Statistical Methods for Causal Inference in Observational Studies

Lateef Babatunde Amusa

February, 2020

Statistical Methods for Causal Inference in Observational Studies

by

Lateef Babatunde Amusa

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
DOCTOR OF PHILOSOPHY
in
APPLIED STATISTICS

Thesis Supervisors: Prof. Temesgen Zewotir
Prof. Delia North

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Lateef Babatunde Amusa, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then

   (a) their words have been re-written but the general information attributed to them has been referenced, or

   (b) where their exact words have been used, then their writing has been placed in italics and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

_____ _____

Lateef Babatunde Amusa (Student) Date

Disclaimer

This document describes work undertaken as a PhD programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Declaration

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal Westville, under the supervision of Professor Temesgen Zewotir and co-supervised by Professor Delia North.

I, Lateef Babatunde Amusa, declare that this thesis is my own and unaided work. It has not been submitted for any academic or examination purpose at any other university.

[redacted] 17/02/2020

Lateef Babatunde Amusa Date

[redacted] February 17, 2020

Professor Temesgen Zewotir Date

[redacted] 17/02/2020

Professor Delia North Date

Abstract

Estimating causal effects is essential in the evaluation of a treatment or intervention. It is particularly straightforward for well-designed experiments. However, when the treatment assignment is complicated by confounders, as in the case of observational studies, such inferences regarding the treatment effects, require more sophisticated adjustments. In this thesis, we investigated different matching techniques in terms of how well they balance the treatment groups on the covariates, as well as their efficiency in estimating treatment effects. We considered the various algorithm variants of these matching techniques, which include propensity score matching, Mahalanobis distance matching, and coarsened exact matching. Secondly, we proposed two new strategies for estimating treatment effects, namely, covariate-balancing rank-based Mahalanobis distance (CBRMD) and an improved version of CBRMD (iCBRMD). We evaluated their performance via simulations and some real-life datasets. Thirdly, we investigated a relatively new optimization-based alternative, known as entropy balancing, which has been used rarely in the applied biomedical literature. We shared our experiences learned from using entropy balancing in non-experimental studies, via Monte Carlo simulations and an empirical application. We further extended the evaluation of entropy balancing to some standard measures of causal treatment effects, namely; difference in means, odds ratios, rate ratios and hazard ratios. We pulled together our evaluations by conducting Monte Carlo simulations, evaluating both well-established methods and the more recently proposed methods. These adjustment techniques were evaluated under different scenarios that align with the practical reality. Finally, we utilized a dataset from a

recently conducted HIV Incidence Provincial Surveillance System (HIPSS) study, to apply the considered techniques to a public health issue in South Africa.

Acknowledgements

First and foremost, many thanks to my supervisors, Professor Delia North and Professor Temesgen Zewotir, for their continued interest and guidance throughout this PhD journey. I am indeed fortunate to have two supervisors, who are not only distinguished scholars, but also parental figures. Amongst other things, they guided me through the transition of being an excellent student to an independent researcher. I cannot thank them enough. Only God can repay them for their good deeds.

The academic staff members of the Department of Statistics (Westville campus) have been awesome. I acknowledge Mrs Alershneey, the extremely hardworking departmental Secretary, who has been supportive and diligent in helping the students, including me. Gratitude further goes to the School of Maths, Statistics, and Computer science, for providing space, time and resources needed in undertaking this research. The high-performance computers provided to the postgraduate students, were extremely helpful.

Finally, I would like to appreciate my parents, for their prayers, encouragement, patience and understanding throughout the period of my academic pursuit. To my wife, Ramat, and my daughter, Faridat, I say thank you for allowing me to take away those many hours that could have been spent with you, while doing my research work.

Publications

The following eight (8) journal articles have been produced from this thesis:

1. Amusa, L., Zewotir, T., and North, D. (2019c). Examination of Entropy balancing technique for estimating some standard measures of treatment effects: A simulation study. *Electronic Journal of Applied Statistical Analysis*, 12(2):491-526 **(Published)**

2. Amusa, L., Zewotir, T., and North, D. (2019b). A weighted covariate balancing method for estimating causal effects in case-control studies. *Modern applied science*, 13(4):40-50. **(Published)**

3. Amusa, L., Zewotir, T., and North, D. (2019a). Evaluation of subset matching methods: Evidence from a Monte Carlo simulation study. *American Journal of Applied Sciences*, 16(3):92-100. **(Published)**

4. Amusa, L., Zewotir, T., and North, D. (2019d). A Simulation Study of some Modern Weighting Methods for Estimating Treatment Effects in Observational Studies. *Journal of Modern Applied Statistical Methods*. **(Accepted)**

5. Amusa, L., Zewotir, T., and North, D. Examination of Entropy balancing technique for estimating causal treatment effects in observational studies. *Songklanakarin Journal of Science and Technology*. **(Accepted)**

6. Amusa, L., Zewotir, T., and North, D. On the estimation of causal effects in observational studies: Can Mahalanobis distance matching be redefined? *Electronic Journal of Applied Statistical Analysis*. **(Under review)**

7. Amusa, L., Zewotir, T., and North, D. Evaluating different techniques for estimating causal treatment effects in observational studies: A simulation study. *Communications in Statistics - Simulation and Computation* **(Under review)**

8. Amusa, L., Zewotir, T., North, D., Ayesha Bm Kharsany, and Lara Lewis. Evaluating the causal treatment effect of medical male circumcision on HIV acquisition: Application of the state-of-the-art matching methods. **(In preparation)**

Contents

List of Figures

List of Tables

Abbreviations and Notations

ASMD	absolute standardized mean difference
ATE	average treatment effect
ATT	average treatment effect on the treated
CBRMD	covariate-balancing rank-based Mahalanobis distance
CEM	coarsened exact matching
CI	confidence interval
CV	coefficient of variation
EPBR	equal percent bias reduction
GLM	generalized linear model
HR	hazard ratio
iCBRMD	improved covariate-balancing rank-based Mahalanobis distance
IPW	inverse probability weighting
MDM	Mahalanobis distance matching
MIB	monotonic imbalance bounding
MSE	mean square error
OR	odds ratio
PS	propensity score
PSM	propensity score matching
RR	rate ratio
RMSE	root mean square error
SE	standard error
TMLE	targeted maximum likelihood estimation
VR	variance ratios

Notations

n	number of treated observations or units
n_t	number of treated units
n_c	number of control units
K	number of covariates
$\mathbf{\Sigma}$	Variance-Covariance Matrix
β_{trt}	conditional treatment effect
δ	marginal treatment effect
Y^1	Potential outcome for the treated group
Y^0	Potential outcome for the control group
$P_{i,trt}$	Probability of treatment selection
ω	Vector of weights
Bolded lower-case latin letters	Vector
Bolded upper-case latin letters	Matrix

Chapter 1

Introduction

Estimation of causal effects has been the motivation for much research in the social, demographic and health sciences (He et al., 2016; Pearl et al., 2009; Imbens & Rubin, 2015). In an observational study, a random assignment of units to treatment groups is not feasible in the investigation of treatment effects. Because randomized experiments are the gold standard for evaluating treatment effects, efforts are being made to make an observational study structured in order to a resemble simple randomized experiment.

Though causality has been initially studied from experiments, Rubin and his colleagues' remarkable work in this area built a framework that allows causality to be studied from observational studies (see Rubin, 2006). The framework, which is referred to as the counterfactual or potential outcomes model of causality, has been refined and as a result, a unified framework for the prosecution of causal questions is now available.

To further understand causality, as suggested by Holland (1986); Dawid (2000); Pearl et al. (2009), we should distinguish between the following questions:
Associational: "Do people take aspirin when they have a headache?"
Interventional (effects of causes): "I have a headache. Will it help if I take aspirin?"

Counterfactual (causes of effects): "My headache has gone. Is it because I took aspirin?"

Note that the first question is non-causal and it was added to distinguish associational inference from causal inference.

Classical statistics, like regression, champion solving the first question. For example, if we observe predictors $\mathbf{X}$ and response $\mathbf{y}$ and wish to estimate $\mathbf{y}$ given $\mathbf{X}$. Commonly, we can use the likelihood principle to infer the parameter by embedding the conditional distribution in a parametric family $P(\mathbf{y}|\mathbf{X}) = P_\theta(\mathbf{y}|\mathbf{X})$. The (Gaussian) linear regression is a typical example:

$$P_{\boldsymbol{\beta},\sigma}(\mathbf{y}|\mathbf{X}) \propto exp\frac{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2}{2\sigma^2} \ . \tag{1.1}$$

This type of inference is most useful in making predictions, where the probability distribution is assumed constant.

As association does not imply causation, the real question is whether we can use statistics techniques to answer causal questions (the interventional and counterfactual questions)? Except for older statistics literature (Neyman, 1923)which answers queries of causality in settings of randomized experiments, the statistical theory has been relatively silent on causality questions.

1.1 The Potential Outcomes Model

The causal language used in the third question asked earlier belongs to the counterfactual or potential outcomes setting. This causal model, which is perhaps the most widely adopted approach by applied researchers, was utilized in this thesis. This model is generally attributed to Rubin's works in the 1970s and 1980s; though Splawa-Neyman et al. (1990) thinks Neyman first used this language in random-

ized experiments. Accordingly, this approach is commonly called the Rubin causal model, or the Neyman-Rubin causal model.

Rubin (1974) defined the causal effect of a treatment using the potential outcomes. The potential outcomes language links observational studies to the more general missing data problem (Rubin, 1976, 1977). The potential outcome model assumes that a population of interest can be exposed to two alternative groups of a cause. A distinct set of conditions, which potentially affects an outcome of interest, characterizes each treatment group, even though each condition can be observed in only one treatment group, at any point in time. When only two treatment groups are considered, they are referred to as 'treated' and 'control' groups. Throughout this thesis, we will conform to this convention.

For example, for the causal effect of participation in the school feeding program on performance (mid-term scores), children who participated in the program have a theoretical what-if mid-term scores under the "not participating" state, while children who did not participate in the feeding program, have a hypothetical what-if mid-term scores under the "participating" state. These "what-if" potential outcomes are counterfactual in the sense that they exist hypothetically, but are not observed.

Formalizing this concept, we assume that there is a binary treatment variable T_i, and $\mathbf{X}_i = (X_{i1}, ..., X_{iK})$ denote a $K-$ dimensional vector of observed pre-treatment covariates associated with unit i. Let $Y_i^t : t \in \{0, 1\}$ be the potential outcome variable value, that is, the value of the outcome variable if $T_i = t$, also known as a counterfactual outcome. Rubin (2005) stated that the assignment mechanism is the most important quantity in observational studies, which can be written as $P(T|\mathbf{X}, Y^0, Y^1)$. He also stated that randomized experiments share a critical property called ignorable,

$$P(T|\mathbf{X}, Y^0, Y^1) = P(T|\mathbf{X}, Y) \, . \tag{1.2}$$

This assumption implies that it is justifiable to ignore the missing values (unobserved potential outcomes). The observed outcome Y is then defined as

$$Y = \begin{cases} Y_i^1, & T_i = 1 \\ Y_i^0, & T_i = 0 \, . \end{cases} \tag{1.3}$$

In other words, $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ is observed.

Though a stronger assumption is the strong ignorability of treatment assignment (Rosenbaum, 1983), also known as strongly ignorable or unconfounded:

$$P(T|\mathbf{X}, Y^0, Y^1) = P(T|\mathbf{X}) \, . \tag{1.4}$$

Equation (1.4) means that the potential outcomes $Y_i(1)$ and $Y_i(0)$ are independent of $T_i|\mathbf{X}_i$. In other words, there are no unmeasured confounders - the treatment assignment model has included all relevant covariates. This assumption enables the treatment assignment to be sufficient for controlling pre-treatment differences between the treated and control groups.

Another important property is that

$$0 < P(T|\mathbf{X}, Y) < 1 \text{ , and } 0 < P(T|\mathbf{X}) < 1 \, . \tag{1.5}$$

Intuitively, Equation (1.5) means that a subject or unit has a non-zero probability of being assigned to the treated group. This is called the overlap assumption in the context of observational studies.

1.2 Aggregate Causal Effects

Equation (1.3) implies that one can never observe the potential outcome under the treated group for those observed in the control group, and the potential outcome

under the control group can never be observed for those in the treated group. Therefore, the calculation of individual-level causal effects is impossible. Attention is usually focused on the estimation of defined aggregate causal effects. Aggregate causal effects are usually defined as averages of individual-level effects.

These aggregate effects can be defined for any subset of the population. In many cases, the average treatment effect (ATE) for the entire population, which is the broadest possible average effect, is the quantity of interest. This effect can be expressed mathematically as

$$\hat{\delta}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^1 - Y_i^0)\,, \tag{1.6}$$

where n is the total number of subjects or observations. Equation (1.6) is defined with reference to a well-defined target population. For example, in the school feeding program scenario described in Section 1.2, the population would be all the primary school children in that community or county under investigation. In this thesis, depending on the nature of outcome variables, we also considered other group-level effects like risk ratios, odds ratios, hazard ratios, i.e. the interest is not just risk or mean differences, as in Equation (1.6).

One of the variants of the ATE is the average treatment effect among the treated (ATT). ATT is defined as

$$\hat{\delta}_{ATT} = \frac{1}{n_t}\sum_{i|T_i=1}^{n_t}(Y_i^1 - Y_i^0), \tag{1.7}$$

where n_t is the number of treated units. It is imperative to note that equations (1.6) and (1.7) both specify the treatment effect after a specific point in time i.e; time is constant and there is no reference of a time effect. ATT is a popular estimand in the medical and health sciences. For example, in determining the effect of intraoperative blood transfusion in cardiac surgical patients. The ATT of interest would be to with-

hold blood transfusion for all the patients who currently receive blood. It would not make sense to estimate the ATE in this case, as it would require contrasting either withholding blood transfusion for all patients or giving blood to all patients.

In the definitions above, units selection is made before the treatment assignment so that the estimands in Equations (1.6) and (1.7) are defined for the fixed n units. It is often more practical for the n units to be viewed as random samples from a large population. Theoretically, it is more convenient to assume an infinite population, so that the units are i.i.d. draws. Therefore, we can define the estimands as

$$\hat{\delta}_{ATE} = E(Y^1 - Y^0), \text{or } \hat{\delta}_{ATT} = E(Y^1 - Y^0 | T = 1) \, . \tag{1.8}$$

In Equation (1.8), the expectations are taken over the joint distribution of (T, Y^1, Y^0). In this spirit, we also implicitly assumed the stable unit treatment value assumption (SUTVA) of Rubin (1980), which loosely refers to the assumption of no interference between units. In other words, the potential outcomes of one unit, are not affected by the potential outcome of another unit. While we do not know if the ignorability assumption (1.2) or strong ignorability (1.3) is correct for an observational study, we are willing to assume them. The question is how we proceed after assuming these conditions to make an observational study which resembles a randomized experiment? This shall be discussed extensively in this thesis.

1.3 Some Methods for Adjustment in Observational Studies

Quite several techniques have been developed and utilized to improve our ability to draw reliable causal inferences from observational studies. Historically, multivariate regression is probably the most straight forward and commonly used approach, where the response variable is modelled as a function of the treatment status and relevant background characteristics. Other methods like propensity scoring (including matching, stratification and weighting), instrumental variables (Heckman, 1997),

machine-learning methods (McCaffrey et al., 2004, 2013; Friedman et al., 2010), and entropy balancing (Hainmueller, 2012) have also joined the repository of available methods (Zagar et al., 2017). Without loss of generality, we describe the propensity score (PS) methods for estimating the ATT in the next section.

The Propensity Score

In a seminal paper, Rosenbaum (1983) changed the approach in which observational researchers can control for observed confounding in estimating causal effects. He showed that if the strongly ignorable assumption is made, then the difference between the treated and control groups at each value of a balancing score is an unbiased estimate of the treatment effect at that value. Consequently, matching, subclassification and covariance adjustment on this balancing score (referred to as the propensity score) can produce unbiased estimates of the average treatment effect.

The propensity score (PS) was defined as the probability of treatment assignment, given the observed baseline covariates. Let T be treatment assignment indicator, $\mathbf{X}$ be the observed baseline covariates and π be the propensity scores, then it can be expressed as

$$\pi(\mathbf{X}) = P(T = 1|\mathbf{X}) \, . \tag{1.9}$$

To simplify the notation, independence is assumed, so that

$$P(T_1 = t_1, ..., T_n = t_n) = \prod_{i=1}^{n} \pi_i^{t_i} (1 - \pi_i)^{1-t_i} \, . \tag{1.10}$$

Although logistic regression has become the standard method for estimating the PS, Lee et al. (2010); Setoguchi et al. (2008) have also estimated the PS using probit models, discriminant analysis and more recently, bagging or boosting, recursive partitioning or tree-based methods, random forests and neural network models.

Propensity Score-adjusted Regression

The PS can be used as a covariate to adjust for the observed covariates in a regression setting (Rosenbaum, 1983), so that we can use a regression model of the form

$$E(Y|T, \mathbf{X}) = \beta_0 + \beta_1 \, \pi(\mathbf{X}) + \beta_2 \, T \, . \tag{1.11}$$

After fitting Equation (1.11), the ATT is estimated as

$$\hat{\delta}_{ATT,reg} = \hat{\beta}_2 + \hat{\beta}_1 \, \overline{\pi(\mathbf{X_t})} \, , \tag{1.12}$$

where $\overline{\pi(\mathbf{X_t})}$ is the mean of the propensity scores for the treated units.

Stratification

Stratification on the PS involves stratifying treated and control units into mutually exclusive subgroups or strata based on their estimated propensity scores. Their estimated propensity score ranks the units, where the units are then stratified into subclasses, based on previously defined cut-offs of the estimated propensity scores. Observations, or units with similar propensity scores, are placed into one stratum. A typical approach is to utilize the quintiles of the estimated propensity scores to separate the units into five equal-sized groups.

Stratification, with five strata based on the quintiles, has been shown to remove about 90% of the bias in estimating treatment effect (Austin, 2011; Rosenbaum & Rubin, 1984). In estimating the ATT, subclassification on the propensity scores, can be done by summing the weighted treatment effect in each stratum, where the weight of the stratum is the proportion of the treated subjects in the stratum over the treated subjects in the sample (Williamson et al., 2012). That is,

$$\hat{\delta}_{ATT,strat} = \sum_{h=1}^{H} \frac{N_{T_h}}{N_t} \hat{\delta_h} \, , \tag{1.13}$$

where H is the number of strata, N_t is the number of treated units, $\hat{\delta_h}$ is the estimated treatment effect for the hth stratum, and N_{T_h} is the number of observations in the $h-$th stratum.

