificing balance control in favor of ensuring feasibility

$$
\begin{aligned}
\underset{\boldsymbol{w},p}{\text{minimize}} \quad & \lambda \|\boldsymbol{w}\|_2^2 + (1-\lambda) \left\| \frac{\sum_{i=1}^{n} w_i X_{ip} \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n} \sum_{i=1}^{n} X_{ip} \right\|_\infty^2 \\
\text{subject to} \quad & \sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Z_i=k\}} \\
& w_i \geq 0 \qquad i = 1, ..., n.
\end{aligned}
\tag{4.10}
$$

where $\lambda \in (0, 1)$ is a tuning parameter that can be chosen by cross validation (Athey et al., 2016) and explicitly represents the trade-off between bias and variance.

In both representations of the problems we are suggesting fixing the sum of weights for each treatment group to be equal to the sample proportion of subjects in each group, giving a total sum of weights of one. Roughly speaking, this is analogous to the use of the stabilized version of the IPTW.

4.4 Simulation study

In this section we conduct a simulation study with two different settings to compare the performance of the proposed weighting method with the standard approach of

using the generalized propensity score (Hirano and Imbens, 2005; Imbens, 2000) to construct these weights. Three methods are used to estimate the generalized propensity score: maximum likelihood, CBPS, and GBM.

4.4.1 Data generating mechanisms

4.4.1.1 Categorical treatment

The first simulation setting was based on one of the designs in Yang et al. (2016). Here the treatment variable had three levels and there were six covariates that were included in the outcome model, as well as in the treatment assignment model. Specifically, $\boldsymbol{X_i} = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})$ was generated as $X_{i1}, X_{i2}, X_{i3} \sim \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0,0,0)^\top$, $\sigma_{jj} = 2,1,1$ for $j = 1,2,3$ respectively, $\sigma_{12} = \sigma_{21} = 1$, $\sigma_{13} = \sigma_{31} = -1$, and $\sigma_{23} = \sigma_{32} = -0.5$; $X_{i4} \sim U[-3,3]$; $X_{i5} \sim \chi^2_1$; and $X_{i6} \sim$ Bernoulli(0.5). The treatment variable Z_i with three levels was constructed from the treatment indicators $(\mathbb{1}_{\{Z_i=1\}}, \mathbb{1}_{\{Z_i=2\}}, \mathbb{1}_{\{Z_i=3\}}) \sim$ Multinom$(p(1|\boldsymbol{X_i}), p(2|\boldsymbol{X_i}), p(3|\boldsymbol{X_i}))$, where $p(z|\boldsymbol{X_i}) = \frac{exp(X_i^\top \eta_z)}{\sum_{\zeta=1}^{3} exp(X_i^\top \eta_\zeta)}$, with $\eta_1 = (0,0,0,0,0,0)^\top$, $\eta_2 = 0.7 \times (1,1,1,-1,1,1)^\top$, and $\eta_3 = 0.4 \times (1,1,1,1,1,1)^\top$. The outcome was given by

$Y_i(1) \mid \boldsymbol{X_i} = -1.5 + X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5} + X_{i6} + \varepsilon_i$

$Y_i(2) \mid \boldsymbol{X_i} = -3 + 2X_{i1} + 3X_{i2} + X_{i3} + 2X_{i4} + 2X_{i5} + 2X_{i6} + \varepsilon_i$

$Y_i(3) \mid \boldsymbol{X_i} = 1.5 + 3X_{i1} + 1X_{i2} + 2X_{i3} - X_{i4} - X_{i5} - X_{i6} + \varepsilon_i,$

where $\varepsilon_i \sim \mathcal{N}(0,1)$. With this generating mechanism, the true dose-response function was $\mathrm{E}\left[Y_i(z)\right] = \mathrm{E}\left[\mathrm{E}\left[Y_i(z) \mid \boldsymbol{X_i}\right]\right] = 0$ for all $z = 1,2,3$, and therefore the average effect was zero when comparing any two treatment levels.

To compare the methods' performance under misspecification of the outcome and treatment models, we simulated data under a second scenario. Here we introduced non-linearities in both models by using the transformed covariates $\boldsymbol{X'_i}$, where $X'_{i2} = sign(X_{i2}) \times |X_{i2}|^{\frac{1}{2}}$, $X'_{i5} = \frac{1}{exp(X_{i5})}$, and $X'_{ij} = X_{ij}$ for $j = 1,3,4,6$ as the observed covariates used to obtain the weights. The true treatment effect remained the same. We generated a total of 1000 replications for both scenarios, each one with a sample

size of $n = 1500$.

4.4.1.2 Continuous treatment

The goal of including this second simulation setting was to evaluate the performance of the proposed method when the treatment is continuous, comparing it with the continuous version, when available, of the same methods as before. For this, we generated data as in Fong et al. (2017) and used two of the scenarios studied in their simulation: both treatment assignment and outcome models linear in the covariates, and both having non-linearities. This design considered covariates that were not related to the outcome and/or the treatment assignment. Specifically, $\boldsymbol{X_i} = (X_{i1}, ..., X_{i10}) \sim \mathcal{N}_{10}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\sigma_{jj} = 1$ for $j = 1, ..., 10$ and $\sigma_{jq} = 0.2$ for $j \neq q$. In the first scenario, outcome and treatment were given by:

$Z_i = X_{i1} + X_{i2} + 0.2X_{i3} + 0.2X_{i4} + 0.2X_{i5} + \epsilon_i$

$Y_i(z) \mid \boldsymbol{X_i} = X_{i2} + 0.1X_{i4} + 0.1X_{i5} + 0.1X_{i6} + z + \varepsilon_i,$

where $\epsilon_i \sim \mathcal{N}(0, 9)$ and $\varepsilon_i \sim \mathcal{N}(0, 25)$. This means that the true dose-response function was $\mathrm{E}\left[Y_i(z)\right] = \mathrm{E}\left[\mathrm{E}\left[Y_i(z) \mid \boldsymbol{X_i}\right]\right] = z$ and thus, the value of the relevant parameter was one.