Matching

Matching is a method that samples from a large reservoir of control units to produce a modest control group in which the distribution of covariates in the treated and control groups is similar. Propensity scores can be used for matching by sampling control group units as a match for the treated units, based on the propensity scores of some distance function between the treatment groups.

The following steps briefly explain propensity score matching: (i) For each treated unit, select a single control group unit having similar or comparable values of the estimated propensity scores, (ii) estimate the within-pair treatment effect by taking the difference in the two outcome groups and (iii) calculate the average within-pair treatment effect estimate. More details about matching on propensity scores, as well as the general concept of matching are discussed in Chapter 2.

Weighting

Weighting based on the propensity scores, otherwise called the inverse probability weighting (IPW) method, is to weight the treated and control observations to make them representative of the population of interest. More details about weighting based on propensity scores, as well as the general concept of weighting, are discussed in Chapter 2.

1.4 Balance Assessment

The aim of pre-processing or adjustment techniques is to adjust the treated and control groups in such a way that the empirical distribution of the covariates $\mathbf{X}$ are as similar as possible between the two treatment groups. Two groups are said to be

balanced with respect to $\mathbf{X}$, if they have identical distributions of $\mathbf{X}$.

Balance measures include visualisations like the q-q plots and histograms of continuous variables, as well as comparisons of high-dimensional distributions like the Kolmogorov-Smirnov (K-S) statistic. Low-dimensional summaries, such as variance ratios and differences in means, are presented here.

Due to the measurement of different covariates on very different scales, the differences in means are usually standardized by dividing them with some pre-adjustment standard deviation, for which the absolute standardized mean difference (ASMD) has been used in the literature. ASMD values above 0.1, may be indicative of covariate imbalance, as suggested by some authors (Mamdani et al., 2005; Normand et al., 2001). Some authors (Ho et al., 2011; Imai et al., 2008; McCaffrey et al., 2013) also suggested that ASMD values above 0.25 are large. The ASMD value for each covariate X_k $(k = 1, 2, ..., K)$ is

$$ASMD = \left| \frac{M_t(X_k) - M_c(X_k)}{\sqrt{L}} \right| \tag{1.14}$$

where $L = \frac{V_t(X_k)+V_c(X_k)}{2}$ or $L = V_t(X_k)$ when ATE or ATT is the estimand, respectively. $M_t(X_k)$ is the mean of the covariate for the treated group and $M_c(X_k)$ is the mean of the covariate for the control group; $V_t(X_k)$ and $V_c(X_k)$ denote the variance of the treated and control groups, respectively. Note that the variance formulation assumes equal variance for the two populations.

The variance ratio (VR), as defined by Stuart (2010), compares the variances between the treated and control groups, using the formula:

$$VR_k = \frac{V_t(X_k)}{V_c(X_k)}, \tag{1.15}$$

Covariates with variance ratios close to 1, are judged to be balanced.

P-values of significance tests that includes information on the sample size (e.g. t-test and Chi-square test for significant differences in the means and proportions, respectively for the two groups) could also be employed (Ali et al., 2015). Though previous studies (Austin, 2007; Imai et al., 2008) have argued that such tests do not contribute to the analysis since balance is inherently an in-sample property, without reference to any broader population, that should be improved as much as possible, regardless of the significance of some test statistic. Furthermore, hypothesis tests can be misleading as balance measures because they often confuse changes in balance with changes in statistical power. For matching, randomly discarding observations can lead to a significant test statistic becoming insignificant despite no systematic improvement of balance having occurred.

In this thesis (except in Chapter 6), we utilized only the absolute standardized mean difference (ASMD) as a measure of balance.

1.5 Research Motivation and Objectives

Estimating average treatment effects is essential in the evaluation of a treatment or intervention. It is particularly straightforward in experiments but very complicated in observational studies, since treatment assignment is not random. The complication comes from the fact that treatment exposure may be associated with background covariates that are further associated with the potential response, which may substantially introduce covariates imbalance in these treatment groups. There is thus a need for methods that adjust for such confounding of the background covariates for a reliable causal inference to be made from observational data.

While there have been several statistical techniques developed, ranging from simple to sophisticated methods, for controlling confounding in observational data analyses, there is a great need for guidance and recommendations regarding the optimal

strategy for making causal inference in specific scenarios.

This study centers around statistical methods for estimating causal treatment effects in observational studies, as well as their variants. In particular, connections with other studies are drawn. These include Monte Carlo simulations, extending existing methodologies and applications to real-world data. The objectives of this thesis, as outlined in the following points, are to

- review the existing methods for estimating causal treatment effects in observational studies;
- empirically evaluate some matching methods;
- propose new adjustment techniques and evaluate them, relative to the gold standard techniques, via Monte Carlo simulations and notable real-life datasets;
- explore some modern weighting techniques via some Monte Carlo simulations and establish some trade-offs;
- provide general recommendations for the optimal strategies or techniques based on findings from a series of simulations for different scenarios.

1.6 Main Research Contributions

The main contributions of this research work are that we developed a rank-based Mahalanobis distance technique, namely, the covariate balancing rank-based Mahalanobis distance (CBRMD) method, that can reduce the effect of confounding in estimating causal treatment effects.

We further improved on the CBRMD methods by introducing a generalized version of the CBRMD technique, which was built on changing how the rank-based Mahalanobis distances were utilized. Our new method has a parameter that is capable of optimizing the desired cost function (covariate balance and efficiency), for a given

dataset.

We provided insights into the use of a highly efficient technique known as entropy balancing. Not many applications of this highly effective method have been utilized in quasi-experimental research designs, particularly in the medical and health literature.

In addition to the mentioned contributions, this thesis avails us the opportunity to thoroughly review the fundamental basis of some of the existing confounder adjustment techniques and offers useful contributions, suggestions and recommendations, based on our experience in this research work.

1.7 Thesis Layout

The rest of the thesis is organized as follows:

Chapter 2 provides a review of matching and weighting methods that are used in the causal inference literature. Chapters 3 - 8 are products of journal articles that have either been published or under review. Chapter 3 is a simulation study evaluating the different matching methods and algorithms reviewed in Chapter 2. In Chapter 4, we proposed a new technique for adjusting confounders in the estimation of causal treatment effects. In Chapter 5, we improved on the proposed method in Chapter 4 and introduced a new version, while providing new insights and some numerical explorations. In Chapter 6, we studied the entropy balancing technique via Monte Carlo simulations and an empirical application. In Chapter 7, we extend the examination of entropy balancing to outcome of different types, estimating some standard treatment effects. Chapter 8 is an overall comparison of some selected methods that were discussed in this thesis. Several simulation studies were carried out and we provided comments on each technique in terms of their strengths and weaknesses.

Chapter 9 is a further application of the utilized adjustment techniques in Chapter 8, by evaluating treatment effects from an HIV study conducted in the Kwazulu-Natal province of South Africa. Chapter 10 gives an overall discussion and conclusion of the thesis. We provided suggestions for further studies, along with a summary of the results achieved in this study.

Chapter 2

Review of Matching and Weighting Methods

Though in Chapter 1, we discussed several approaches to estimating causal effects in observational studies, our contributions in this thesis have revolved around matching and weighting methods. Accordingly, in this chapter, a detailed description of matching and weighting methods, that have been proposed and utilized in the literature, is provided. We outlined a review of the basic variants of each of these methods, with a focus to estimate the average treatment effect among the treated (ATT).

2.1 Matching Methods

In observational studies, matching is a non-parametric method used to control the confounding influence of pretreatment covariates. Stuart (2010) broadly defined matching as any technique that aims to equalize (or "balance") covariates distribution in the treated and control groups. Matching methods find control group units with similar distributions of the observed covariates as that of the treated units. As a result, control group units that were not selected by the matching algorithm are discarded. A successfully matched dataset does not require any further controlling for covariates $\mathbf{X}$ and causal effects can be estimated with less model dependence

and reduced statistical bias.

Matching can be viewed methodologically as a strategic sub-sampling from among treated and control units. Following the lead of Smith & Todd (2005), all matching estimators of the ATT can be expressed as some variation of

$$\hat{\delta}_{ATT,match} = \frac{1}{n_t} \sum_i [(y_i|T_i = 1) - \sum_j w_{i,j}(y_j|T_j = 0)], \tag{2.1}$$

where n_t is the number of treated units, i is the index over the treated units, j is the index over the control units and $w_{i,j}$ represents a set of scaled weights measuring the distance between each control unit and the treated unit. In this chapter, we will broadly divide matching methods into raw matching, propensity score matching and coarsened exact matching methods.

2.1.1 Raw Matching

We refer to the term raw matching as methods that use only the raw covariates and do not attempt to model the treatment assignment mechanism. Once a distance metric d based on the raw covariates is selected, a matching algorithm can then be applied. The very first example is exact matching, which matches identical covariate values of the treated group with the control group. As required for any matching method, a distance metric d is specified on the covariates. For exact matching:

$$d(\mathbf{X}_t, \mathbf{X}_c) = \begin{cases} 0, & \mathbf{X}_t = \mathbf{X}_c \\ \infty, & \mathbf{X}_t \neq \mathbf{X}_c \, . \end{cases} \tag{2.2}$$

Equation (2.2) implies that two units are matched only if they have exactly the same pre-treatment covariates. While estimating the ATT, exact matching uses weights:

$$w_i = \begin{cases} \frac{1}{g_i}, & \text{matched control units} \\ 0, & \text{unmatched control units,} \end{cases} \tag{2.3}$$

where g_i is the number of exact matches identified for each treated unit. The inherent property of exact matching is often too stringent and may not be realistic in a high-dimensional covariate space. Thus, the Mahalanobis distance is an alternative measure that has found usage in matching, for this scenario. In calculating the Mahalanobis distance, categorical variables could be converted to a series of binary indicators. Gu & Rosenbaum (1993) showed that the performance of this measure is not optimal for matching when the covariates are not normally distributed, or in the presence of a bountiful number of covariates. The Mahalanobis distance between two units, $\mathbf{X}_t$ and $\mathbf{X}_c$ is defined as

$$d(\mathbf{X}_t, \mathbf{X}_c) = (\mathbf{X}_t - \mathbf{X}_c)^T \, \mathbf{\Sigma}^{-1} \, (\mathbf{X}_t - \mathbf{X}_c), \tag{2.4}$$

where $\mathbf{X}_t$, $\mathbf{X}_c$ denote the treated and control group covariates, respectively and $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$. Once the distance measure $d(.)$ is selected, a matching algorithm can then be applied. For Mahalanobis distance matching, the $c:1$ nearest neighbour algorithm estimates the ATT. This algorithm matches each treated unit to c $(c \geq 1)$ control units that are closest in terms of d, while the unselected control units are discarded. It is possible to utilize the nearest neighbour matching with replacement, where a control unit can be chosen multiple times as a match, or without replacement, where the algorithm proceeds greedily and there is an emphasis on which treated units are matched first.

A more sophisticated algorithm, optimal matching (Rosenbaum, 2012), is also an option. It involves selecting the overall best match of the data among the candidate set by minimizing a global distance measure. A good example is genetic matching, which has an implementation in the R package, *Matching* (Sekhon, 2011). Genetic matching involves the introduction of a generalized form of the Mahalanobis distance by including an additional weight parameter W (Diamond & Sekhon, 2013). A genetic search algorithm is then used to choose W in the generalized Mahalanobis distance, given a specified criterion of covariate imbalance to ensure that the matched

samples optimize the specified imbalance criterion. Formally,

$$d(\mathbf{X}_t, \mathbf{X}_c) = \sqrt{(\mathbf{X}_t - \mathbf{X}_c)^T (\mathbf{\Sigma}^{-\frac{1}{2}})^T \, W \, \mathbf{\Sigma}^{-\frac{1}{2}} \, (\mathbf{X}_t - \mathbf{X}_c)} \, . \tag{2.5}$$

2.1.2 Propensity Score Matching

As mentioned in Section 2.1.1, an alternative to matching based on raw covariates, is to model the treatment assignment mechanism. Rosenbaum (1983) introduced the propensity score - a balancing score, which summarizes all of the covariates into one scalar measure. An estimated propensity score gives the probability of treatment assignment, given the observed covariates: $\pi = P(T = 1|\mathbf{X})$. The value of π is commonly estimated from a logistic regression model.

In addition to logistic regression, there are several statistical techniques that can be used to estimate the PS, such as linear and quadratic discriminant analysis, probit models; although more recently, nonparametric methods such as boosted CART and generalized boosted models (GBM), recursive partitioning or tree-based methods, random forests and neural network models Lee et al. (2010); Setoguchi et al. (2008) have also been used and often showed good performance.

The introduction of propensity scores was motivated by the inherent difficulty of extending raw matching methods to a high-dimensional covariate space. In light of this, propensity scores are an essential tool to adjust for covariate imbalance. The PS is a scalar and can be used for matching using the two standard ways to define the propensity score distance:

$$d(\mathbf{X}_t, \mathbf{X}_c) = |\hat{\pi}(\mathbf{X}_t) - \hat{\pi}(\mathbf{X}_c)| \text{ or}$$
$$d(\mathbf{X}_t, \mathbf{X}_c) = |logit(\hat{\pi}(\mathbf{X}_t)) - logit(\hat{\pi}(\mathbf{X}_c))| \, .$$

After the propensity scores are estimated, one can apply either the nearest neighbour (with or without replacement) or optimal matching algorithms described in Section

2.1.1. Additionally, for Mahalanobis distance and PS matching, PS calipers can be used with the nearest neighbour algorithm. Calipers provide a restriction imposed on the distance between the covariates of the two treatment groups, such that a control group unit is matched if

$$d(\mathbf{X}_t, \mathbf{X}_c) < \xi,$$

where ξ is the user-specified caliper. Rosenbaum & Rubin (1985) proposed using a caliper of size a quarter of the propensity scores standard deviation.

Subclassification, which involves stratifying units into mutually exclusive strata, based on the estimated propensity scores, can also be used for matching (Rosenbaum & Rubin, 1984). Further, full matching (Hansen, 2004), is a more sophisticated form of subclassification, which uses all the sample units and creates a series of matched sets which contain at least one treated unit and at least one control unit.

2.1.3 Coarsened Exact Matching

The earlier mentioned matching techniques are known as "equal percent bias reducing" (EPBR) methods. For these methods, improvements in the bound of balance for one covariate will affect each of the other covariates. They also do not guarantee a certain level of imbalance reduction for any given dataset. Further, these methods are associated with a tedious process of continuous matching and re-matching, until a sufficient level of balance is achieved on all covariates.

Iacus et al. (2012) introduced coarsened exact matching (CEM), which avoid the shortcomings of EPBR methods. CEM is a special case of Monotonic Imbalance Bounding (MIB) methods. These methods improve the bounds on the balance of one covariate in isolation and do not affect the maximum imbalance of each of the other covariates. Unlike other matching methods, where balance is being checked continually, CEM inverts the process and thus guarantees that the covariate imbal-

ance between the matched groups will not exceed the user's pre-specified level.

CEM essentially coarsens each variable as reasonably as possible, by subclassifying them into distinct groups, either by using automated choices of coarsening, using the Sturges rule (Scott, 2009) or a user-defined coarsening. The "exact matching" algorithm is then applied to the coarsened data to select matches and prune unmatched units. In other words, after coarsening, the algorithm creates a set of strata, $s \in S$, each with the same coarsened **X** values. Units in the strata, containing at least one treated and at least one control unit, are retained, while units in the remaining strata are removed from the sample.

We use T^s and C^s to denote the set of treated and control units, respectively in stratum s; $m_T^s = \#T^s$ as the number of treated units in T^s, similarly $m_C^s = \#C^s$ is the number of control units in C^s. The number of matched units are $m_T = \bigcup_{s \in S} m_T^s$ and $m_C = \bigcup_{s \in S} m_C^s$, for the treated and control groups, respectively. The unmatched units receive zero weight, while to each matched unit i in stratum s, CEM assigns the following weights:

$$w_i = \begin{cases} 1, & i \in T^s \\ \frac{m_C m_T^s}{m_T m_C^s}, & i \in C^s \end{cases}. \tag{2.6}$$

2.2 Weighting Methods

We will discuss weighting methods for estimating causal effects in observational studies in this section. In the sense of non-technicality, the matching methods discussed in Section 2.1.1 are also weighting methods with discrete weights. The weighting methods discussed in this section produce continuous weights and are inherently different from matching. There are two general weighting approaches in causal inference. One does not directly make covariate balance its primary objective, i.e. it focuses on modelling the data to get probabilities, from which weights that reduce

the imbalance to some considerable extent, can be obtained. The other approach, also known as automated covariate balancing methods, directly use some minimization algorithm to choose weights that perfectly balance the covariates, subject to some specified constraints.

2.2.1 Propensity Score Weighting

In this thesis, we refer to propensity score weighting, commonly called inverse probability weighting (IPW) in observational studies literature, as weighting methods which agree with the first general weighting approach described above. Propensity score weighting (Crump et al., 2009; Hirano & Imbens, 2001; Hirano et al., 2003; Imbens, 2004) is the continuous counterpart of propensity score matching, which originates from a survey sampling problem. The idea was formed from the Horvitz-Thompson weight (Horvitz & Thompson, 1952), which for each sample unit, is the inverse of the probability of such unit being assigned to the observed group.

In estimating the ATT, the IPW technique weights by the odds of the unit being assigned to the treatment group of interest. (Imbens, 2004). Formally,

$$w_i = \begin{cases} 1, & T_i = 1 \\ \frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}, & T_i = 0 \ , \end{cases} \tag{2.7}$$

where $\pi(\mathbf{X})$ is the estimated propensity score. Equation (2.7) implies that each unit is assigned a weight that equals the reciprocal of the probability of receiving the treatment that the unit received. It is unlikely for the propensity score to be known in practice, so it is routinely estimated using a parametric model like the logistic model. Therefore, the success of the IPW largely rests on the correct specification of the propensity score model. Slight misspecification of the PS model will result in substantial bias of the estimated treatment effects (Kang & Schafer, 2007). If π_i is the

inclusion probability of the sample Y_i, the Horvitz-Thompson estimator is given by

$$\hat{\mu_i} = \frac{1}{n}\sum_{i=1}^{n} {\pi_i}^{-1} Y_i \, .$$

In observational studies, the ATT, $E(Y_i^1 - Y_i^0 | T = 1)$ can be viewed as estimating two population means over the treated units. Therefore, the IPW estimator for the ATT, while extending the Horvitz-Thompson estimator, is given as

$$\hat{\mu}_{ATT,IPW} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1 - T_i) Y_i \, \hat{\pi}(\mathbf{X}_i)/1 - \hat{\pi}(\mathbf{X}_i)}{\sum_{i=1}^{n} (1 - T_i) \, \hat{\pi}(\mathbf{X}_i)/1 - \hat{\pi}(\mathbf{X}_i)} \tag{2.8}$$

2.2.2 Empirical Calibration Weighting

Weighting methods generally seek non-negative weights w, which ensures that the weighted empirical distributions of the covariates for the treated and control groups are as close as possible. The standardized difference (Rosenbaum & Rubin, 1985; Austin & Stuart, 2015) typically measures the distance between these weighted distributions concerning some covariate function ϕ:

$$d_{sd,\phi(.)}(F_w(0), F_w(1)) = \frac{E_{F_w(1)}[\phi(\mathbf{X})] - E_{F_w(0)}[\phi(\mathbf{X})]}{Var_{F_w(0)+F_w(1)}\phi(\mathbf{X})}, \tag{2.9}$$

where $F_w(t, x)$ (t= 0 or 1) is the empirical distribution of the covariates for the treated and control groups.

In the earlier section, the approach was to assert a parametric model and then search for a balance-maximizing specification, by trying one after the other, which is stressful and error-prone. In other words, specifying a propensity score model which has a high level of covariate balance, is a cautionary approach. We now turn to empirical calibration weighting, which we have used as a general term for weighting methods that seek optimal balance. These methods are built on the notion of achieving efficiency by solely balancing the covariate distributions, without a direct estimation of the propensity score or outcome regression function. A similar example of this idea

applied to matching is the genetic matching of Diamond & Sekhon (2013), which searches for the matches that achieve the best possible balance.

Formally, empirical calibration weighting (EBCW) seeks weights w, such that the standardized difference (2.9) is zero or small for some pre-specified functions ϕ_k, $k = 1, 2, ..., m$. The weights w attempt to solve the equation

$$\sum_{i|T_i=1} w_i \phi_k(\mathbf{X}_i) = \sum_{i|T_i=0} w_i \phi_k(\mathbf{X}_i), k = 1, ..., m \; . \tag{2.10}$$

Finally, the resulting weights are normalized by requiring

$$\sum_{i|T_i=0} w_i = 1 \; . \tag{2.11}$$

In general, the class of calibration estimators minimizes the overall distance $\sum_i D(w_i, v_i)$ between the final weights to a given vector of design weights, subject to exact balance (2.10) and the normalization (2.11), where v_i is a set of uniform weights.

Though calibration estimators were initially used in survey sampling (Deville & Särndal, 1992; Tan, 2010), researchers have found the usage thereof in observational studies (Hainmueller, 2012; Chan et al., 2016; Zubizarreta, 2015). Specific examples include entropy balancing (Hainmueller, 2012), which considered estimating ATT with D being the negative Shannon entropy; stable balancing weights (Zubizarreta, 2015), which uses the squared norm of w as the objective function, whilst allowing for inexact balance and Chan et al. (2016) proved that the empirical calibration estimators for ATT and ATE could achieve the semiparametric efficiency bound. Imai & Ratkovic (2014) offered a related (but not precisely the same) approach, but it is one that integrates covariate balancing into the estimation of the propensity score.

Chapter 3

Evaluation of Subset Matching Methods: A Simulation Study

This chapter is an extension of Amusa et al. (2019a), which was attached in the Appendix (10.1). So far, we have introduced and reviewed several techniques to estimate causal effects in observational studies. We focus on matching methods in this chapter. While many simulation studies have compared the performance of different matching methods, it cannot be taken for granted that their findings are transferable to other data scenarios (Franklin et al., 2014). Even though there have been a few notable studies that have examined the performance of matching techniques in terms of how well they balance the groups on the covariates, only a few of them have extended the evaluation to outcome analyses (Austin, 2014; Jacovidis, 2017; Stone & Tang, 2013).

In Amusa et al. (2019a), we empirically compared the performance of matching methods, namely; Mahalanobis distance matching, Propensity score matching and coarsened exact matching. In this chapter, we build upon the limitation of the previous study by extending the simulation scenarios, as well as introducing variations of matching algorithms on the Mahalanobis distance and the propensity score. The theory of these methods and algorithms have been presented in Chapter 2.

3.1 Background

There are several matching methods existing in the literature, each employing different distance measures, algorithms and rules for selecting control group units. Each technique could potentially choose various control group units from the overall control pool to create the matched group. The matched control group composition could, therefore, vary considerably depending on the particular matching algorithm used (Jacovidis, 2017).

Matching techniques have been applied either using covariate (Miksch et al., 2010) or propensity score matching (Stock et al., 2010; Windt & Glaeske, 2010; Drabik et al., 2012), with some authors providing evidence for the superiority of propensity score (PS) matching (Drabik et al., 2012). The causal inference literature has shown that PS matching is not necessarily the gold standard (Fullerton et al., 2016). Depending on the scenario considered, other matching techniques can induce a better balance on the covariates, as well as producing more efficient treatment effect estimates. Furthermore, the performance of PS matching depends on the correct specification of the propensity score model, choice of covariates and the matching algorithm used. (Dehejia & Wahba, 2002; King et al., 2011; Rosenbaum & Rubin, 1984).

Accordingly, we evaluate the performance of different matching methods and algorithms via a series of Monte Carlo simulations. We examine covariate balance, average matched sample size and efficiency of treatment effect estimates. Without loss of generality, we compared coarsened exact matching, propensity score matching (with and without caliper), Mahalanobis distance matching (with and without caliper) and full matching (on Mahalanobis distance and propensity score). For the nearest neighbour matching methods, we assumed a 1:1 pair matching and without replacement. Pair matching selects for each treated unit the control unit with the smallest distance from that treated unit. By matching without replacement, we mean that controls can be used as matches for only one treated unit.

3.2 Simulation Study

We used a data generation scheme derived from previous studies (Lee et al., 2010; Setoguchi et al., 2008), where ten randomly generated ten baseline covariates, $X_1 - X_{10}$ had standard normal distributions. Some pair of covariates were induced with specified levels of dependence. $X_1, X_3, X_5, X_6, X_8, X_9$ were dichotomized. In the course of this study, we tried many different simulation conditions, but for simplicity sake, we decided to present three of them, which mostly align with practice reality. The following scenarios were considered:

Scenario 1 (S1): Not all covariates were confounders i.e., related to both treatment and outcome. X_1, X_2, X_3, X_4 are associated with both treatment and outcome, X_5, X_6, X_7 are predictors of the treatment variable only, while X_8, X_9, X_{10} are predictors of the outcome variable only. The treatment status was generated from a Bernoulli distribution: $T_i \sim Ber(P_{i,trt})$, where the probability of treatment selection $P_{i,trt}$ was determined from

$$log\left(\frac{P_{i,trt}}{1-P_{i,trt}}\right) = \alpha_{0,trt}+\alpha_1 X_1+\alpha_2 X_2+\alpha_3 X_3+\alpha_4 X_4+\alpha_5 X_5+\alpha_6 X_6+\alpha_7 X_7\,. \quad (3.1)$$

The outcome model was simulated linearly as
$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$.

Scenario 2 (S2): Having common covariates affecting both treatment selection and outcome models, including extraneous variables. The treatment selection model in Equation (3.1) still holds, while we generated the outcome variable as
$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{trt} T_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$.

Scenario 3 (S3): Following the recommendation that model complexity is likely to

improve estimators that rely on the propensity score (d'Agostino, 1998), we considered interactions among covariates in both treatment choice and outcome models as

$$\begin{aligned} log\left(\frac{P_{i,trt}}{1-P_{i,trt}}\right) &= \alpha_{0,trt} + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 \\ &+ \alpha_6 X_6 + \alpha_7 X_7 + \alpha_2 X_2^2 + \alpha_4 X_4^2 + \alpha_7 X_7^2 + 0.7\ \alpha_2 X_2 X_4 + \\ &0.5\ \alpha_5 X_5 X_7 + 0.7\ \alpha_2 X_2 X_3 + 0.5\ \alpha_3 X_3 X_4 + 0.5\ \alpha_4 X_4 X_5\ . \end{aligned} \tag{3.2}$$

The outcome model was simulated non-linearly as
$Y_i = e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10}} + \beta_{trt} T_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$.
For the three scenarios, the coefficients, $\alpha_1 - \alpha_7$ were based on real-life data utilized in a previous study (Setoguchi et al., 2008), while
$\beta_1 = \beta_2 = \beta_3 = log(2), \beta_4 = \beta_5 = \beta_6 = log(1.75)$ and $\beta_7 = log(1.5)$ to reflect very high, high, and moderate effect sizes (Austin, 2007, 2014).

By varying the values of $\alpha_{0,trt}$, we fixed two values of the percentage of subjects who received the treatment (subsequently referred to as prevalence of treatment) at 25% and 33%. The chosen treatment prevalence rates 25%, 33% correspond to $r = 3:1, 2:1$, respectively, where r is the ratio of controls to one treated unit. We also varied sample sizes between $n = 500, 800$, and 1000. Overall, we generated a total of 1000 replications of each dataset and matched them with each method.

3.2.1 Performance Measures

We evaluated the performance of the methods in terms of covariate balance, average size of the matched sample and treatment effect estimates.

For covariate balance, the methods were compared in terms of their ability to induce balance in the covariates, between treated and control groups. This was achieved using the absolute standardized mean difference formula of Equation (1.14) for each covariate. It has been suggested that a standardized mean difference of at most 0.1 is

quite sufficient at balancing a given covariate between the treatment groups (Austin, 2007; Normand et al., 2001).

In terms of the size of matched samples, the number of successfully matched treated and control group units was examined for each matching method, relative to the raw data.

Finally, the ATT estimates were obtained and evaluated according to the absolute bias and root mean square error (RMSE) of the estimated treatment effects.

Table 3.1: Description of matching strategies and software implementation

Method	Label	R package	R function and parameters
Unmatched data	RAW	-	-
Coarsened exact matching	CEM	cem	cem (with default parameters)
Propensity score matching	PS1	MatchIt	matchit (NN matching, WOR,) distance=logit)
Propensity score matching (with PS caliper)	PS2	MatchIt	matchit (NN matching, WOR, distance=logit,caliper=0.25 SD)
Mahalanobis distance matching	MD1	Matching	Match (NN matching, WOR)
Mahalanobis distance matching (with PS caliper)	MD2	Matching	Match (NN matching, WOR, caliper=0.25 SD)
Full matching on the PS	FUL1	optmatch	matchit(optimal full matching, distance=logit)
Full matching on the Mahalanobis distance	FUL2	optmatch	matchit(optimal full matching, distance=Mahalanobis)

Note: NN denotes Nearest neighbour; WOR: without replacement; SD is the standard deviation of the logit(PS).

3.2.2 Methods

The aim of this chapter is to investigate the comparative performance of several matching methods used for estimating the ATT. In estimating the ATT, the data were partitioned into a collection of subclasses or matched sets according to the estimated propensity score or/and Mahalanobis distance of each observation. Discrete weights for each observation in the simulated dataset were then derived based on the matched set membership, and the ATT estimates are obtained via weighted lin-

ear regression of the outcome on treatment assignment. A list of the considered methods, with references to specific software implementation and the relevant literature, is given in Table 3.1.

3.3 Results

In this section, we provide results obtained from analyzing the simulated datasets in each of the considered scenarios. We present the results for each performance measure under separate subsections. Results for the two considered prevalence rates, $r = 3:1, 2:1$, had similar comparisons among the methods; hence, we report our findings based on any one of them. As a form of sensitivity analysis, we ran simulations for other treatment prevalence rates, but we do not present the results, as no qualitative differences were observed in the relative performance of the method.

Covariate Balance

In this section, we evaluate the matching methods in terms of covariate balance by presenting the results of S1, for sample size n=500 and the two prevalence rates, 25% and 33%. In Figure 3.1, we showed boxplots of the absolute standardized mean differences (ASMD) of the covariates, averaged across the simulated datasets. The other simulation conditions produced a similar pattern (results not shown). It is observable from Figure 3.1 that the original simulated data is substantially imbalanced with the ASMD value as high as 0.6. The two prevalence rates produced comparatively similar results. Full matching on the Mahalanobis distance (FUL2) performed poorly: not only did it produce ASMD values all above 0.1, but it also appeared to have similar results with the unmatched data. Though Mahalanobis distance matching (MD1) performed quite satisfactorily in balancing the covariates, it did not induce balance on all the covariates, and its performance was sub-optimal as compared to other methods. PS matching with caliper (PS2) and full matching on the propensity score (FUL1) had the best performance, with both producing ASMD

values close to zero on all the covariates. Coarsened exact matching (CEM), Mahalanobis distance with PS caliper (MD2) and propensity score matching (PS1) performed excellently.

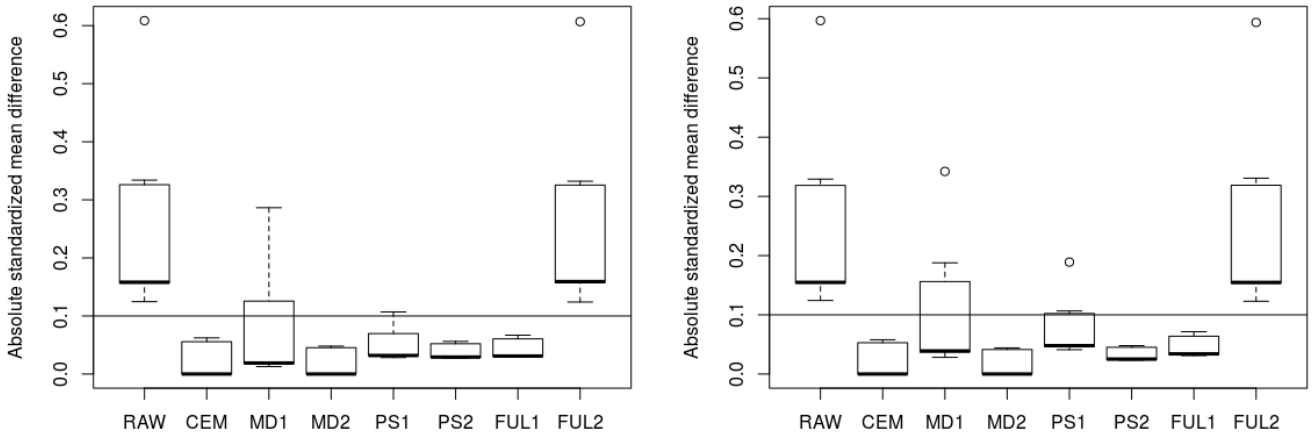

Figure 3.1: Boxplots of absolute standardized differences in means in Scenario 1 when r = 3:1 (leftpanel) and r=2:1 (right panel)

Sample Size

Figure 3.2 presents the average matched sample size from each of the matching methods. We presented results for the two considered prevalence rates, under S1 and S3. Findings from both scenarios are similar. As expected, the average number of matched pairs when the treatment prevalence rate was 33%, is higher than the average number of matched pairs when the treatment prevalence rate was 25%. As expected, CEM produced a minimal matched sample size, since the CEM algorithm leads to extreme coarsening of variables and discards any subjects without exact matches on those coarsened variables. Mahalanobis distance matching with PS caliper (MD2) resulted in the smallest matched sample size. PS matching with caliper (PS2), resulted in a smaller matched sample size compared to PS matching without caliper. The number of discarded units depend on the width of the PS

caliper. Full matching (both on PS and Mahalanobis distance) retained the original sample size, as expected.

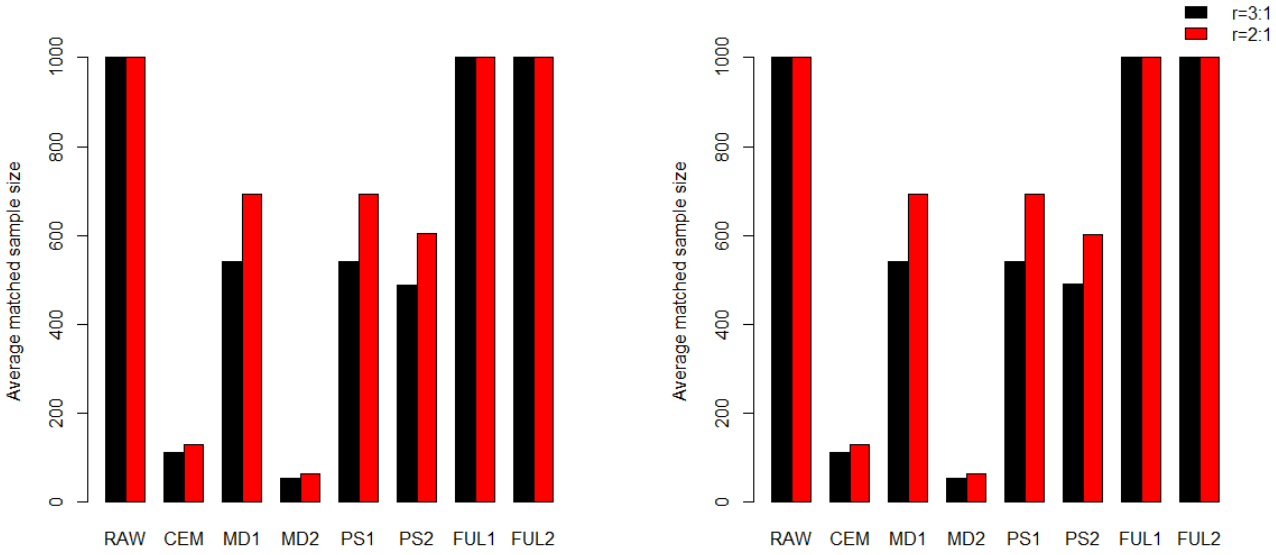

Figure 3.2: Sample size after applying the matching methods for Scenario 1 (left panel) and Scenario 2 (right panel)

Treatment Effect Estimates

Overall, the increase in sample size did not affect the bias of treatment effect estimates. Full matching on the Mahalanobis distance (FUL2), in most cases, either did not reduce the bias or slightly increased it. Except for full matching (on the propensity score and Mahalanobis distance), an increase in sample size correspondingly decreased the RMSE of the different methods, across all the considered scenarios. For full matching, it is not surprising that increasing sample size did not affect the RMSE, since theory, as well as results from Figure 3.2, showed that full matching does not discard sample units and the initial sample size is retained. As expected, the sample size effect is more substantial for coarsened exact matching (CEM), since there is the need to have an abundant sample size to make up for the many units that were discarded from the CEM algorithm. Consequently, the efficiency of the

outcome estimation for CEM might depend heavily on the sample size.