In the second scenario,

$Z_i = (X_{i2} + 0.5)^2 + 0.4X_{i3} + 0.4X_{i4} + 0.4X_{i5} + \epsilon_i$

$Y_i(z) \mid \boldsymbol{X_i} = 2(X_{i2} + 0.5)^2 + 0.5X_{i4} + 0.5X_{i5} + 0.6X_{i6} + z + \varepsilon_i.$

Here, the true dose-response function was $\mathrm{E}\left[Y_i(z)\right] = \mathrm{E}\left[\mathrm{E}\left[Y_i(z) \mid \boldsymbol{X_i}\right]\right] = 2.5 + z$ so the relevant parameter remained equal to one. Again, 1000 replications were generated with sample size $n = 1500$.

4.4.2 Estimation methods

The first method we used to estimate the generalized propensity score is GLM. In the categorical treatment setting, we fit the linear multinomial logistic regression model with the `R` function `multinom` from the `nnet` package, while in the continuous

treatment setting, we fit the linear model with the `lm` function. Once the necessary probabilities were estimated, the stabilized versions of the weights were computed. The second method is CBPS and we used the `R` function `CBPS` from the `CBPS` package (Fong et al., 2016). The weights were obtained directly from the function output. The third method is GBM and it was fit using the function `mnps` from the `twang` package (Ridgeway et al., 2016) and the corresponding weights were obtained applying the `get.weights` function to the output from `mnps`. For these methods, the model included the covariates linearly, which in the first scenario in each setting means that the model was correctly specified and in the second scenario, misspecified. All the functions used were set to their default values.

For the stable balancing weights, the weights were obtained solving the version of the optimization problem that includes the covariate balance in the objective function with $\lambda = \frac{1}{2}$. For the first scenario in both simulation settings, only mean balance was required, while in the second scenario, both mean and fine balance with ten categories were required.

Since `twang` package does not include an option for a continuous treatment and our method is designed to accommodate categorical treatments, the treatment in the second setting was approximately balanced by dividing it into a categorical variable with ten levels created from the treatment deciles. Additionally, the naive estimator where no weights were used was also included in the comparison as a measure of original bias.

In the first setting, the estimated quantities were $\tau_{1,2} = \mathrm{E}[Y_i(2) - Y_i(1)]$, $\tau_{1,3} = \mathrm{E}[Y_i(3) - Y_i(1)]$, and $\tau_{2,3} = \mathrm{E}[Y_i(3) - Y_i(2)]$, and in the second, β_0 and β_1 from the MSM $\mathrm{E}[Y_i(z)] = \beta_0 + \beta_1 z$. The methods were compared by the bias and root mean square error (RMSE) of the estimators of these parameters, as well as the average length and coverage of the 95% confidence intervals obtained using the robust sandwich variance estimator.

4.4.3 Results

4.4.3.1 Categorical treatment

Figure 4.1 shows the distributions of the estimated treatment effects across repetitions for each method and Table 4.1 their bias and RMSE. The dashed line denotes the true parameter value. There we can see that the estimated treatment effect of treatment 3 over treatment 1 was not heavily biased in the unweighted sample and non of the methods exacerbated this bias. However, given the instability of the GLM weights, the variability of its corresponding estimator was increased considerably, resulting in a RMSE twice as large as the one of the unweighted (Unwt) estimator. For the other two treatment effects where the unweighted estimator provided more biased results, GBM improved in terms of bias and RMSE, but not quite as much as the other methods. In terms of bias, GLM performed better than CBPS, but the larger variance represented larger RMSE. SBW provided the lowest bias and RMSE on all three parameter estimators. These results are transferred to the confidence interval performance (Table 4.2). As we can see, the actual coverage of the intervals obtained using GBM reached values as low as 56.80%. On the other hand, the SBW interval covered the true parameter value in every simulation repetition, while preserving the shortest lengths. The interval coverage for GLM and CBPS was closer to the nominal 95%.

Figure 4.1: Boxplots of the estimated parameters in the first setting, first scenario

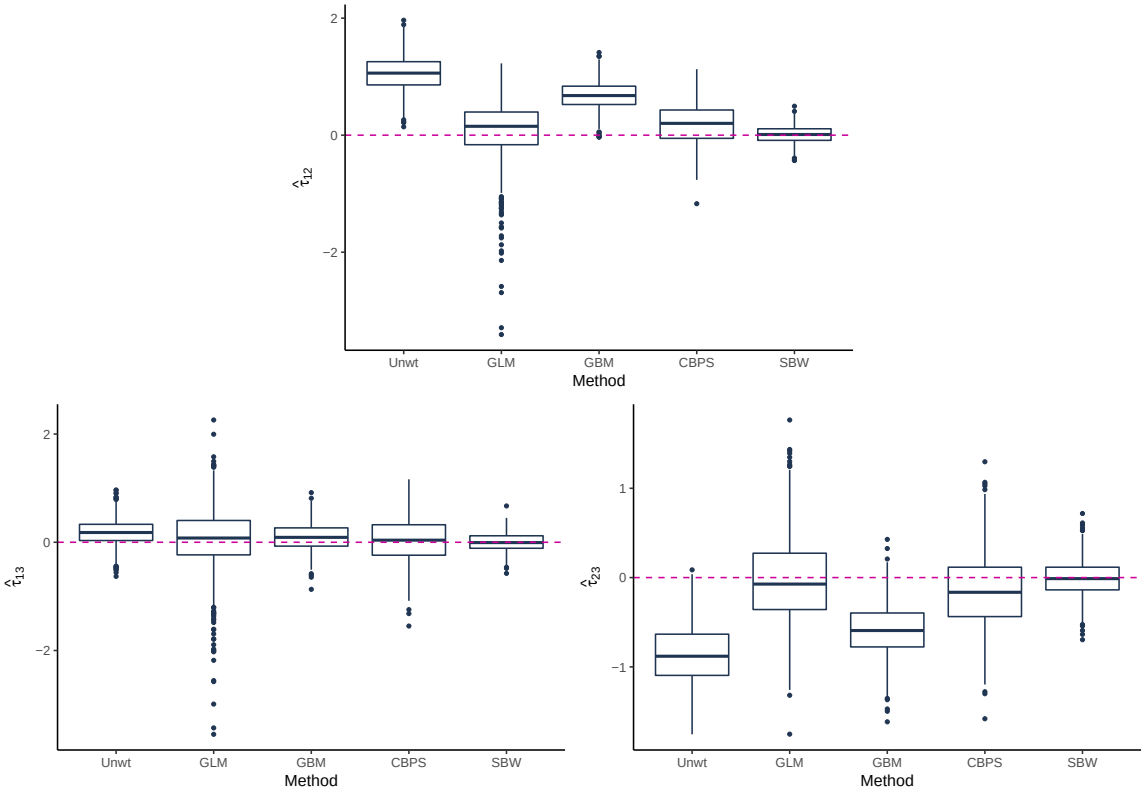


Table 4.1: Bias and RMSE of the estimated parameters in the first setting, first scenario

Method	$\hat{\tau}_{1,2}$		$\hat{\tau}_{1,3}$		$\hat{\tau}_{2,3}$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unwt	1.05	1.09	0.18	0.30	−0.87	0.93
GLM	0.08	0.51	0.05	0.59	−0.02	0.48
GBM	0.67	0.71	0.09	0.27	−0.58	0.65
CBPS	0.19	0.39	0.03	0.41	−0.16	0.43
SBW	0.01	0.14	0.00	0.17	−0.01	0.20

Table 4.2: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the first setting, first scenario

Method	$\tau_{1,2}$		$\tau_{1,3}$		$\tau_{2,3}$	
	Coverage	Length	Coverage	Length	Coverage	Length
Unwt	14.40	1.48	98.90	1.57	22.80	1.23
GLM	97.90	2.06	97.80	2.20	97.80	2.23
GBM	56.80	1.45	98.90	1.35	75.20	1.57
CBPS	97.00	1.88	98.50	1.96	97.40	2.06
SBW	100.00	1.44	100.00	1.30	100.00	1.48

Figure 4.2 and Tables 4.3 and 4.4 show the results for the scenario with model misspecification in this setting. The general trend of these was similar to the previous scenario, except for GBM. The flexibility of these models made them robust to model misspecification, providing practically the same results when the original covariates were observed or when only a non linear transformation of them was observed. Nonetheless, the bias and RMSE for this method were still the largest for $\hat{\tau}_{1,2}$ and $\hat{\tau}_{2,3}$, which were the ones with larger bias on the unweighted sample. The other three methods behaved similarly among each other with respect to original bias, obtaining better results in bias, RMSE, and coverage with SBW than with the other two methods.

Figure 4.2: Boxplots of the estimated parameters in the first setting, second scenario

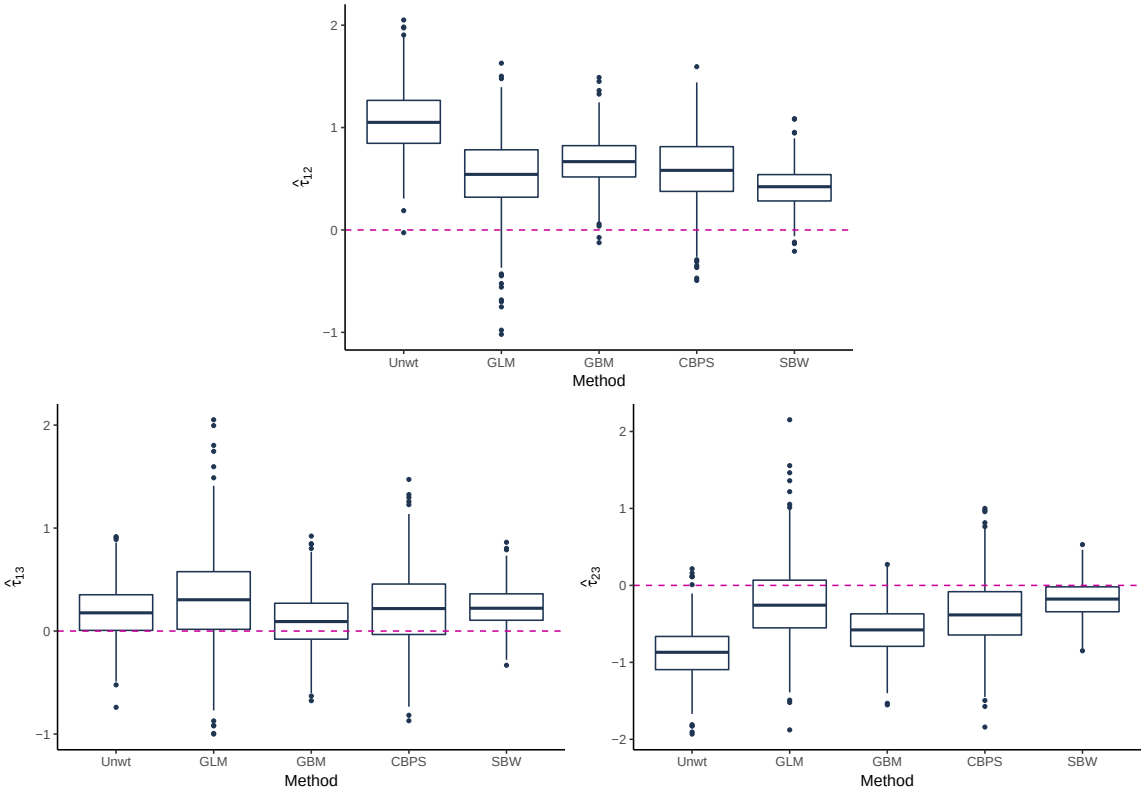


Table 4.3: Bias and RMSE of the estimated parameters in the first setting, second scenario