Results of S1, where a covariate is either related to the treatment or the outcome or both, are summarized in Figure 3.3. In terms of bias, PS matching with caliper (PS2), CEM, Mahalanobis distance matching with PS caliper (MD2), full matching on the PS (FUL1), reduced the absolute bias considerably well, with most of them producing near-zero bias values. Mahalanobis distance matching (MD1) and FUL2 produced considerable bias. In terms of RMSE, MD2 produced the worst results, followed by CEM and FUL2. They increased the RMSE relative to the original data. However, for the sample size (n=1000), CEM outperformed MD1 and FUL2. The other four methods performed considerably well and produced similar results in most cases.

Results of S2, where all covariates are included in the treatment assignment model and the outcome regression model in a linear fashion, are summarized in Figure 3.4. As compared to S1, results from this scenario produced higher bias and RMSE values. In terms of bias, PS2, CEM, MD2, FUL1 reduced the absolute bias considerably well, with most of them producing near-zero bias values. Though MD1 and PS1 also reduced the absolute bias, PS1 produced slightly smaller values. For RMSE, MD2 produced the worst results, followed by CEM and FUL2. They increased the RMSE relative to the original data. However, for the sample size (n=1000), CEM outperformed MD1 and competed favourably with FUL2. The other four methods performed considerably well and produced similar results.

Results of S3 are summarized in Figure 3.5. It shows the performance of the matching methods when interactions of some covariates, as well as non-linearity in the outcome model. As compared to S1 and S2, results from this scenario produced higher bias and RMSE values. This suggests that results introducing model complexities could make or mar the efficiency of estimates if the right model terms are

not specified. The two methods that depend on the propensity score (PS1 and PS2) produced similar bias and RMSE values in most cases.

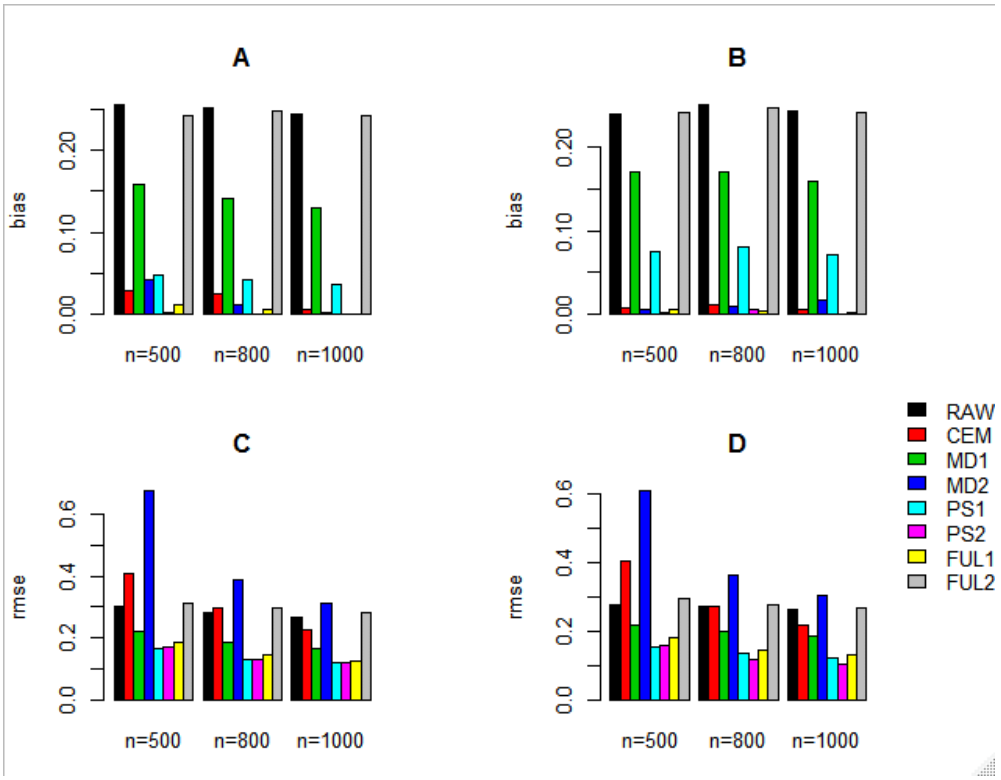

Figure 3.3: Absolute Bias (Panel A: r=3:1, Panel B: r=2:1) and RMSE (Panel C: r=3:1, Panel D: r=2:1) of the estimated treatment effects, for Scenario 1

3.4 Discussion and Conclusion

This simulation study aimed to evaluate seven matching methods, each selecting a subset of treated and control units, namely; coarsened exact matching, propensity score matching (with and without PS caliper), Mahalanobis distance matching (with and without PS caliper) and full matching (on Mahalanobis distance and propensity score). The performance assessment of these methods was based on (i) the ability to induce balance on measured background covariates; (ii) the sample size retained after matching and (iii) the performance of treatment effect estimates.

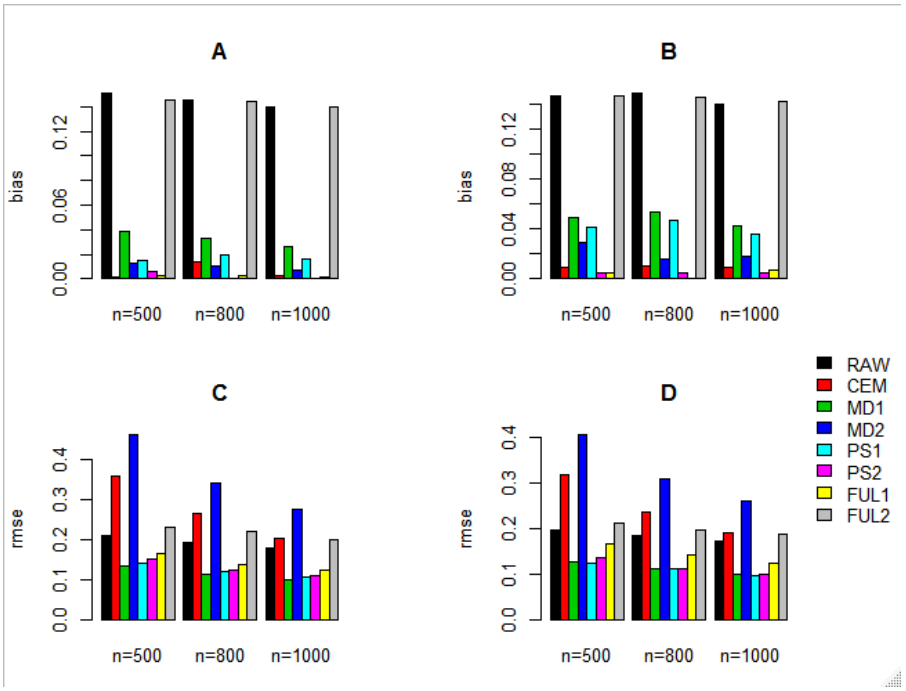

Figure 3.4: Absolute Bias (Panel A: r=3:1, Panel B: r=2:1) and RMSE (Panel C: r=3:1, Panel D: r=2:1) of the estimated treatment effects, for Scenario 2

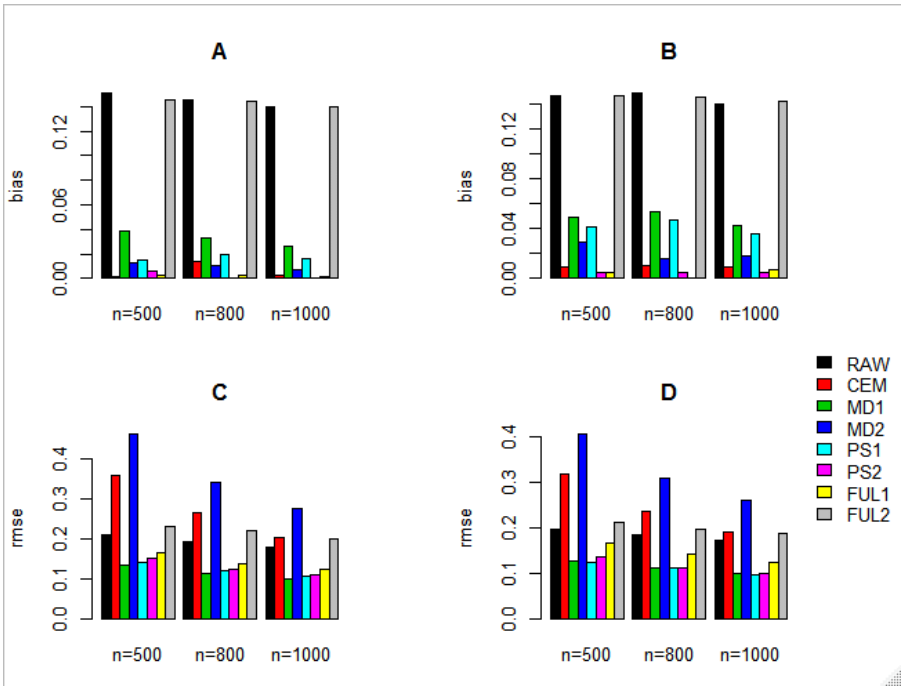

Figure 3.5: Absolute Bias (Panel A: r=3:1, Panel B: r=2:1) and RMSE (Panel C: r=3:1, Panel D: r=2:1) of the estimated treatment effects, for Scenario 3

Though our findings suggest that no particular technique was overall superior to others, full matching on the Mahalanobis distance, consistently produced a sub-optimal performance in covariate balance and treatment effects estimation. In terms of the size of the matched samples, the average number of matched units was minimal for CEM, given the chosen automated coarsening. In terms of its corresponding effect on outcome estimation, results from the simulation study revealed that the choice of degree of coarsening, could make or mar the performance of CEM. If the elements of the coarsening values are too small, then too many observations may be discarded, which may then lead to not finding a solution or result in inefficient outcome estimations. In contrast, if the elements of the coarsening values are set too high, though more units will be retained, useful information that might produce better matches may be missed. Furthermore, more covariate imbalances, model dependence and statistical bias, will be introduced (Iacus et al., 2012). However, in practice, variable coarsening are selected based on the substance of the variable.

When matching by covariates, Austin (2008) recommended using PS calipers to improve the matching. Our simulations provided results supporting this claim for PS matching; however, using PS calipers with Mahalanobis distance matching, considerably reduced the efficiency of treatment effects. Full matching on the propensity score performs excellently well; however, full matching on the Mahalanobis distance is not recommended, as evident from the results of this simulation study. Worthy of note concerning PS matching, is the correct specification of the PS model. This was evident in the optimal performance of PS methods in S1, where the PS model is correctly specified, as compared to S2, where the PS model is most likely incorrect. In practice, an excellent alternative to 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 (Lee et al., 2010).

The inclusion of interaction and squared terms in both the treatment and outcome model, has been advised in previous studies (de los Angeles Resa & Zubizarreta, 2016; Zagar et al., 2017). However, results from our simulation study showed that this does not necessarily improve the performance of PS methods, unless the correct model complexity terms are chosen.

While several simulation studies have compared different matching methods, most of them were centred on methods that rely on the propensity score. To our knowledge, this simulation study is one of the very few that accommodates different variations of matching, in terms of the nearest neighbour and optimal algorithm, as well as the propensity score and Mahalanobis distances.

Efficient estimation of treatment effects is the ultimate goal of causal inference in an observational study. Therefore, it is essential to note that covariate balance is only a means to an end - not an end in itself. Accordingly, we align with the thought of a previous study (de los Angeles Resa & Zubizarreta, 2016), which recommended that the widely used threshold ASMD value of 0.1 for balancing covariates, should not be used rigidly. Instead, it should be as flexible as possible, depending on the available data and sample size after matching, such that a sufficient amount of data which can ensure accuracy of treatment effect estimates within a reasonable level of covariate balance, is retained from the matching process. Optimal methods like full matching, which do not discard observations, do not share this problem.

Our simulation findings are reliable and generalizable because the simulations were based on traditional study designs that mimic practice reality. However, our simulation results might be limited to the scenarios considered by our simulation data. Therefore, the results cannot be generalized to settings that have not been evaluated.

Chapter 4

Tailoring the Mahalanobis Distance Matching

This chapter is a slight modification of (Amusa et al., 2019b), which was attached in the Appendix (10.1). In Amusa et al. (2019a); Zagar et al. (2017), as well as Chapter 3, while empirically evaluating the performance of some matching methods, we observed the sub-optimal performance of the Mahalanobis distance matching (MDM), even when it was used with propensity score calipers. Accordingly, we take an interest in this particular matching technique and redefine how the Mahalanobis distance can be used for matching, to reduce covariate imbalance and improve the efficiency of treatment effect estimates.

We provide a basic description of the proposed method in Section 4.2. The performance of the proposed method is evaluated through a series of Monte Carlo simulations in Section 4.3, as well as an empirical application in Section 4.5.

4.1 Background

The Mahalanobis metric was the first choice of a distance measure that was utilized for matching (Rubin, 1980). While taking account of the correlations among variables $\mathbf{X}$, the Mahalanobis distance ensures that a difference of one standard de-

viation counts the same for each covariate in $\mathbf{X}$. Like the propensity score, it is also easy to implement. However, the Mahalanobis distance was initially developed for use with multivariate normal data, and for that data type, it works fine. With non-normal data, the Mahalanobis distance can exhibit some rather odd behaviour. If a covariate contains extreme outliers, its variance will be exaggerated, and the Mahalanobis distance will tend to ignore that covariate in matching. For binary indicators, the variance is most massive for events that occur about half the time, and it is smallest for events with probabilities of occurrence near zero and one. Consequently, the Mahalanobis distance gives higher weight to binary variables, with probabilities of occurrence near zero or one, as compared to those with probabilities of occurrence closer to 0.5.

4.2 Proposed Methodology

In this chapter, we introduce a rank-based Mahalanobis distance matching approach, namely, the covariate balancing rank-based Mahalanobis distance (CBRMD) method, to estimate treatment effects in the presence of confounding factors. We show how to use a modified form of the Mahalanobis distance - the rank-based Mahalanobis distance, proposed by Rosenbaum (2002), as matching weights that can reduce covariate imbalance between treated and control groups, and can efficiently estimate treatment effects.

Consider a random sample of $n = n_t + n_c$, with each i $(i = 1, 2, ..., n)$ belonging to only one of a binary group T_i, for which the estimation of causal effects is of interest. The ith unit received the treatment of interest (treated) if $T_i = 1$, and $T_i = 0$ if the treatment was not received (control). Let $\mathbf{X}_i = (X_{i1}, ..., X_{iK})$ denote a $K-$ dimensional vector of observed pre-treatment covariates associated with unit i.

According to Stuart (2010), matching methods are characterized by four steps:
(i) Utilize a distance measure that can measure closeness in determining whether a

treated unit is a good match for a control unit.
(ii) Implement a matching method based on the defined distance measure in (i).
(iii) Evaluate the matching quality, and perhaps iterating with Steps (i) and (ii) until well-matched samples are produced.
(iv) Given the matched samples, treatment effects are estimated and outcome analysis is carried out.

Recall that in conducting Mahalanobis distance matching, the Mahalanobis distance between covariates of the treated unit $\mathbf{X}_t$ and the control unit covariates $\mathbf{X}_c$, can be defined as

$$d(\mathbf{X}_t, \mathbf{X}_c) = (\mathbf{X}_t - \mathbf{X}_c)^T \, \mathbf{\Sigma}^{-1} \, (\mathbf{X}_t - \mathbf{X}_c), \tag{4.1}$$

where $\mathbf{X}_t$, $\mathbf{X}_c$ denote the treated and control group covariates, respectively, and $\mathbf{\Sigma}$ is estimated as the sample covariance matrix of $\mathbf{X}$ in the treated group, since ATT is our estimand of interest. The treated unit $\mathbf{X}_t$ is matched with control unit $\mathbf{X}_c$ with the closest $d()$. In other words, the algorithm matches each treated unit to $c(c \geq 1)$ control units that are closest in terms of $d()$, while the unselected control units are discarded.

To avoid the rather odd behaviour of the Mahalanobis distance for non-normal and outliers-present data, we replaced it with a rank-based Mahalanobis distance, defined as follows:

$$d_{\text{rank}}(\mathbf{X}_t, \mathbf{X}_c) = (r(\mathbf{X}_t) - r(\mathbf{X}_c))^T \, adj\mathbf{\Sigma}^{-1} \, (r(\mathbf{X}_t) - r(\mathbf{X}_c)), \tag{4.2}$$

where $r(\mathbf{X}_t)$, $r(\mathbf{X}_c)$ are the ranks of each of the covariates belonging to the treated and control groups, respectively. Average ranks are used for ties.
Further, note that $adj\mathbf{\Sigma}$ denotes the adjusted covariance matrix, which adjusts the $\mathbf{\Sigma}$ (variance-covariance matrix of the ranked covariates) by pre-multiplying and post-multiplying the covariance matrix of the ranks by a diagonal matrix whose diagonal

values are the ratios of untied ranks standard deviation to the tied ranks standard deviations of the covariates. In other words, $adj\mathbf{\Sigma} = \mathbf{D}\ \mathbf{\Sigma}\ \mathbf{D}$,

$$\mathbf{D} = \begin{bmatrix} \frac{S_u}{S_{t1}} & & \\ & \ddots & \\ & & \frac{S_u}{S_{tK}} \end{bmatrix},$$

where S_u is the standard deviation of untied ranks, and S_{tK} is the standard deviation of tied ranks for the kth covariate. Adjusting the covariance matrix was to prevent heavily tied covariates, such as rare binary variables, from having increased influence due to reduced variance.

From the matrix d_{rank} with dimension $t \times c$, the proposed algorithm extracts the control units and its corresponding rank-based Mahalanobis distance on each row of the matrix.

Finally, sample weights for the treated units are fixed at unity. For each treated unit, the control group units that have the smallest rank-based Mahalanobis distance from the individual treated units, are assigned the maximum weight value of one, while the remaining control units are down-weighted equally by a constant factor.