Method	$\hat{\tau}_{1,2}$		$\hat{\tau}_{1,3}$		$\hat{\tau}_{2,3}$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unwt	1.06	1.10	0.18	0.31	−0.88	0.93
GLM	0.54	0.66	0.31	0.52	−0.24	0.54
GBM	0.67	0.71	0.10	0.28	−0.57	0.65
CBPS	0.59	0.68	0.22	0.42	−0.37	0.56
SBW	0.42	0.46	0.23	0.30	−0.18	0.30

Table 4.4: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the first setting, second scenario

Method	$\tau_{1,2}$		$\tau_{1,3}$		$\tau_{2,3}$	
	Coverage	Length	Coverage	Length	Coverage	Length
Unwt	14.30	1.48	99.30	1.57	20.50	1.23
GLM	82.00	1.85	96.40	1.89	94.70	2.19
GBM	60.60	1.45	98.50	1.34	73.50	1.57
CBPS	77.40	1.75	97.70	1.74	91.80	2.04
SBW	99.80	1.92	99.70	1.53	100.00	2.05

4.4.3.2 Continuous treatment

The results for the setting with a continuous outcome in the correctly specified scenario are presented in Figure 4.3 and Tables 4.5 and 4.6. The general observations are similar to those in the previous setting. Bias and RMSE were lowest with SBW, followed by both GLM and CBPS, with lower bias but higher variance for GLM, and the most biased estimator was the one that used the weights obtained using GBM. Interval coverage was close to the nominal 95% on all methods except for GBM, providing SBW the second largest coverage with the shortest lengths.

In the second scenario of this setting, where treatment probability and outcome models were misspecified, non of the methods was able to provide the desired results (Figure 4.4, Table 4.7, 4.8). GLM and CBPS were the most sensitive to model misspecification, while the flexibility of GBM and the fine balance requirement of SBW allowed them to perform slightly better. Bias and RMSE were comparable among these two methods, however, the GBM 95% confidence intervals covered the true parameter value only on 4.80% of the repetitions, while SBW had a – still low – 14.90% coverage. It is worth noting that these last two methods were balancing covariates on the 10-levels categorical version of the treatment variable, and not

directly on the continuous treatment as the other two methods.

Figure 4.3: Boxplots of $\hat{\beta}_1$ in the second setting, first scenario

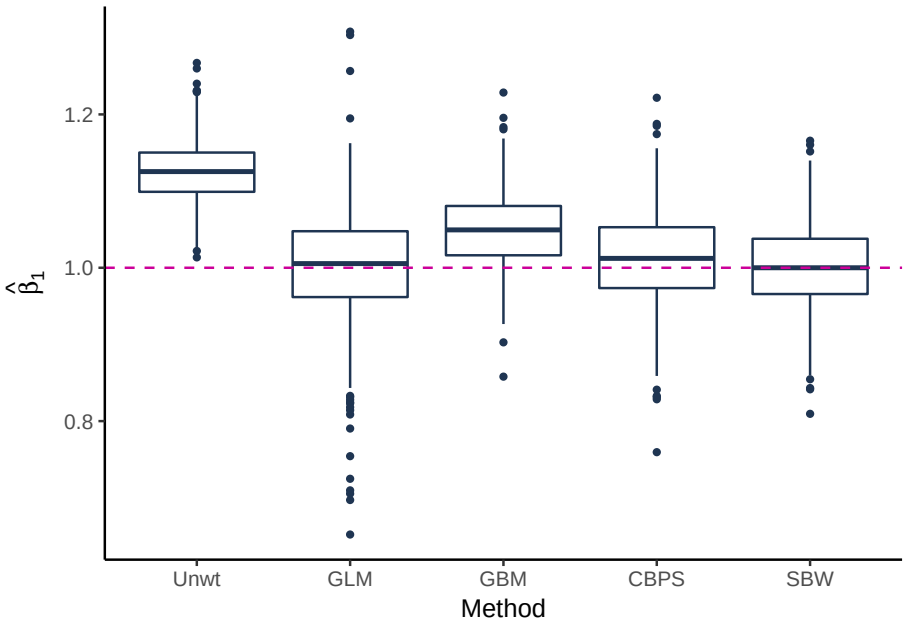


Table 4.5: Bias and RMSE of the estimated parameters of the MSM in the second setting, first scenario

Method	$\hat{\beta}_0$		$\hat{\beta}_1$	
	Bias	RMSE	Bias	RMSE
Unwt	−1.00	1.01	0.12	0.13
GLM	−1.00	1.02	0.00	0.07
GBM	−0.99	1.01	0.05	0.07
CBPS	−1.00	1.01	0.01	0.06
SBW	−1.00	1.01	0.00	0.05

Table 4.6: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the second setting, first scenario

Method	β_0		β_1	
	Coverage	Length	Coverage	Length
Unwt	95.20	0.52	9.00	0.15
GLM	94.60	0.68	93.90	0.25
GBM	95.00	0.63	84.80	0.19
CBPS	94.30	0.64	92.90	0.22
SBW	93.90	0.59	93.60	0.20

Figure 4.4: Boxplots of $\hat{\beta}_1$ in the second setting, second scenario

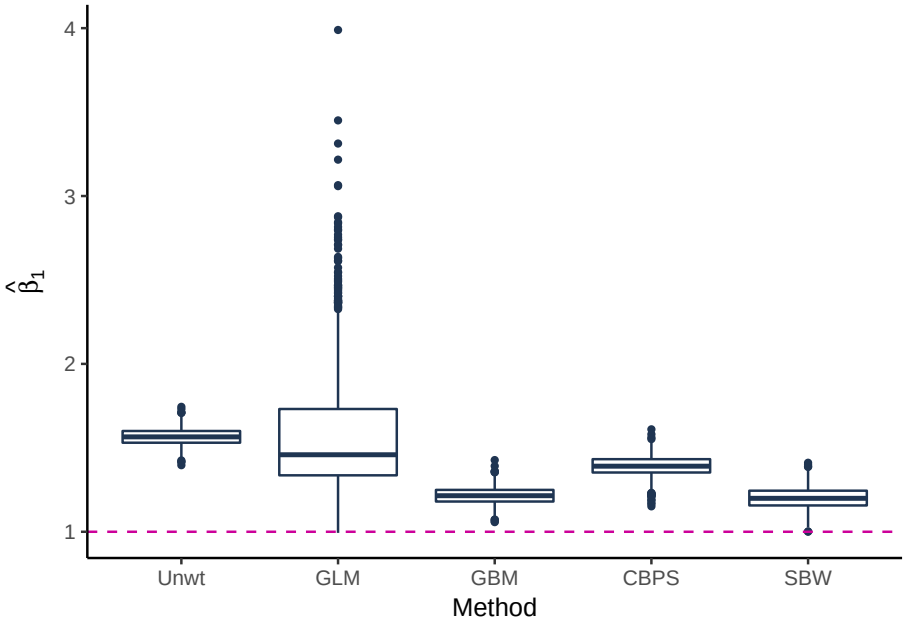


Table 4.7: Bias and RMSE of the estimated parameters of the MSM in the second setting, second scenario

Method	$\hat{\beta}_0$		$\hat{\beta}_1$	
	Bias	RMSE	Bias	RMSE
Unwt	0.79	0.80	0.57	0.57
GLM	0.99	1.02	0.59	0.70
GBM	0.77	0.79	0.22	0.23
CBPS	0.74	0.76	0.39	0.40
SBW	0.95	0.97	0.20	0.21

Table 4.8: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the second setting, second scenario

Method	β_0		β_1	
	Coverage	Length	Coverage	Length
Unwt	0.00	0.63	0.00	0.17
GLM	0.10	0.81	4.80	0.55
GBM	0.00	0.72	2.30	0.22
CBPS	0.00	0.64	0.10	0.21
SBW	0.00	0.72	14.90	0.26

These simulations seem to indicate that GLM and CBPS are more sensitive to probability model misspecifications, giving close to unbiased, although variable, estimates when correctly specified but largely biased ones when not. Contrary to this, GBM seems to perform comparatively better than the previous methods when the transformed covariates were used, but seems to fail to balance covariates correctly even when the correct model is assumed. SBW was able to outperform the rest of the methods in these simulation settings, both in bias and variance reduction, even though the difference was small in the misspecified scenario of the second setting.

In this case, fine balance with finer categories could be required in the optimization problem or the tuning parameter λ could be changed to try to improve the results.

4.5 Case study: 2010 earthquake effects on post-traumatic stress in Chile

To illustrate the use of the method both on a multi-valued categorical treatment and a continuous treatment, we use available data before and after the 2010 earthquake in Chile. This magnitude 8.8 earthquake took place on February 27, 2010, shortly after the 2009 national socioeconomic survey (CASEN) had been completed. Being an earthquake of such a large magnitude, followed by a tsunami, the catastrophic event provoked economic losses of $30 billion USD, including the destruction of houses, schools, and hospitals, and killed more than 500 people (USGS, 2011b).

Survivors of disasters commonly experience mental health consequences (Norris et al., 2002). One of the most frequently observed and studied effects is posttraumatic stress disorder (PTSD). Studies suggest that the severity and type of symptoms that survivors present may differ due to the intensity of the earthquake experienced, as well as to socioeconomic and cultural backgrounds (USDVA, 2016).

To study the impact the earthquake had on the affected population, the Ministry of Planning (MIDEPLAN) reinterviewed between May and June of 2010 a subsample of 22,456 households from the 71,460 households interviewed for CASEN in 2009. This postearthquake survey provides the opportunity to study posttraumatic stress symptoms well before they disappear, while also providing, in conjunction with CASEN data, a set of pre-exposure variables for the same subjects without introducing recall bias.

While Zubizarreta et al. (2013) use matching to compare, in terms of PTSD, residents of areas that experienced high degree of shaking to those that were similar to them but lived in areas that were almost untouched by the earthquake, in this study

we evaluate the effect of the earthquake on PTSD as a function of the earthquake intensity. This intensity was measured using peak ground acceleration values from the United States Geological Survey (USGS, 2011a, 2014). Two versions of the treatment were considered, intensity as a categorical variable with low [0,0.08), medium [0.08,0.25), and high levels ($\geq$0.25), and as a continuous variable using directly the observed peak ground acceleration.

4.5.1 Data and design

The outcome used to assess PTSD level was the score obtained from the self-rated Davidson Trauma Scale (Davidson et al., 1997) questions included in the postearthquake survey. These included two items rated on a five-point scale for each one of the 17 symptoms of PTSD from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. Of each pair of items, one referred to frequency (1 = "not at all" to 5 = "every day") and the other referred to severity (1 = "not at all distressing" to 5 = "extremely distressing"). Adding the responses of every item, the total score ranged from 34 to 170.

Table 4.9 shows the pre-earthquake covariates that were balanced. Mean balance with an absolute standardized difference in means lower than 0.001 for the treatment with three levels and 0.005 for the continuous treatment was required for all the covariates. For continuous covariates (Age, Household size, Education, Individual work income, Household own per capita income, and Household total per capita income) fine balance restrictions of a 10-levels categorical version of the variables were included in addition to the mean balance ones. Subjects for which no outcome or treatment variables were available were excluded from the study, leaving a sample size of $n = 23,322$.