If any control unit does not have the smallest rank-based Mahalanobis distance from any treated unit, the CBRMD procedure does not give it a weight of zero. Instead, it only down-weights them. When there are ties in the control units that have the least rank-based Mahalanobis distance from any treated unit, the weight is approximately equally distributed among them so that every sample unit contributes to the estimation.

The weights allocation of the ith control unit can be expressed mathematically as

follows:

$$w_i = \begin{cases} 1, & i \in \min\limits_{\forall c} d_{\text{rank}} \\ \frac{1}{n_c}, & i \notin \min\limits_{\forall c} d_{\text{rank}}, \end{cases} \tag{4.3}$$

where n_c is the number of control units. The proposed algorithm is described by the following steps:

Algorithm 1 iCBRMD Algorithm

Step 1: Sort the data in order of the treatment indicator, with corresponding unit identification number. Rank the data.
Step 2: Compute the rank-based Mahalanobis distances of each of the treated units with the control group units, using Equation (4.2), and store the distances in a matrix with t rows and c columns.
Step 3: Create a vector which stores the column number of the control unit that has the smallest distance with the treated units in each row.
Step 4: Extract a frequency distribution based on Step 3 to identify the number of times each control unit had the smallest distance. Control units with zero frequencies are down-weighted approximately equally.
Step 5: Treated units have weights that are fixed at 1, while control units have weights based on Step 4.

Adopting the Neyman-Rubin causal model, let $Y_i(t)$, t = 0, 1 be the potential outcomes and the observed outcome $Y_i = T_i Y_i(1) + (1 - T_i)\, Y_i(0)$. The average treatment effect on the Treated (ATT) can be estimated as ATT = $\hat{\delta}_{ATT} = E(Y^1 - Y^0 | T = 1)$. The expectations can then be obtained from marginal computations:

$$E[Y^1 = 1] = \frac{1}{n_t} \sum_{i=1}^{n_t} w_i \, y_i,$$
$$E[Y^0 = 1] = \frac{1}{n_c} \sum_{i=1}^{n_c} w_i \, y_i,$$

where w_i denotes the ATT weights induced by CBRMD. For non-linear outcomes, a model-based approach, which involves incorporating the CBRMD weights by a weighted regression of the outcome on the treatment status indicator, can be used.

4.3 Simulation Study

In this section, we study the numerical performance of the CBRMD methodology via a series of Monte Carlo simulations in two phases: (i) we study the large-sample properties of the CBRMD technique; (ii) evaluate its effectiveness in balancing covariates and efficient estimation of treatment effects. The CBRMD technique is compared to Mahalanobis distance matching, as well as the inverse probability weighting (IPW). The choice of IPW as one of the methods for benchmarking the performance of the CBRMD method is due to its familiarity, and its representativeness of methods that produce continuous weights.

Phase 1: The first phase of simulation follows the Kang & Schafer (2007) design, which has been used in the causal inference literature to evaluate the performance of pre-processing techniques. We replicate the Kang and Schafer simulation to study the large-sample performance of the CBRMD method, relative to the raw data. We conducted 1000 Monte Carlo simulation runs for sample sizes, 200, 500, 1000, and 2000. In brief, the data generation of the design is as follows:
$Y_i = 210 + 27.4X_{i1} + 13.7(X_{i2} + X_{i3} + X_{i4}) + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$. Note that the $X_i's$ are independent, standard normally distributed and the treatment assignment model was generated with probability $\pi_i = \frac{1}{1+e^{-(X_{i1}+0.5X_{i2}-0.25X_{i3}-0.1X_{i4})}}$.

Phase 2: In the second phase, the simulations were made to be as realistic as possible, by simulating from real-life data. We generated treatment and outcome variables from the covariates of the Lalonde-PSID data. The data is a hybrid of program participants (treated units) from the experimental data of LaLonde (1986), while the control group was drawn from the Panel Study of Income Dynamics (PSID) data. The dataset comprises ten covariates, including age (age), indicator variables for unemployment in 1974 (u74) and 1975 (u75), marital status (married), lack of a high school diploma (nodegree), number of years of education (education), hispanic race (hispanic), black race (black), and real earnings in 1974 (re74) and 1975 (re75). The

outcome was the actual earnings in 1978. This data has a reputation of being used as a benchmark in the causal inference literature. The presence of stark imbalance motivated our choice of this data. Furthermore, the data will enable us to evaluate how well the CBRMD method can recover the treatment effect estimates from the experimental version of the data.

We considered outcomes that have linear, as well as non-linear functions. In specific terms, we discussed four types of outcomes: continuous (normal distribution), binary (binomial distribution), count (Poisson distribution), and time-to-event (survival distribution). The idea is to mirror some outcome variables that are typically encountered in the applied sciences. For example, household income, presence or absence of diseases, number of antenatal care visits by pregnant women and time to remission from cancer, are usually described by the normal, binomial, Poisson, and survival distributions, respectively.

4.3.1 Data Generation

Like (Iacus et al., 2012), we retain the covariates of the Lalonde-PSID data; we generate the treatment variable $T_i \sim Bernoulli(\pi_i)$, where π_i is obtained from a logistic regression model, with coefficients being the ones from fitting such a model to the real data, as shown in Equation (4.4).

$$\begin{aligned} logit(\pi_i) = \alpha_0 + \alpha_1\, age + \alpha_2\, edu + \alpha_3\, re74 + \alpha_4\, re75 + \alpha_5\, married + \\ \alpha_6\, black + \alpha_7\, hispanic + \alpha_8\, nodegree + \alpha_9\, u74 + \alpha_{10}\, u75\,. \end{aligned} \tag{4.4}$$

The outcome variables were generated for each of the following:

- Continuous outcomes: $Y_i = \sum_{k=1}^{10} \beta_k X_k + \beta_{trt} T_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$.
- Binary outcomes: $Y_i \sim Bernoulli(P_i)$, where $log\left(\frac{P_i}{1-P_i}\right) = \sum_{k=1}^{10} \beta_k X_k + \beta_{trt} T_i$.
- Count outcomes: $Y_i \sim Poisson(\eta_i)$, where $log(\eta_i) = \sum_{k=1}^{10} \beta_k X_k + \beta_{trt} T_i$.

- Time-to-event outcomes: For time-to-event or duration outcomes, we used a data-generating process, described by a previous study (Bender et al., 2005). Survival times t_i were generated as $t_i = \left(\frac{-log(U_i)}{\lambda e^{LP}}\right)^{\frac{1}{v}}$, where $U_i \sim Uniform(0,1)$, and the linear predictor, LP = $\sum_{k=1}^{10} \beta_k X_k + \beta_{trt} T_i$. We set $v = 2$ and $\lambda = 0.000001$.

This process generates survival times from a Cox-Weibull distribution. We assumed that all event times are observed for this chapter.

The model coefficients for the outcome models are set as $\beta_0 = 0, \beta_1 = \beta_2 = \beta_3 = \beta_4 = log(1), \beta_5 = \beta_6 = log(2), \beta_7 = \beta_8 = log(1.75)$, and $\beta_9 = \beta_{10} = log(1.5)$.

4.3.2 Assessing the Performance of Treatment Effect Estimates

While incorporating the weights generated from the CBRMD method, treatment effect estimates were determined by regressions of the particular outcome types on the treatment variable. For phase 2 simulations, the performance of treatment effect estimates was compared to the Mahalanobis distance matching and inverse probability weighting.

Except for the estimation of the difference in means for continuous outcomes, the other three treatment effects, namely; odds ratio (binary outcomes), rate ratios (count outcomes) and hazard ratios (duration outcomes), are not collapsible. Collapsibility refers to the coincidence of conditional and marginal treatment effects. Thus, for each of the conditional treatment effects (odds ratios, hazard ratios and rate ratios), we determined their corresponding marginal treatment effects δ.

For each phase, 1000 datasets were simulated. The performance of estimated treatment effects was assessed by calculating the absolute bias as $\left|\bar{\hat{\delta}} - \delta\right|$ and RMSE as $\sqrt{(\bar{\hat{\delta}} - \delta)^2 + var(\hat{\delta})}$, where $\bar{\hat{\delta}}$ is the mean of the 1000 regression coefficients.

4.4 Simulation Results

In this section, we present and explain the results obtained from analysing the simulated datasets. In terms of the large-sample performance of the CBRMD method, relative to the raw data, Figure 4.1 shows that increasing the sample size decreased the bias and RMSE of the estimated treatment effects for the CBRMD technique.

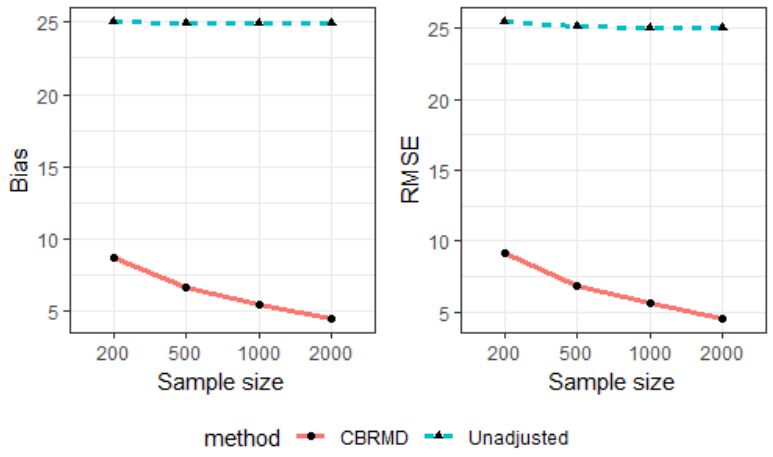

Figure 4.1: Assessment of large-sample properties of the CBRMD method, based the Kang and Schafer design

Table 4.1 shows the absolute bias and RMSE of the estimated treatment effects under the different outcome types considered. Figure 4.2 also supports the findings of Table 4.1, as it visualizes the deviation of the estimated treatment effects from the true effect, across the simulated datasets. Across the different types of estimated treatment effects, in terms of absolute bias and RMSE, there is evidence of a better performance of the CBRMD method, relative to the raw data.

Across the board, the CBRMD method outperformed MDM alone. For the difference in means, it resulted in about 97% reduction in both absolute bias and RMSE. For OR, it further reduced by 52% and 40% in absolute bias and RMSE, respectively. For RR, it resulted in about 59% reduction in both absolute bias and RMSE. For HR, 89% reduction in both absolute bias and RMSE was observed. When estimating hazard

ratios, it was observed that the CBRMD technique outperformed both MDM and IPW.

Table 4.1: Comparison of the ATT estimates for the different outcome types.

Method	Difference in means		Odds ratios		Rate ratios		Hazard ratios	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unadjusted	21813.29	21814.84	9.26	11.04	7.61	7.61	5.19	5.23
CBRMD	630.12	694.99	4.42	6.61	3.15	3.19	0.56	0.69
IPW	51.15	490.49	4.12	6.96	3.05	3.11	1.97	2.18
MDM	3329.73	3355.56	4.95	6.4	3.57	3.58	2.34	2.39

Note: Bias is measured in the absolute form.

4.5 Case Study: The Lalonde Data

In this section, we illustrate the CBRMD technique in real data setting. We reanalyzed the Lalonde-PSID data, whose covariates the simulation datasets were generated from. This data has a reputation of being used as a benchmark in the causal inference literature. The unique availability of experimental version of the data helps us to find out how well the application of the iCBRMD technique on the observational dataset recovers the experimental target.

As a form of validation exercise, we applied CBRMD to the Lalonde dataset that has been described in the simulation study. Figure 4.3 visualises balance of each of the ten covariates from the simulation data after applying the adjustment techniques. We superimposed a horizontal line on the panel to denote ASMD of 0.25, as some authors have suggested that ASMD values that exceed this threshold may indicate significant imbalance (Ho et al., 2011; Imai et al., 2008; McCaffrey et al., 2013). The Lalonde-PSID data had ASMD values ranging from 0.114 to 5.446. The CBRMD method substantially improved the balance on the ten covariates, with average ASMD values ranging from 0.002 to 0.215. Except for one covariate, IPW considerably reduced the covariate imbalance. MDM did not perform well, as it only induced sufficient amount of balance on five covariates.

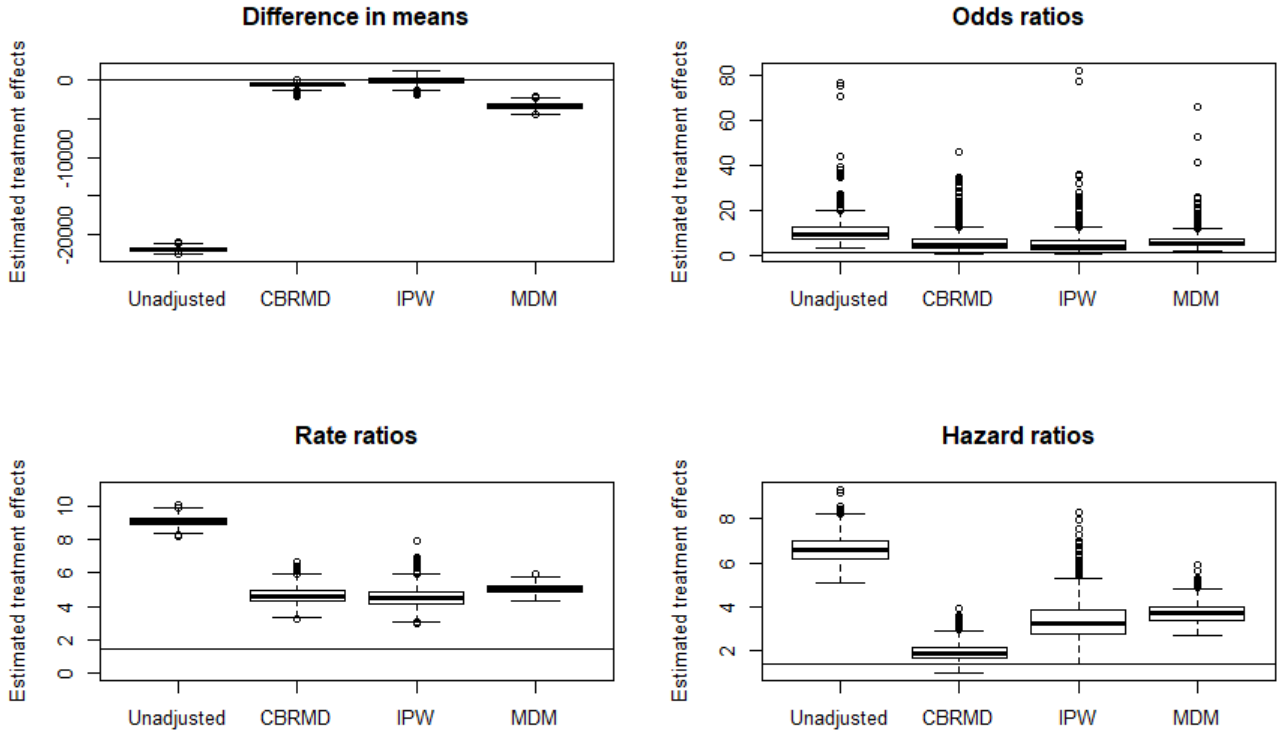

Figure 4.2: Boxplot of estimated treatment effects for the different outcome types

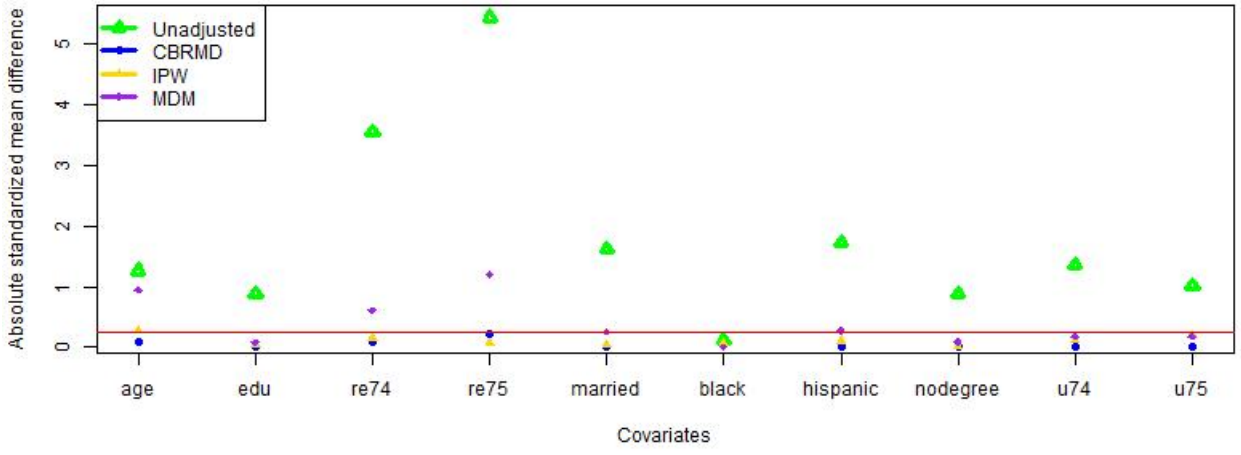

Figure 4.3: Covariate balance in the Lalonde data

A simple difference in means of the experimental version of the data yielded an average of $1794 with a 95% confidence interval of [551, 3038]. Table 4.2 shows the difference in means estimates and their associated 95% confidence interval from the adjustment methods. CBRMD and IPW produced statistically significant estimates, Mahalanobis distance matching did not. The difference in means between the treated group and the reweighted control group from the CBRMD method, yields an ATT estimate of $2064.5, with a 95% confidence interval of [150, 4803] - an estimate that is close to the experimental target.

The CBRMD and IPW adjusted estimates suggest that the job training programs significantly increased post intervention earnings. This is in agreement with findings from previous studies (Hainmueller, 2012; Diamond & Sekhon, 2013).

Table 4.2: Causal effect estimation in the Lalonde data

Estimator	Difference in means	95% Confidence Interval
Unadjusted	-15204.8	(-17468.80, -12940.75)
CBRMD	2064.5	(150, 4803)
IPW	2796.2	(1304, 4964)
MDM	-1482.1	(-3266.2, 301.9)

Note: Standard errors of the weighted estimators were bootstrapped with 2000 replicates

4.6 Discussion and Conclusion

In this chapter, we proposed a new pre-processing technique, which is based on computations from a rank-based Mahalanobis distance. Through simulations and an empirical application, we showed the effectiveness of the proposed method in terms of improvement in covariate balance, and efficient estimation of treatment effects.

It is noteworthy that the CBRMD method performs better with increasing sample

size. This was evident from the results obtained from the simulations and the case study that was carried out in this chapter. Large sample sizes are typical of epidemiological studies and national surveys.

The fact that we considered the estimation of different outcome-specific treatment effects is also a strength of the study in this chapter. We acknowledge that the Mahalanobis distance matching and IPW methods were only used as a benchmark for evaluating the performance of the CBRMD technique. The combination of CBRMD, with other pre-processing methods, may be considered in future studies.

When causal effects are of interest in the presence of confounding variables, as in the case of observational studies, the proposed covariate balancing rank-based Mahalanobis Distance (CBRMD) method, is a viable alternative method that can improve covariate balance, bias reduction and efficient estimation of treatment effects. However, we identified a gap in the utilized approach of obtaining the weights in the CBRMD technique. Therefore, in the next chapter, we will discuss this gap and improve on it by modifying the CBRMD technique.

Chapter 5

Improving the CBRMD Technique

In Chapter 4, we provided a new approach to estimating causal treatment effects, termed the covariate balancing rank-based Mahalanobis distance (CBRMD) matching technique, which is based on redefining how the Mahalanobis distance can be used for matching. By allocating equal weights of a constant factor to control units that could not attain the smallest rank-based Mahalanobis distances from any treated group, we believe that the CBRMD method does not explicitly create a fair play scenario. Accordingly, this chapter considers improving on the CBRMD method by providing new insights and numerical explorations. In addition, a web-based Shiny application written in R statistical language, was developed and deployed online to demonstrate the implementation of the proposed method.

5.1 Background

The CBRMD methodology is primarily based on weights computed from a rank-based Mahalanobis distance. In this chapter, we build on the previous proposal of the CBRMD methodology and introduce a slight modification, by considering a proper and more explicit definition of down-weighting control units, that do not have the smallest rank-based Mahalanobis distances from any treated unit. In addition, unlike the CBRMD, which allocates weights for the control group unit based on the number of times a control unit has the smallest rank-based Mahalanobis distance

from the individual treated units, we consider obtaining the control group weights from the ranked distances in the raw form. They are then raised by a constant factor λ, within the range (0, 1).

The general framework of the proposed method is explained in Section 5.2. In Section 5.3, we used a real-life dataset to study the effect of varying the values of parameter λ, on the stability of the weights and covariate balance. In Section 5.4, the comparative performance of the proposed method was performed using the Kang & Schafer (2007) simulation design. Additionally, a bias-variance trade-off of the different values of parameter λ, was explored using the simulated dataset. We report on bias and root mean squared error (RMSE). The simulation results are presented in Section 5.5. We analyzed a case study in Section 5.6.

5.2 Methodology Proposed

The definitions made in Section 4.2 still holds in this new proposal. The central message of this chapter is to redefine how the weights are computed from the rank-based Mahalanobis distances, differently from the CBRMD algorithm described in Section 4.2. In the CBRMD methodology, the control group is allocated weights based on the frequency distribution of units having the smallest rank-based Mahalanobis distances, while the remaining control group units, which could not take values of 1, are uniformly allocated weights by a constant factor, as in Equation (4.3).

By allocating weights of equal values to control group units that did not have the smallest rank-based Mahalanobis distances, the weights allocation in Equation (4.3) implies that the algorithm does not take into account the magnitude of the corresponding distances of the remaining control units. This is improved upon by utilizing the matrix d_{rank} in Equation (4.2), as described in the following algorithm:

The above-described methodology, termed as the improved covariate balancing rank-

Algorithm 2 iCBRMD Algorithm

Step 1: Sort the data in order of the treatment indicator, with corresponding unit identification number. Rank the data.
Step 2: Compute the rank-based Mahalanobis distances of each treated units with the control units, using Equation (4.2), and store the distances in a matrix with t rows and c columns.
Step 3: For each row of matrix d_{rank}, the distances are ranked in ascending order of magnitude.
Step 4: The ranks obtained in Step 2 are raised to a constant factor $\lambda : \lambda \in (0, 1)$ as λ^{j-1}, where j is the rank of the corresponding control unit.
Step 5: The optimal value of λ that yields the minmax standardized mean difference on all the covariates is selected.
Step 6: Treated units have weights that are fixed at 1, while control units have weights based on Step 5.

based Mahalanobis distance matching method (iCBRMD), have weights for the control group units with $j = 1$ being the same as that of the CBRMD. This implies that we are still able to maintain a clear demarcation of control units that have the smallest distances with each treated unit from others. In other words, for $j = 1$, λ^{j-1} attains its maximum for any value of λ, but decreases at a constant rate, with increasing values of j.

The proposed methodology can be implemented through a web-based Shiny application written in R statistical language, which is being hosted by the RStudio server at https://amusasuxes.shinyapps.io/iCBRMD/.

5.3 Exploring Parameters of the Proposed Method with Real-life Data

The Ohio Heart Health Center (OHHC) operators at the Lindner Christ Hospital in Cincinnati, Ohio carried out an observational study in 1997. In brief, the Lindner dataset comprises information on 996 patients who received an initial Percutaneous Coronary Intervention (PCI), received at the health facility at that time. The treated group are patients who received the PCI with an additional treatment abciximab

(abcix) - an expensive, high-molecular-weight IIb/IIIa cascade blocker, while the control group are those who received the PCI alone. Covariates include, an indicator for recent acute myocardial infarction (acutemi), left ventricle ejection fraction (ejecfrac), height, number of vessels involved in initial PCI (ves1proc), an indicator for coronary stent insertion (stent), gender (female), diabetic indicator (diabetic) and an indicator for survival at six months (sixMonthSurvive). Details of this dataset and its analysis have been published elsewhere (Abdia et al., 2017; Kereiakes et al., 2000).

We apply the proposed methodology to the Lindner dataset provided in the twang package (McCaffrey et al., 2013) of R software. Specifically, we examined covariate balance and stability of weights, at different λ values. The aim is to explore the effect of changing these parameter λ values, on the analysis of interest.

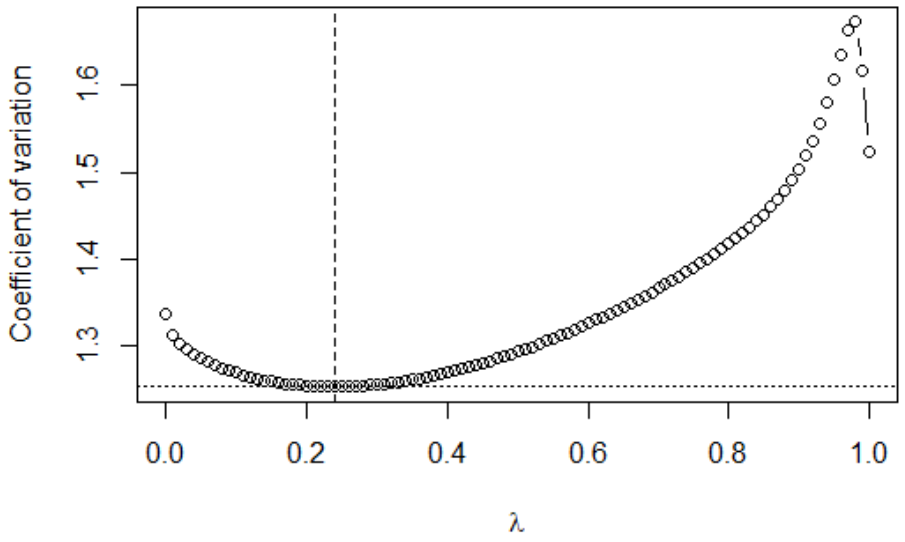

Figure 5.1: Relationship between the adjustment parameter and the variability of the weights in the Lindner study

We report the coefficient of variation of the weights (Figure 5.1) and the corresponding covariate balance (Table 5.1), from varying the values of parameter λ for the proposed method. Both Figure 5.1 and Table 5.1 are an indication of the importance

Table 5.1: Assessment of covariate balance for different levels of weights adjustment in the Lindner study

	Absolute standardized mean difference				
Covariates	Unadjusted	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
stent	0.25	0.08	0.25	0.25	0.25
height	0.00	0.02	0.00	0.00	0.00
female	0.11	0.01	0.11	0.11	0.11
diabetic	0.15	0.08	0.15	0.15	0.15
acutemi	0.37	0.19	0.37	0.37	0.37
ejecfrac	0.18	0.01	0.18	0.18	0.18
veslproc	0.43	0.03	0.43	0.43	0.43
survival	0.19	0.03	0.19	0.19	0.19

of the parameter λ. Figure 5.1, for example, shows how changing λ from 0 to 1 can vary the coefficient of variation of the weights between 1.25 and 1.67. The optimal λ value appears to be 0.24 for this dataset. There seems to be an increasing trend from that optimal point, as well as a turning point at about $\lambda = 0.95$. For Table 5.1, the effect of changing the λ values on the improvement or otherwise of the balance induced on the covariates, was observed. The least absolute standardized mean difference (ASMD) values were observed for the covariates at $\lambda = 0.01$. However, λ values of at least 0.1, increased the covariate imbalances and produced high ASMD values. Further, λ values in this range did not provide substantially different ASMD values.

It is noteworthy that the λ value which produced the most stable weights ($\lambda = 0.24$), contributed high ASMD values and worsened the covariate imbalances (results not shown). Indeed, there is a clear trade-off between balance and variability. It is therefore essential to explore and regulate this trade-off for a given dataset.

5.4 Simulation Study

We modified the Kang & Schafer (2007) simulation study, which has been described in Chapter 4 and has been the standard for evaluating the performance of pre-

processing techniques in causal inference (Amusa et al., 2019b; Imai & Ratkovic, 2014; Zubizarreta, 2015). We randomly generated 1000 samples, each of which comprised sizes of $n = 200, 1000$. In alignment with practice reality, we assumed that the treatment assignment model is unknown, by misspecifying the coefficients of the model terms by assuming $\pi_i = \frac{1}{1+e^{-(X_{i1}+X_{i2}-X_{i3}-X_{i4})}}$.

We calculated the absolute bias and root mean square error (RMSE) and assessed the bias-variance trade-off for a grid of equally spaced λ values, between 0 and 1. Finally, the stability of the proposed method weights was evaluated relative to the IPW and CBRMD weights, in terms of the mean coefficient of variation (CV). Extreme or outlying weights were assessed with the mean 95th and 99th percentiles, calculated across the simulations. Also, for outcome assessments, the performance of the iCBRMD method was evaluated, relative to the other methods, in terms of absolute bias and RMSE of the estimated treatment effects.

5.5 Simulation Results

Figure 5.2 show the absolute bias (left panel) and RMSE (right panel) of estimated treatment effects for $n = 200$ and $n = 1000$, respectively. These figures show how the values of λ can drastically vary the absolute bias and RMSE of estimated treatment effects. Optimal λ values for the absolute bias were observed at 0.85 and 0.43, for the corresponding sample sizes n =200 and n = 1000, respectively. The RMSEs, on the other hand, had optimal λ values at 0.99, for both sample sizes. There was further a trend of increasing λ values, resulting in reduced RMSEs. Irrespective of the λ values, the proposed method had smaller bias and RMSE values for the larger sample size ($n = 1000$).

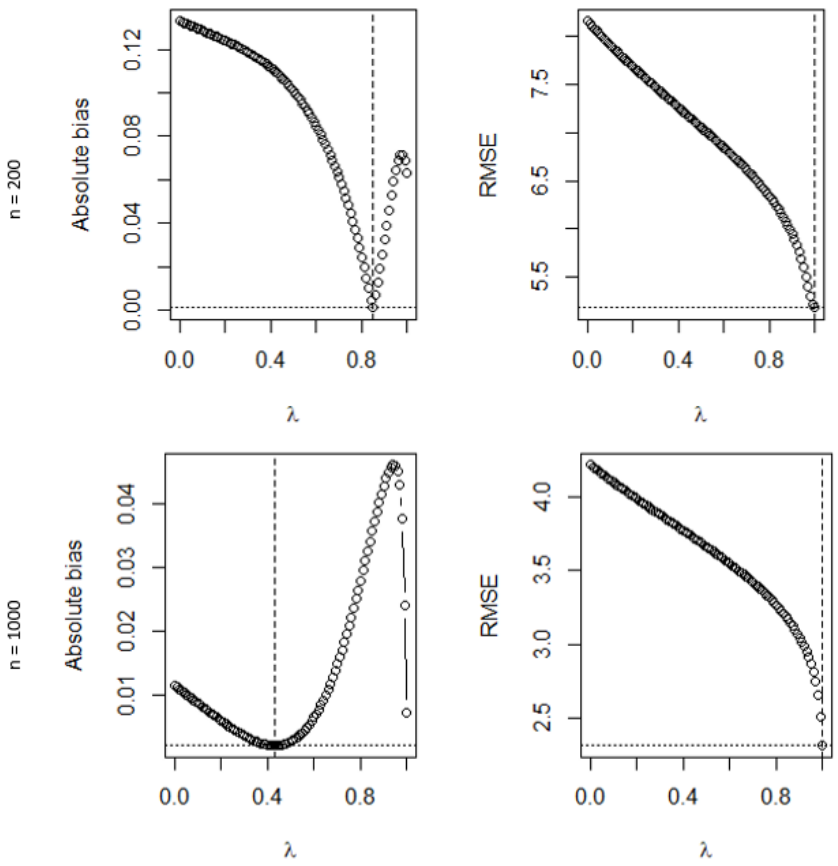

Figure 5.2: Absolute bias (left panel) and RMSE (right panel) of the iCBRMD method across different levels of adjustment from the simulation study

In terms of stability, the CV of weights from the proposed method (at any λ value) was always lower than the other adjustment techniques. The proposed technique, with higher values of λ, produced more stable weights in this simulation study. Aberrant or outlying weights were measured using the 95th and 99th percentiles. While the IPW method produced less aberrant weights, the CBRMD method had the lowest 99th percentile value for sample size $n = 200$, the proposed method provided massively extreme weights at higher λ values. Smaller λ values, like $\lambda = 0.01$, produced reasonable weights.

Results from Table 5.2 show that the adjustment techniques indicated an overall reduction in absolute bias and RMSE, as compared to the unadjusted data. The absolute bias and RMSE of the proposed method, regardless of the values of parameter

Table 5.2: Weight diagnostics and performance assessment of different methods in the simulation study

		Stability			Outcome Assessment	
Sample size	Method	CV	P95	P99	Absolute bias	RMSE
n=200	Unadjusted	-	-	-	25.00	25.43
	IPW	1.73	2.05	6.54	1.72	8.51
	PSM	-	-	-	3.08	5.91
	CBRMD	1.28	2.95	6.42	0.13	8.15
	iCBRMD (λ=0.01)	1.28	2.98	6.46	0.13	8.12
	iCBRMD (λ=0.1)	1.00	1001.70	1001.70	0.07	5.19
	iCBRMD (λ=0.5)	1.00	200.34	200.34	0.06	5.19
	iCBRMD (λ=0.9)	0.99	100.17	100.17	0.06	5.19
n=1000	Unadjusted	-	-	-	24.94	25.03
	IPW	1.99	2.14	6.91	0.53	5.14
	PSM	-	-	-	3.98	6.11
	CBRMD	1.53	2.95	7.53	0.01	4.21
	iCBRMD (λ=0.01)	1.53	2.97	7.58	0.01	4.20
	iCBRMD (λ=0.1)	1.00	1109.00	1109.00	0.01	2.32
	iCBRMD (λ=0.5)	1.00	587.12	587.12	0.01	2.32
	iCBRMD (λ=0.9)	1.00	499.05	499.05	0.01	2.32

λ, are smaller than those of the other methods. From Table 5.2, it is worth noting that even though higher values of λ for the proposed method produced better results in terms of bias and efficiency. It is pertinent to explore within reason, values of λ that produced stable and less aberrant weights.

5.6 Case Study: The Lalonde Data

We further applied the proposed iCBRMD method to the Lalonde-PSID data that was described in Chapter 4. The effectiveness of iCBRMD for reducing selection bias, was evaluated along with the PSM, IPW, and CBRMD methods. In the causal inference literature, the primary advice to this point, has been to select the method that yields the best balance (Harder et al., 2010; Ho et al., 2007; Rubin, 2006). Though defining the best balance is complex, since it involves trading off balance on multiple covariates. For this case study, we fixed the optimal value of λ that resulted in the fewest number of large absolute standardized mean differences (ASMD).

While we superimposed a horizontal line to denote ASMD of 0.25, Figure 5.3 shows a summary of selection bias (measured by ASMD) before applying the adjustment techniques and reduction in the covariate imbalance post adjustment. Before any adjustment, there was a high degree of covariate imbalance in the original data, with ASMD values ranging from 0.114 to 5.446. The iCBRMD method substantially improved the balance on the ten covariates, with average ASMD values ranging from 0.001 to 0.220. Except for two covariates, IPW considerably reduced the covariate imbalance. PSM did not perform well, as it only induced sufficient amount of balance on three covariates. Both CBRMD and iCBRMD methods performed remarkably well in reducing covariate imbalances, as they both achieved ASMD values well below the threshold of 25%. Overall, iCBRMD outperformed the other three methods.