Since the categorical version of the treatment is an ordinal variable, the parameters of interest to be estimated were $\tau_{1,2} = \mathrm{E}[Y_i(2) - Y_i(1)]$ and $\tau_{2,3} = \mathrm{E}[Y_i(3) - Y_i(2)]$ for the categorical version of the treatment, and β_0 and β_1 from the model $\mathrm{E}[Y_i(z)] =$

$\beta_0 + \beta_1 z$ for the continuous version, where Y_i represents the PTSD score of subject i. Again, balance for the continuous treatment was approximated by transforming the continuous variable into a 10-levels categorical variable constructed using deciles. 95% confidence intervals were obtained using the robust sandwich variance estimator.

4.5.2 Results

Figure 4.5 summarizes the balance of the pre-earthquake covariates before and after weighting in the categorical setting. It shows the absolute standardized differences in means of the 45 covariates in the three treatment groups. We can see that even though the unweighted sample was not drastically unbalanced, there were several instances in which the absolute standardized differences in means were well beyond acceptable levels. The stable balancing weights were able to obtain a very tight mean balance for all covariates.

Having achieved that level of balance, the treatment effects and their corresponding 95% confidence intervals were estimated (Table 4.10). We can see that experiencing the earthquake with the second intensity level increased on average the PTSD score by 7.70 points compared to the lowest intensity level. The increase is slightly higher when changing from the second category to the third one, with an estimated average increase of 8.27 points on the PTSD score. This means that the average accumulated increase when changing from the first to the third level would be 16.07 points.

Figure 4.6 shows a summary of the covariate mean balance before and after weighting for the case where the treatment was transformed into a 10-levels categorical variable. With ten treatment categories, the pre-earthquake covariates had more imbalances across treatment levels. These were practically removed by weighting the sample with the SBW obtained. Table 4.11 shows the estimated treatment effect and its corresponding 95% confidence interval. Each additional intensity unit on the earthquake has an estimated average effect of 60.88 points on the PTSD score. The

Table 4.9: Pre-earthquake covariates

Demographic covariates	Housing structure
Age (years)	Acceptable
Women	Reparable
Indigenous ethnic group	Irreparable
Household size	Overcrowding
Marital status	No
Married or cohabitating	Medium
Divorced or widow	Critical
Singe	Health before the earthquake
Socioeconomic covariates	Health problem (last month)
Education (years)	Hospitalized (last year)
Employment status	Has a psychiatric problem
Employed	Self-rated health
Unemployed	Poor
Inactive	Fair
Individual work income (1000 pesos)	Good
Household own per capita income (1000 pesos)	Missing
Household total per capita income (1000 pesos)	Health insurance
Poor	Public (FONASA)
Housing before the earthquake	Private (ISAPRE)
Housing status	Other
Own housing or paying to own it	No
Rented housing	Unknown
Ceded housing	Disability
Irregular use of housing	Self-sufficient or low
Housing rent per year (pesos)	Moderate or severe
0-25,000	No
25,001-50,000	Other
50,001-75,000	Rural zone
>75000	

range of the earthquake intensity variable was 0-0.32, with a mean value of 0.2 and a standard deviation of 0.1. This represents a maximum average increase of 20 points on the PTSD score for that particular earthquake.

Figure 4.5: Boxplots of absolute standardized differences in means before and after weighting for the treatment with three categories

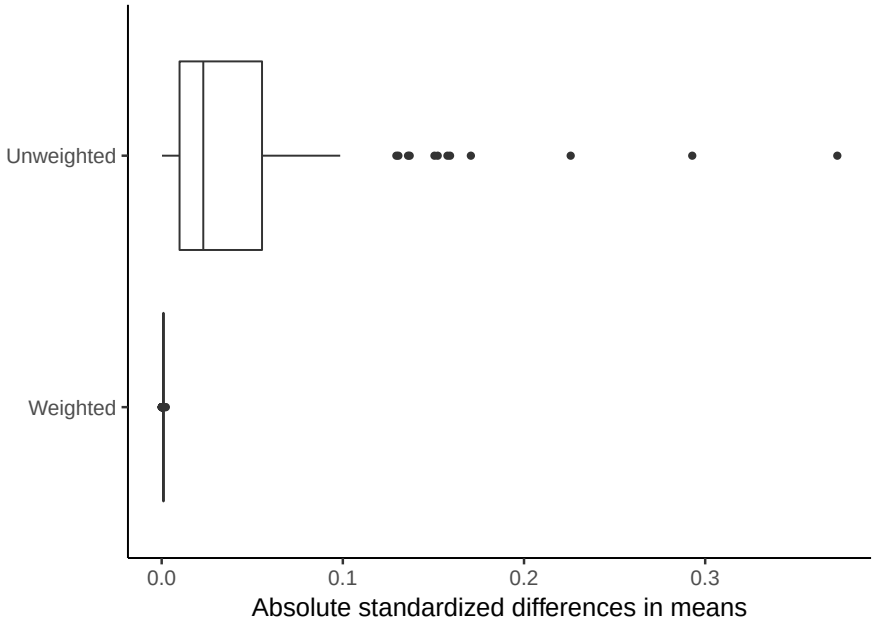


Table 4.10: Estimated effect between incremental earthquake intensity categories on posttraumatic stress

	$\tau_{1,2}$	$\tau_{2,3}$
Estimate	7.70	8.37
95% C.I.	(6.95,8.45)	(7.49,9.26)

Figure 4.6: Boxplots of absolute standardized differences in means before and after weighting for the continuous treatment