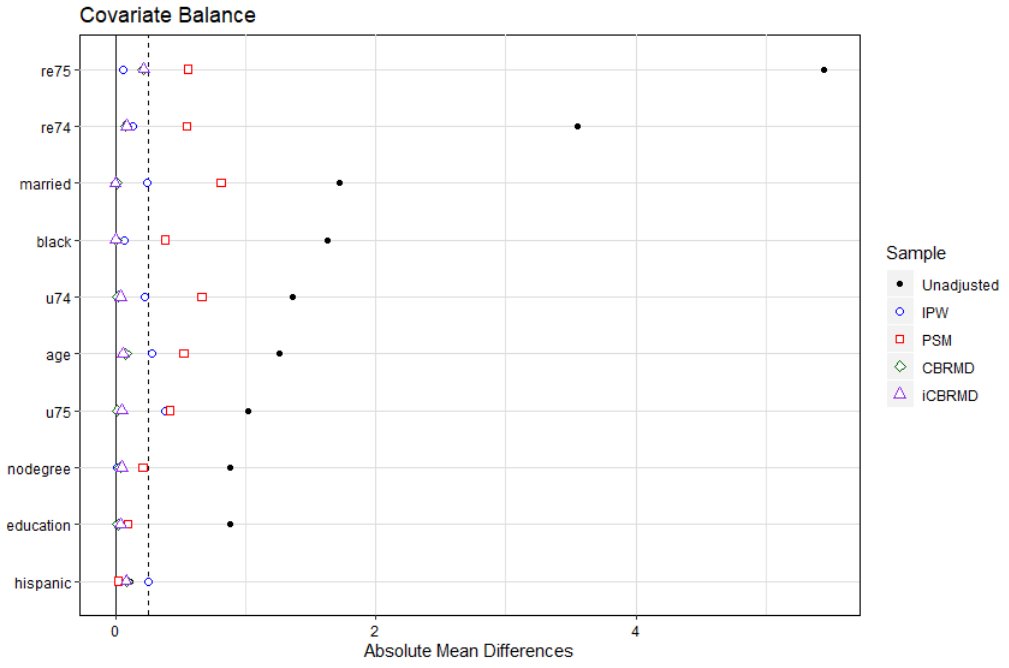

Figure 5.3: Assessment of covariate balance

We next is to evaluate the treatment effects. Due to the stark imbalance of the Lalonde data, they are generally regarded as a tricky adjustment problem (the unadjusted difference in mean outcomes is far away from the experimental target at

$-15204.8). A simple difference in means of the experimental version of the data yielded an average of $1794 with a 95% confidence interval of [551, 3038]. Using weights obtained from the adjustment techniques, we calculated a weighted difference in means for each of them. For PSM, we calculated a simple difference in means for the matched data. We performed bootstrapping (2000 samples) to produce 95% confidence intervals (CIs), which has been shown to account for uncertainty in the matching procedure (Stuart, 2010). Table 5.3 shows the difference in means estimates and their associated 95% confidence interval from the adjustment methods, relative to the raw data. The difference in means between the treated group and the reweighted control group from the iCBRMD method, yielded an ATT estimate of $1752.2, with a 95% confidence interval of [15, 3876] - an estimate that is close to the experimental target. Except for PSM, the adjusted estimates suggested that the job training programs significantly increased postintervention earnings. This is thus an agreement with findings from previous studies (Hainmueller, 2012; Diamond & Sekhon, 2013).

Table 5.3: Recovering the Lalonde's Experiment using its nonexperimental version

Estimator	Difference in means	95% Confidence Interval
Unadjusted	-15204.8	(-17468.80, -12940.75)
PSM	-977.1	(-2797.2, 1061.8)
IPW	2796.2	(1304, 4964)
CBRMD	2064.5	(150, 4803)
iCBRMD	1752.2	(15, 3876)

Note: Standard errors of the weighted estimators were bootstrapped with 2000 replicates

5.7 Discussion and Conclusion

In this chapter, we propose iCBRMD - a technique for estimating causal effects in observational studies. It was built on the previous proposal of the CBRMD methodology (Amusa et al., 2019a). We have demonstrated numerically, the satisfactory performance of the proposed iCBRMD method to induce balance on background co-

variates, as well as demonstrating a striking decrease in the bias and an increase in efficiency of the estimated treatment effects. The proposed method produced more balanced data than did the other matching methods considered. It also competed favourably with the other considered methods in terms of the accuracy of the estimated treatment effects. Both small and large samples produced similar results in terms of comparison with the other considered methods.

The proposed method further performs better with increasing sample sizes. It has a particularly interesting parameter that provides insight into the behaviour of less variable and less aberrant weights that can maximize covariate balance. Our proposal has opened doors to explore the option of λ values within reason, to maximize the desired objectives of the analysis of a given dataset. If the coefficient of variation does not increase substantially, then it may be worth tightening the covariate balance. For outcome analysis, the obtained weights must be assessed for aberrance. However, achieving sufficient covariate balance is the ultimate goal, hence, if the optimal value of λ in that regard produces aberrant or outlying weights, it might be necessary to trim the weights, as was done for PS weighting in previous studies (Lee et al., 2011; Stürmer et al., 2010).

The PS model specifications were the default, off-the-shelf versions of each of the methods since that is what most applied researchers would likely do. For example, logistic regression with carefully chosen interactions may perform better than the simple main effects-only model used here. We, therefore, acknowledge that the propensity score methods could perform better if the PS model specifications were tweaked; thus, we avoid overstating the superiority of the proposed method over PS methods. Future work may include the consideration of extensions and applications of the proposed method to a variety of other settings. Most importantly, and currently being explored by us, is the automation of selection of the optimal parameter λ value that can simultaneously produce stable weights that maximizes

covariate balance, while reducing bias and increase efficiency. Further, It is recommended to extensively study the effect of trimming weights obtained from the proposed method on its performance.

This new proposal does not only serve as an improved version of the CBRMD method, but can also be regarded as its generalized version, as it allows us to vary parameter values that optimize the desired cost function (covariate balance and efficiency) that may be of interest to the analyst.

A major limitation of the method proposed in this chapter, as well as the other methods investigated in Chapters 3 and 4, is that they cannot directly incorporate covariate balance into the weight function applied to the sample units. Methods that exploit this forehand knowledge about the data typically reweights observations appropriately to achieve balance, but at the same time, keeps the weights as close as possible to the base weights to prevent loss of information, while retaining efficiency for the subsequent analysis. We will thus explore one of these methods in the next chapter.

Chapter 6

Optimal Balance Weighting Methods: Entropy Balancing

In the previous chapters, the introduced adjustment techniques require going through a diagnostic step to ensure that the covariates are sufficiently balanced. Even when these methods provide adequate covariate balance on many variables, one or two of them might not be adequately balanced. In this chapter, we introduce entropy balancing, an adjustment technique, which belongs to the family of empirical calibration weighting (EBCW) methods discussed in Section 2.2.2. These EBCW methods, otherwise known an optimization-based methods, have been utilized in the literature (Li et al., 2018; Chan et al., 2016; Zubizarreta, 2015; Imai & Ratkovic, 2014; Hainmueller, 2012). These methods are automated covariate balancing methods, they have have an inbuilt facility of directly incorporating a balance condition for the moments (not just the means) of the covariates in the estimation procedure, thereby ensuring perfect covariate balance. Accordingly, the conventional balance checking is not necessary for such methods.

6.1 Background

More recently, weighting methods have taken centre stage in efficiently estimating treatment effects when treatment assignment is confounded with background

covariates. Though non-technically speaking, the matching methods described in Chapters 3, 4 and 5 are also weighting methods with discrete weights and can only produce finite many possible weights. The weighting methods discussed in this chapter, do not have this constraint and are inherently different from matching.

Weighting is a nonparametric balancing strategy, which applies weights to sample units to match the distribution of a target population. The literature on weighting methods which agree with the first general weighting approach described above, has been dominated by the inverse probability weighting (IPW) method, originating from survey research (Crump et al., 2009; Hirano & Imbens, 2001; Hirano et al., 2003; Imbens, 2004). IPW method is the most common weighting adjustment to applied researchers and practitioners, especially in the medical and health sciences (Austin & Stuart, 2015).

Despite their popularity and relatively high usage, propensity score (PS) methods, with specific reference to IPW, rely heavily on the correct specification of the PS model, as slight misspecification of the PS model will result in a substantial bias of the estimated treatment effects (Kang & Schafer, 2007). It takes a highly skilled user to specify what is close to a correct PS model. Consequently, iteratively tweaking the PS models, until the measured baseline covariates are balanced, can be quite tedious. Despite this cycle of attempting to fit the correct PS model, achieving a sufficient level of covariate balance can occasionally be elusive and additional imbalances may be introduced when using the IPW method.

We present entropy balancing - an optimization-based weighting method, which shares the spirit of the first general weighting approach described above. Entropy balancing (Hainmueller, 2012) achieves covariate balance remarkably. Relative to the other optimization-based weighting methods, we were particularly interested in entropy balancing, because it is the oldest and more computationally attractive.

Prior studies (Harvey et al., 2017; Setodji et al., 2017; Amusa et al., 2019d,c) that compared some EBCW techniques with the IPW method, confirmed the better performance and computational simplicity of the entropy balancing technique.

We aim to provide an intensive exploration of the use of entropy balancing for health services/outcomes research. Not many applications of this highly effective method have been utilized in the medical and health literature for balancing in quasi-experimental research designs. An in-depth search from the Web of Science Core Collection, excluding methodology-based articles, identified 170 published articles that utilized entropy balancing, and only a few of them (26.19%) were in the medical and health sciences. A majority of the applications of entropy balancing have been in the social sciences. A few of these applications in the medical and health literature can be found in Adhikary et al. (2016); Brettschneider et al. (2017); Grupp et al. (2017); Mattke et al. (2015); Pearson et al. (2014).

Using the IPW method as a benchmark, the performance of entropy balancing was examined via Monte Carlo simulations, modelling situations typical of the medical and health sciences. Finally, we illustrate the application of entropy balancing with an empirical case study, exploring changes in its various parameters, as well as its effect on achieving balance on the measured baseline covariates, further focusing also on accuracy and precision in estimating treatment effects.

6.2 Entropy Balancing Technique

The definitions made in Section 4.2 are still applicable here. In this section, while estimating the average treatment effect among the treated (ATT), we describe the entropy balancing, for adjusting the inherent non-randomization of treatments that is characterized by an observational study. The IPW method has been briefly described in Section 2.2.1.

Entropy balancing is a preprocessing method that utilizes a maximum-entropy reweighting scheme to directly incorporate covariate balance in terms of means or/and higher-order moments into the weight function (Hainmueller, 2012). In other words, it assigns a scalar weight to each sample unit such that the reweighted groups satisfy a set of balance constraints that are imposed on the sample moments of the covariate distributions. Entropy balancing can, therefore, guarantee perfect covariate balance, as well as maximum retention of information (Parish et al., 2018). The reweighting scheme belongs to the family of maximum-entropy methods, which has roots in information theory and applied statistics (Kullback, 1959; Golan, 2018; Ciavolino & Carpita, 2015; Aria et al., 2018). The weights w_i are selected to minimize the relative entropy:

$$\min_{w_i} H(\omega) = \min_{w_i} \sum_{i|T=0} w_i \, log(w_i/q_i) \, , \tag{6.1}$$

subject to the constraints:

$$\sum_{i|T=0} w_i \, c_{ri} \, (X_i) = m_r, \text{for } r = 1, \ldots, R \tag{6.2}$$

$$\sum_{i|T=0} w_i = 1 \tag{6.3}$$

$$w_i \geq 0, \forall i, \text{ for } T = 0. \tag{6.4}$$

The above optimization problem minimizes the loss function $H(\omega)$ to obtain weights that satisfy the balance conditions for the user-specified covariate functions $c_{ri} \, (X_i) = m_r$ imposed on the covariate moments of the reweighted control group.

The loss function $H(\omega)$ is a distance metric defined by the directed Kullback (1959) entropy divergence, with estimated weights w_i and base weights q_i. A vector of uniform weights, with $q_i = \frac{1}{n_c}$ is usually set as the base weights. Let $\mathbf{X}_i = (X_{i1}, ..., X_{iK})$ denote a $K-$ dimensional vector of observed pre-treatment covariates associated with unit i. We denote $m_r = \frac{1}{n_t} \sum_{i|T=1} c_{ri} \, (X_i)$ as the formulation containing the rth order moment of a given variable $X \in \{X_1, ..., X_k\}$ from the treated group,

while the moment functions are specified for the control group as $c_{ri}\ (X_i) = X_i^r$ or $c_{ri}\ (X_i) = (X_i - E(X_i))^r$, with mean $E(X_i)$. Equation (6.2) is the balance constraint specified in terms of the rth moment to be achieved on all covariates. Equation (6.3) is the normalization constraint, which ensures that the weights sum to a normalized constant of one. Equation (6.4) is the non-negativity constraint, because the distance metric is not defined for negative weight values.

To obtain the entropy balancing weights, we minimized the loss function $H(\omega)$ subject to the constraints given in Equations (6.2) to (6.4). Using the Lagrange multiplier, the primal optimization problem is given as

$$\begin{aligned}\min_{\omega,\lambda_0,\lambda} L^p = \sum_{i|T=0} w_i\ log(w_i/q_i) + \sum_{r=1}^{R} \lambda_r \left(\sum_{i|T=0} w_i\ c_{ri}\ (X_i) - m_r \right) \\ + (\lambda_0 - 1) \left(\sum_{i|T=0} w_i - 1 \right),\end{aligned} \tag{6.5}$$

where $\lambda_0 - 1$ is the Lagrange multiplier for the normalization constraint, and λ_r is the Lagrange multiplier for the rth balance constraint.

Given that the dimensionality of the system of equations in (6.5) is $n_c + R + 1$, it is computationally inconvenient. Therefore, we construct the optimization problem as a dual formulation so that the dimension in (6.5) can be reduced. To do so, we first obtained the optimal solution for each weight w_i^*, by using the Karush-Kuhn-Tucker (KKT) condition, also known as the first derivative test:

$$\frac{\partial L^p}{\partial w_i} = (log(w_i/q_i) + 1) + \sum_r \lambda_r\ c_{ri}\ (X_i) + (\lambda_0 - 1) = 0 \tag{6.6}$$

and

$$\frac{\partial L^p}{\partial \lambda_0} = \sum_{i|T=0} w_i - 1 = 0. \tag{6.7}$$

Equation (6.6) becomes

$$log(w_i/q_i) = -1 - \sum_r \lambda_r \, c_{ri} \, (X_i) - (\lambda_0 - 1),$$

from which it follows that

$$w_i = \frac{q_i \, e^{\left(-\sum_r \lambda_r \, c_{ri} \, (X_i)\right)}}{e^{\lambda_0}}. \tag{6.8}$$

Since $\sum_{i|T=0} w_i = 1$ from (6.7), we have

$$\sum_{i|T=0} q_i e^{\left(-\lambda_0 - \sum_r \lambda_r \, c_{ri} \, (X_i)\right)} = 1$$

$$e^{\lambda_0} = \sum_{i|T=0} q_i e^{\left(-\sum_r \lambda_r \, c_{ri} \, (X_i)\right)}$$

$$\lambda_0 = log \left(\sum_{i|T=0} q_i e^{(-\sum_r \lambda_r \, c_{ri} \, (X_i))} \right). \tag{6.9}$$

Plugging (6.9) into (6.8), we have the optimal solution for each weight as

$$w_i^* = \frac{q_i \, e^{\left(-\sum_{r=1}^{R} \lambda_r \, c_{ri} \, (X_i)\right)}}{\sum_{i|T=0} \, q_i \, e^{\left(-\sum_{r=1}^{R} \lambda_r \, c_{ri} \, (X_i)\right)}}. \tag{6.10}$$

Since $\sum_{i|T=0} w_i = 1$, we next formulate the dual problem as

$$L^d = \sum_{i|T=0} w_i \, log(w_i/q_i) + \sum_{r=1}^{R} \lambda_r \left(\sum_{i|T=0} w_i \, c_{ri} \, (X_i) - m_r \right).$$

The superscript d indicates it is a dual. We justify its duality by inserting (6.10) back into (6.5), which eliminates the constraints and leads to an unrestricted dual problem, as follows:

$$L^d = \sum_{i|T=0} w_i \, log \left(\frac{q_i \, e^{-\sum_{r=1}^{R} \lambda_r \, c_{ri} \, (X_i)}}{q_i \, \sum_{i|T=0} q_i \, e^{-\sum_{r=1}^{R} \lambda_r \, c_{ri} \, (X_i)}} \right) + \sum_{r=1}^{R} \lambda_r \left(\sum_{i|T=0} w_i \, c_{ri} \, (X_i) - m_r \right)$$

$$L^d = \sum_{i|T=0} w_i \left(-\sum_{r=1}^{R} \lambda_r \, c_{ri}\,(X_i) - log(\sum_{i|T=0} q_i \, e^{-\sum_{r=1}^{R} \lambda_r \, c_{ri}\,(X_i)}) \right) + \sum_{r=1}^{R} \lambda_r \left(\sum_{i|T=0} w_i \, c_{ri}\,(X_i) - m_r \right).$$

It then follows that

$$L^d = -\left(log \sum_{i|T=0} q_i e^{-\sum_{r=1}^{R} \lambda_r \, c_{ri}\,(X_i)} + \sum_{r=1}^{R} \lambda_r m_r \right). \tag{6.11}$$

Since duality holds, the dual of a minimization problem will be a maximization problem. In other words, we minimize $-L^d$. Thus, we multiply (6.11) by -1 to get the right hand side of (6.12) as

$$\min_{\lambda} L^d = log \left(\sum_{i|T=0} q_i \, e^{\left(-\sum_{r=1}^{R} \lambda_r \, c_{ri}\,(X_i)\right)} \right) + \sum_{r=1}^{R} \lambda_r \, m_r. \tag{6.12}$$

Equation (6.12) is composed of only two components: the entropy objective function and a linear combination of the rth order moments with their associated Lagrange multipliers. The solution to the dual problem λ^* solves the primal problem. This dual problem is an unconstrained optimization problem, which is much more tractable than a constrained one. Additionally, the dimension of the problem decreases substantially, as $n_c + R + 1$ is reduced to a system of nonlinear equations in the R Lagrange multipliers. Moreover, if there exists a solution, it will be unique, since L^d is strictly convex.