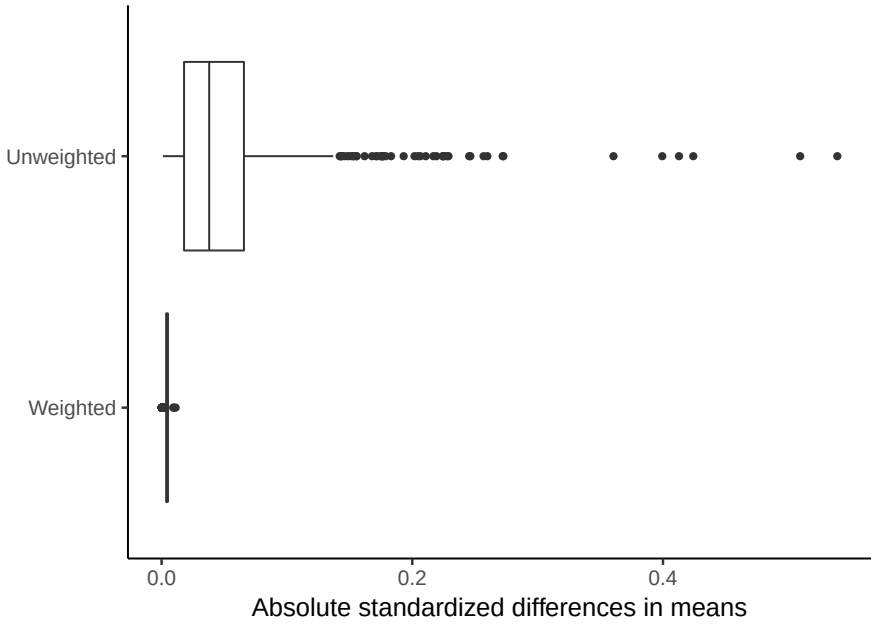


Table 4.11: Estimated effect of earthquake intensity as a continuous variable on posttraumatic stress

	β_0	β_1
Estimate	3.37	60.88
95% C.I.	(2.86,3.88)	(58.04 ,63.72)

Given that the original treatment variable was continuous, the last results provide a more detailed behavior of the effect of earthquake intensity on PTSD. However, when we use less categories, we get larger bins and therefore more observations are assigned to each category, which allows to achieve better balance. For this reason, the approach employed should depend on the real question of interest.

4.6 Summary and concluding remarks

The problem of estimating causal effects in observational studies with a binary treatment has been studied for many years, and the methods available are widely used in applications. Some advances have been made in generalizing these methods to the multiple treatment case, however, it is still common practice to dichotomize the treatment and analyze it using methods designed for binary treatments.

One method that is naturally extended to the multiple treatment case is weighting observations to address confounding, and this has typically be done by using the IPTW. The main drawback of these methods is the large variability in the estimates that result from having extreme weights. Additionally, these weights can fail to balance covariates when the probability model is misspecified, when samples are small, or when extreme weights are truncated.

In this chapter we extended the stable balancing weights of Zubizarreta (2015) to the case with multiple treatments, including a suggestion on how to use them in the continuous case. We determined the conditions the weights need to satisfy in order to provide close to unbiased treatment effect estimates with a reduced variability and defined the convex optimization problem that can be solved to obtain them.

A simulation study was implemented to compare the proposed method with the most common way to estimate the generalized propensity score (GLM) and other recent methods that incorporate the covariate balance in the estimation process in different ways (GBM and CBPS). The simulation included two different settings, one with a categorical treatment with three categories and one with a continuous treatment. Both settings were analyzed in a scenario where the probability and/or outcome model were correctly specified and a scenario where they were misspecified.

From this simulation we found that the flexibility of GBM made it more robust to model misspecification, but its performance depended directly on how biased was the unweighted data. Results from CBPS were similar to the ones obtained when estimating the generalized propensity score with GLM, with some reduction in variability

and bias in some instances. SBW produced the least biased and variable estimates in practically every setting, even when the treatment was continuous, for which these weights were not specifically designed.

Finally, we applied the proposed method to Chilean data available from before and after the 2010 earthquake to estimate the effect that different levels of earthquake intensity have on posttraumatic stress disorder. The earthquake intensity was considered both as a multi-valued categorical variable and a continuous variable. In both cases, it was concluded that there was a statistically significant positive effect of the intensity level of the earthquake on posttraumatic stress.

Bibliography

Ahuja, R., Magnanti, T., and Orlin, J. (1993). *Network Flows: Theory, Algorithms and Applications*. Prentice Hall, New Jersey.

Anand, P., Mizala, A., and Repetto, A. (2009). Using school scholarships to estimate the effect of government subsidized private education on academic achievement in chile. *Economics of Education Review*, 28:370–381.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Athey, S., Imbens, G. W., and Wager, S. (2016). Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. *ArXiv e-prints*.

Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and monte carlo simulations. *Biometrical Journal*, 51(1):171–184.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6):1057–1069.

Baiocchi, M. (2011). Designing robust studies using propensity score and prognostic score matching. *Chapter 3 in Methodologies for Observational Studies of Health Care Policy, Dissertation, Department of Statistics, The Wharton School, University of Pennsylvania*.

Bertsekas, D. P. (1981). A new algorithm for the assignment problem. *Mathematical Programming*, 21:152–171.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to Linear Optimization*. Athena Scientific, Dynamic Ideas.

Burkard, R., Dell'Amico, M., and Martello, S. (2009). *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Cochran, W. and Rubin, D. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Cochran, W. G., Moses, L. E., and Mosteller, F. (1983). *Planning and analysis of observational studies*. Wiley, New York.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Davidson, J. R., Book, S. W., Colket, J. T., Tupler, L. A., Roth, S., David, D., Hertzberg, M., Mellman, T., Beckham, J. C., Smith, R. D., Davison, R. M., Katz, R., and Feldman, M. E. (1997). Assessment of a new self-rating scale for post-traumatic stress disorder. *Psychological Medicine*, 27(1):153–60.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(443):1053–1062.

Dehejia, R. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F., and Small, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514):447–458.

Fong, C., Hazlett, C., and Imai, K. (2017). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements.

Fong, C., Ratkovic, M., Hazlett, C., Yang, X., and Imai, K. (2016). *CBPS: Covariate Balancing Propensity Score*. R package version 0.13.

Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B., and Schneeweiss, S. (2014). Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, 33(10):1685–1699.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

Hansen, B. B. (2007). Flexible, optimal matching for observational studies. *R News*, 7:18–24.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.

Haviland, A., Nagin, D., and Rosenbaum, P. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12(3):247.

Heller, R., Rosenbaum, P. R., and Small, D. S. (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101.

Heller, R., Rosenbaum, P. R., and Small, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, 64(4):299–309.

Hill, J. L. (2008). Discussion of research using propensity-score matching: Comments on Ôa critical appraisal of propensity-score matching in the medical literature between 1996 and 2003Õ by peter austin, statistics in medicine. *Statistics in Medicine*, 27(12):2055–2061.

Hirano, K. and Imbens, G. W. (2005). *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hsieh, C.-T. and Urquiola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from chile's voucher program. *Journal of Public Economics*, 90(8):1477–1503.

Hsu, J. Y., Zubizarreta, J. R., Small, D. S., and Rosenbaum, P. R. (2015). Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, 102(4):767–782.

Iacus, S. M., King, G. K., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171(2):1–22.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76(1):243–263.

Imai, K. and Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, 110(511):1013–1023.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jackson, J. W. (2016). Diagnostics for confounding of time-varying and other joint exposures. *Epidemiology*, 27(6):859–869.

Keele, L., Titiunik, R., and Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A*, 178:223–239.

King, G., Lucas, C., and Nielsen, R. A. (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2):473–489.

Kuhn, H. (1955). The Hungarian Method for the Assignment Problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Lara, B., Mizala, A., and Repetto, A. (2011). The effectiveness of private voucher education: Evidence from structural school switches. *Educational Evaluation and Policy Analysis*, 33(2):119–137.

Lechner, M. (2001). *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*, pages 43–58. Physica-Verlag HD, Heidelberg.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *The Review of Economics and Statistics*, 84(2):205–220.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *ArXiv e-prints*.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.

McEwan, P. J. (2001). The effectiveness of public, catholic, and non-religious private schools in chile's voucher system. *Education Economics*, 9:183–219.

MINEDUC (2016). Estadísticas de la educación 2015. `http://centroestudios.mineduc.cl/tp_enlaces/portales/tp5996f8b7cm96/uploadImg/File/Estadisticas/Anuario_2015.pdf`.

Mizala, A. and Romaguera, P. (2001). Factors explaining secondary education outcomes in chile. *El Trimestre Economico*, 272:515–549.

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statistical Science*, 5(5):463–480.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54(4):387–398.

Norris, F. H., Friedman, M. J., Watson, P. J., Byrne, C. M., and al, e. (2002). 60,000 disaster victims speak: Part i. an empirical review of the empirical literature, 1981-2001. *Psychiatry*, 65(3):207–39. Copyright - Copyright Guilford Publications, Inc. Fall 2002; Last updated - 2016-03-19; CODEN - PSYCAB.

Papadimitriou, C. (1994). *Computational Complexity*. Addison-Wesley, Reading (Mass.).

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., and Burgette, L. (2016). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*. R package version 1.4-9.5.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. and Hernan, M. A. (2008). *Estimation of the causal effects of time-varying exposures*, chapter 23, pages 553–599. Chapman and Hall/CRC.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A*, pages 656–666.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84:1024–1032.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. (2005a). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B*, 67(4):515–530.

Rosenbaum, P. R. (2005b). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21(1):57–71.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83.

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

Rosenbaum, P. R. and Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, 41(1):103–116.

Rosenbaum, P. R. and Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29:159–183.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Sapelli, C. and Vial, B. (2002). The performance of private and public schools in the chilean voucher system. *Cuadernos De Economia*, 39(423-454).

Sapelli, C. and Vial, B. (2005). Private vs public voucher schools in chile: New evidence on efficiency and peer effects. Working Paper 289, Catholic University of Chile, Instituto de Economía.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*. Iowa State University Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21.

Stuart, E. A. and Rubin, D. B. (2007). *Matching Methods for Causal Inference*, chapter 11, pages 155–176. Thousand Oaks, CA: Sage Publications.

Traskin, M. and Small, D. (2011). Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences*, 3:94–118.

USDVA (2016). Traumatic effects of specific types of disasters. `https://www.ptsd.va.gov/professional/trauma/disaster-terrorism/traumatic-effects-disasters.asp`.

USGS (2011a). Magnitude 8.8 - offshore bio-bio, chile. `http://earthquake.usgs.gov/earthquakes/recenteqsww/Quakes/us2010tfan.php`.

USGS (2011b). Report on the 2010 chilean earthquake and tsunami response. `https://pubs.usgs.gov/of/2011/1053/`.

USGS (2014). Largest earthquakes in the world since 1900. `http://earthquake.usgs.gov/earthquakes/world/10_largest_world.php`.

Yang, F., Zubizarreta, J. R., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2014). Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician*, 68(4):253–263.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065.

Zhang, K., Small, D. S., Lorch, S., Srinivas, S., and Rosenbaum, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *Journal of the American Statistical Association*, 106(494):511–524.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

Zubizarreta, J. R., Cerdá, M., and Rosenbaum, P. R. (2013). 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. and Keele, L. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *Journal of the American Statistical Association*, 112(518):547–560.

Zubizarreta, J. R. and Kilcioglu, C. (2016). `designmatch`*: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design.* `R` package version 0.2.0.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *Annals of Applied Statistics*, 8(1):204–231.