A Levenberg-Marquardt scheme is used to find λ^* for the dual problem in (6.12). Let $\omega = [w_1, ..., w_{n_c}]'$ and $\mathbf{q} = [q_1, ..., q_{n_c}]'$. The constraints are written in matrix form by

defining the $(R \times n_c)$ constraint matrix

$$\mathbf{C} = \begin{bmatrix} c_{11}(X_1) & c_{12}(X_2) & \cdots & c_{1n_c}(X_{n_c}) \\ c_{21}(X_1) & c_{22}(X_2) & \cdots & c_{2n_c}(X_{n_c}) \\ \vdots & \vdots & \ddots & \vdots \\ c_{R1}(X_1) & c_{R2}(X_2) & \cdots & c_{Rn_c}(X_{n_c}) \end{bmatrix}$$

and the vector of moments $\mathbf{m} = (m_1, \ldots, m_R)'$, while $\lambda = (\lambda_1, \ldots, \lambda_R)'$ is a vector of Lagrange multipliers for the balance constraints.

$$e^{-\mathbf{C}'\lambda} = \begin{bmatrix} e^{-\sum_{r=1}^{R} \lambda_r c_r(X_1)} \\ e^{-\sum_{r=1}^{R} \lambda_r c_r(X_2)} \\ \vdots \\ e^{-\sum_{r=1}^{R} \lambda_r c_r(X_{n_c})} \end{bmatrix}$$

The rewritten problem is thus given as

$$\min_{\lambda} L^d = log(\mathbf{q}'\, e^{-\mathbf{C}'\lambda}) + \mathbf{m}'\lambda. \tag{6.13}$$

Equation (6.10) can be rewritten as

$$w_i^* = \frac{(q_i\, e^{-\mathbf{C}'\lambda})}{(\mathbf{q}'\, e^{-\mathbf{C}'\lambda})}. \tag{6.14}$$

The balance constraints are stated as $\mathbf{C}\omega = \mathbf{m}$. The gradient is given as $\nabla_\lambda(L^d) = \frac{\partial L^d}{\partial \lambda} = \mathbf{m} - \mathbf{C}\omega$ and the Hessian matrix $\nabla^2_\lambda(L^d) = \frac{\partial^2 L^d}{\partial \lambda^2} = \mathbf{C}[D(\omega) - \omega\omega']\mathbf{C}'$, where $D(\omega)$ is a n_c-dimensional diagonal matrix, with ω in the diagonal. This 2nd-order information is then utilized in a Newton iteration method as follows:

Initialize $\lambda^0 = (\mathbf{C}\mathbf{C}')^{-1}\mathbf{m}$.

Iterate $\lambda^{new} = \lambda^{old} - l\, (\nabla^2_\lambda\, (L^d))^{-1}\, (\nabla_\lambda\, (L^d))$,

where $l \in (0, 1)$ is a scalar denoting the step size, which changes at every iteration.

The step size l quantifies how big a step is taken in the direction of the minimum.The optimal step size is then selected for each iteration. This iterative algorithm is globally convergent as long as the problem is feasible.

The ATT weights (Parish et al., 2018) are defined for the entropy balancing as fixing treated units weight at unity and reweighting the control group units using the algorithm described above.

6.3 Simulation Study

We conducted a set of Monte Carlo simulations to examine the performance of entropy balancing, relative to the IPW method. We made the simulations to be typical of biomedical studies by considering binary outcomes (Austin et al., 2010; Austin & Stuart, 2017). Entropy Balancing was performed with the R-package *ebal* (version 0.1-6) (Hainmueller, 2014).

6.4 Data Generation

Following the simulation structure of previous studies (Lee et al., 2010; Setoguchi et al., 2008), we generated data from ten covariates that are standard normal distributed, with an inducement of specified levels of dependence between some pairs of covariates to be more reflective of practical settings. Six of the covariates were dichotomised. Figure 6.1 gives the causal structure of the simulations.

As shown in Figure 6.1, the simulation study aligns with practice reality: there are confounders associated with both treatment and outcome (X_1, X_2, X_3, X_4), predictors of the treatment variable only (X_5, X_6, X_7) and predictors of the outcome variable only (X_8, X_9, X_{10}).

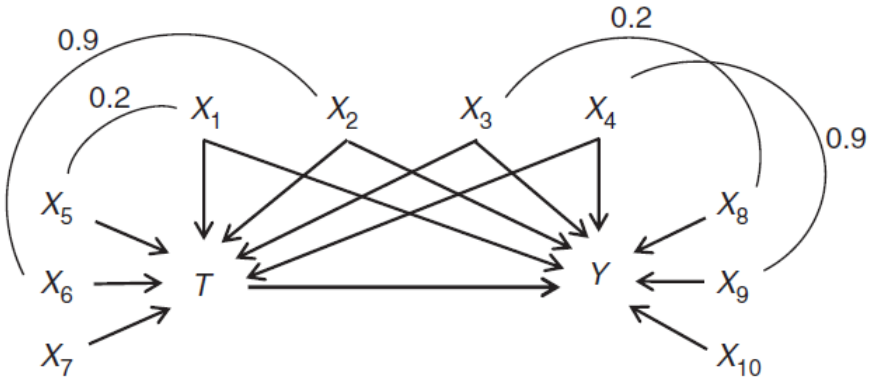

Figure 6.1: Data structure of the simulation study, where $X_1, X_3, X_5, X_6, X_8, X_9, Y$ are binary.

The treatment variable T, was modelled as a logit model of the form

$$log\left(\frac{P_{i,trt}}{1-P_{i,trt}}\right) = \alpha_{0,trt}+\alpha_1 X_1+\alpha_2 X_2+\alpha_3 X_3+\alpha_4 X_4+\alpha_5 X_5+\alpha_6 X_6+\alpha_7 X_7\,. \quad (6.15)$$

The outcome variable Y, was modelled as a logit model of the form

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_8+\beta_6 X_9+\beta_7 X_{10}+\beta_{trt} T_i\,. \quad (6.16)$$

The coefficients α and β were based on real-life data utilized in a previous study (Setoguchi et al., 2008).

We simulated two potential outcomes, Y^1 and Y^0, for treated and control groups, respectively. In each simulated dataset, we computed the marginal risk difference as $\bar{Y}^1_{T=1} - \bar{Y}^0_{T=1}$, where $\bar{Y}^1_{T=1}$ and $\bar{Y}^0_{T=1}$ denote the mean potential outcome under the treated and control groups, respectively, in those units which ultimately belong to the treated group.

6.4.1 Varying Factors

Our simulations varied, based on two factors, (i) Sample size: $n = 500, 2000$; (ii) the proportion of units who received the treatment (prevalence of treatment). The

value of $\alpha_{0,trt}$ was selected such that the prevalence of treatment was fixed at $\pi = 25\%, 33\%, 50\%$, and 67%, corresponding to treated:control units ratio of $1:3, 1:2$, $1:1$, and $2:1$, respectively. We developed the following iterative algorithm, which was used to determine the value of $\alpha_{0,trt}$ that induced targeted prevalence π (Amusa et al., 2019a):

We varied values of $\alpha_{0,trt}$ within reason (-3 to 3 in this case) and simulated n units. For all the considered $\alpha_{0,trt}$ values, the corresponding individual values were computed using Equation (6.15), while the treatment variables $T_i \sim Ber(P_{i,trt})$ were generated, and the mean of each T_i correspond to π. This process was repeated 1000 times to increase the precision of the estimation, while the value of $\alpha_{0,trt}$ which correspond to the desired π was chosen.

Even though there are not many study designs where the proportion of the treated group is higher than the control group, we included it to satisfy our curiosity. The intercept of the treatment assignment model was modified to ensure that the treatment variable had the specified target prevalence in the simulated datasets.

6.4.2 Analyses and Performance Assessment of Estimates

For each of the considered scenarios, we simulated 1000 datasets and obtained ATT weights each for entropy balancing and IPW methods, and further estimated the risk differences obtained from the weighted regressions of Y on T. The risk difference estimates were then averaged over the simulation runs, denoted by δ_i. The marginal risk difference, denoted by δ is the true ATT.

We utilized the absolute standardized mean difference (ASMD) to examine covariate balance, while absolute bias and MSE were used to examine the accuracy of estimated treatment effects. Finally, we calculate the standard errors based on a robust sandwich-type variance estimator, as well as the 95% confidence interval (CI)

coverage. CI coverage is defined as the proportion of times the estimated confidence intervals contain the specified parameter value (Burton et al., 2006). Accordingly, we calculated the percentage of the 1000 estimated confidence intervals containing the true risk difference, for the two weighting strategies.

6.5 Results

We present the results for the simulation study according to each of the performance metrics explained in Section 6.4.2. We emphasise on the results of the entropy balancing method, using the IPW results as a threshold for gauging the performance of entropy balancing.

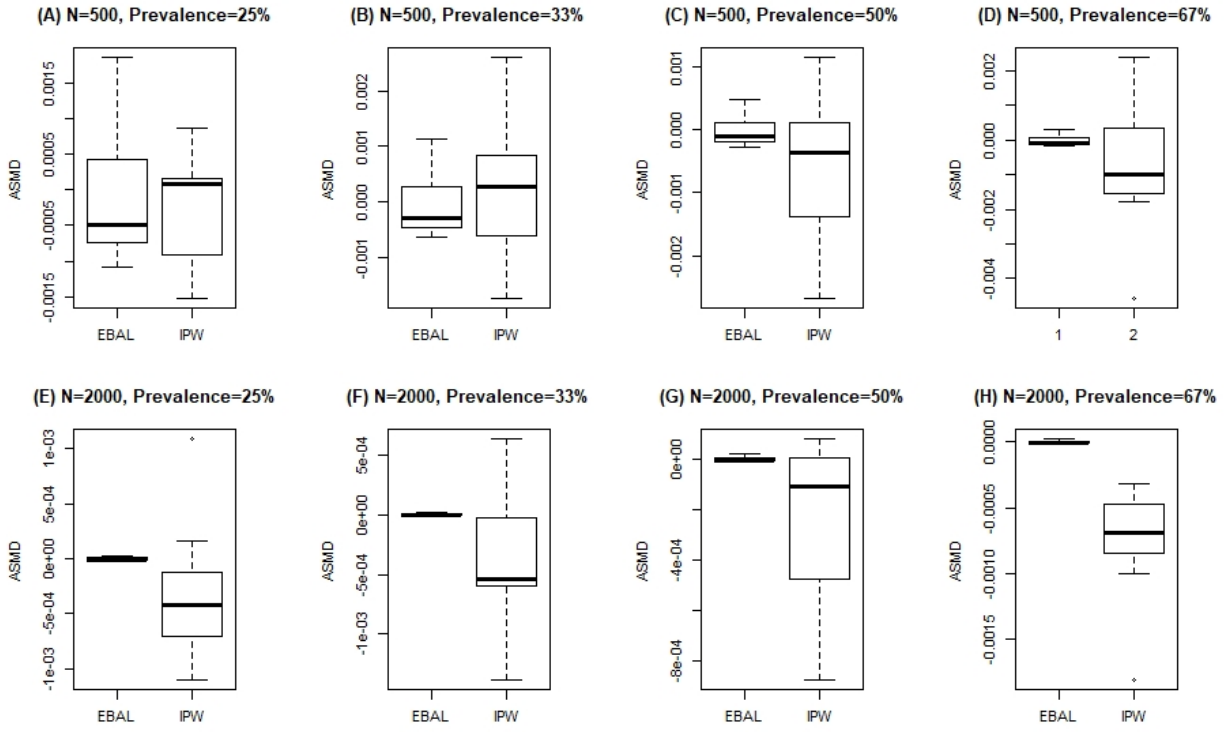

Figure 6.2: Boxplots for the absolute standardized mean difference for covariates. The values for each covariate were averaged from the simulations.

Some authors suggested that ASMD values above 10% may be indicative of covariate imbalance (Mamdani et al., 2005; Normand et al., 2001). As shown in Figure 6.2, both entropy balancing and IPW methods performed remarkably well in reducing

covariate imbalances, as they both achieved ASMD values well below the threshold of 10%. However, entropy balancing outperformed the IPW method, with marginal improvements observed for smaller treatment prevalences (25% and 33%), but a substantial outperformance was evidenced for higher treatment prevalences (50% and 67%). Additionally, entropy balancing produced better balance for large sample size ($n = 2000$), with ASMDs achieving a perfect balance (ASMD=0) on almost all of the covariates, across the rates of treatment prevalence.

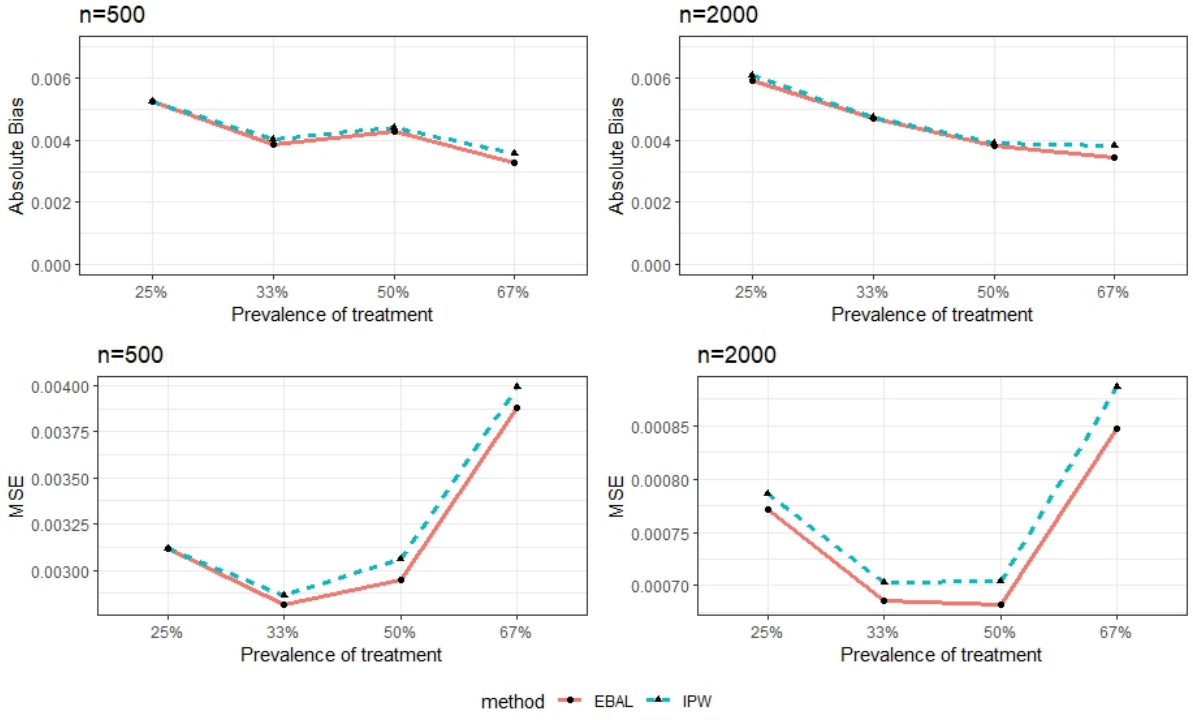

Figure 6.3: Absolute Bias (top panel) and MSE (bottom panel) of estimated treatment effects

In terms of bias, Figure 6.3 shows that entropy balancing produced less biased estimates across the board. The MSE of the treatment effect estimates is described in Figure 6.3. Entropy balancing resulted in estimates with substantially lower MSE values. Furthermore, there was no apparent effect of the prevalence of treatment on both the bias and MSE.

As shown in Figure 6.4, the two methods produced very similar standard errors.

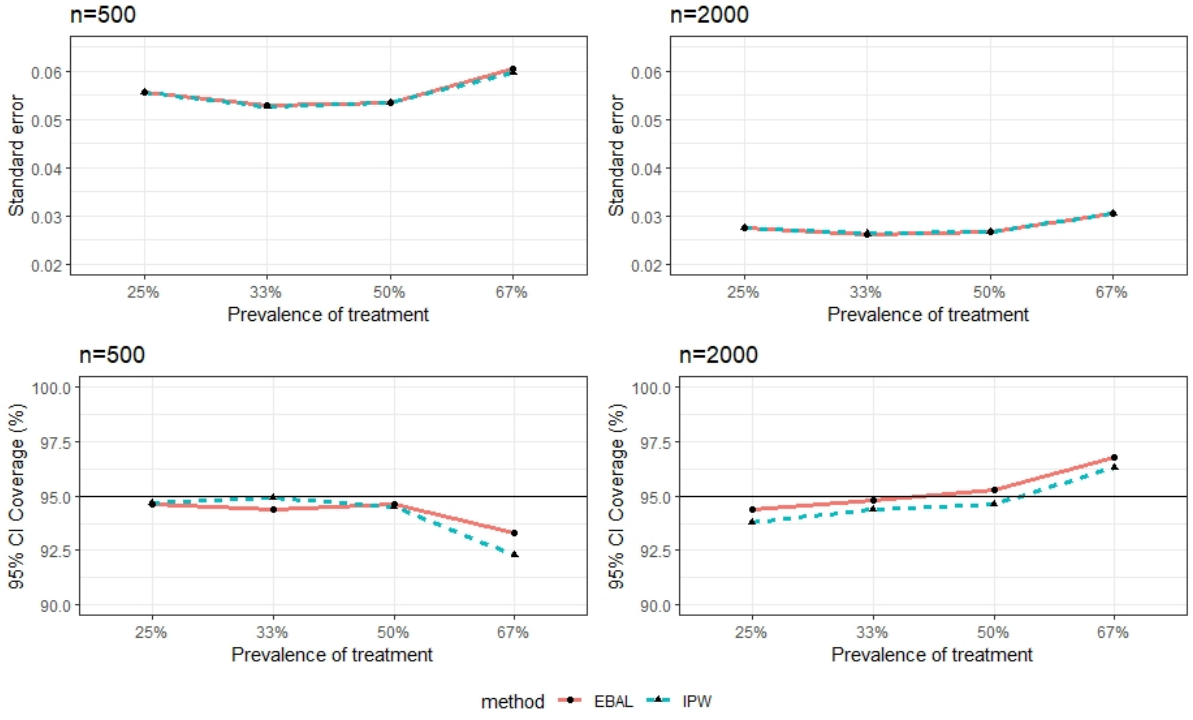

Figure 6.4: SE (top panel) and 95% CI coverage (bottom panel) of estimated treatment effects

Superior 95% CI coverages were observed for entropy balancing when the sample size is relatively large ($n = 2000$). However, for a relatively smaller sample size ($n = 500$), entropy balancing produced higher coverage for higher treatment prevalence (50% and 67%), as shown in Figure 6.4.

6.6 Case study: A Re-analysis of Data on Right Heart Catheterization

We further explored the entropy balancing technique by analyzing observational data (Murphy & Cluff, 1990) to study the effectiveness of right heart catheterization (RHC) for critically ill patients. A few influential studies have also re-analyzed the data using different adjustment methods (Crump et al., 2009; Hirano & Imbens, 2001; Li et al., 2018; Rosenbaum, 2012). In brief, the dataset comprises information on 5735 patients, 2184 (38.1%) of were treated with RHC ($T_i = 1$) within 24 hours

of admission and 3551 (61.9%) did not receive the RHC treatment ($T_i = 0$). The outcome of interest was mortality at 30 days of admission. Full details of this data, including the variable description and its summary statistics, have been published elsewhere (Connors et al., 1996; Hirano & Imbens, 2001).

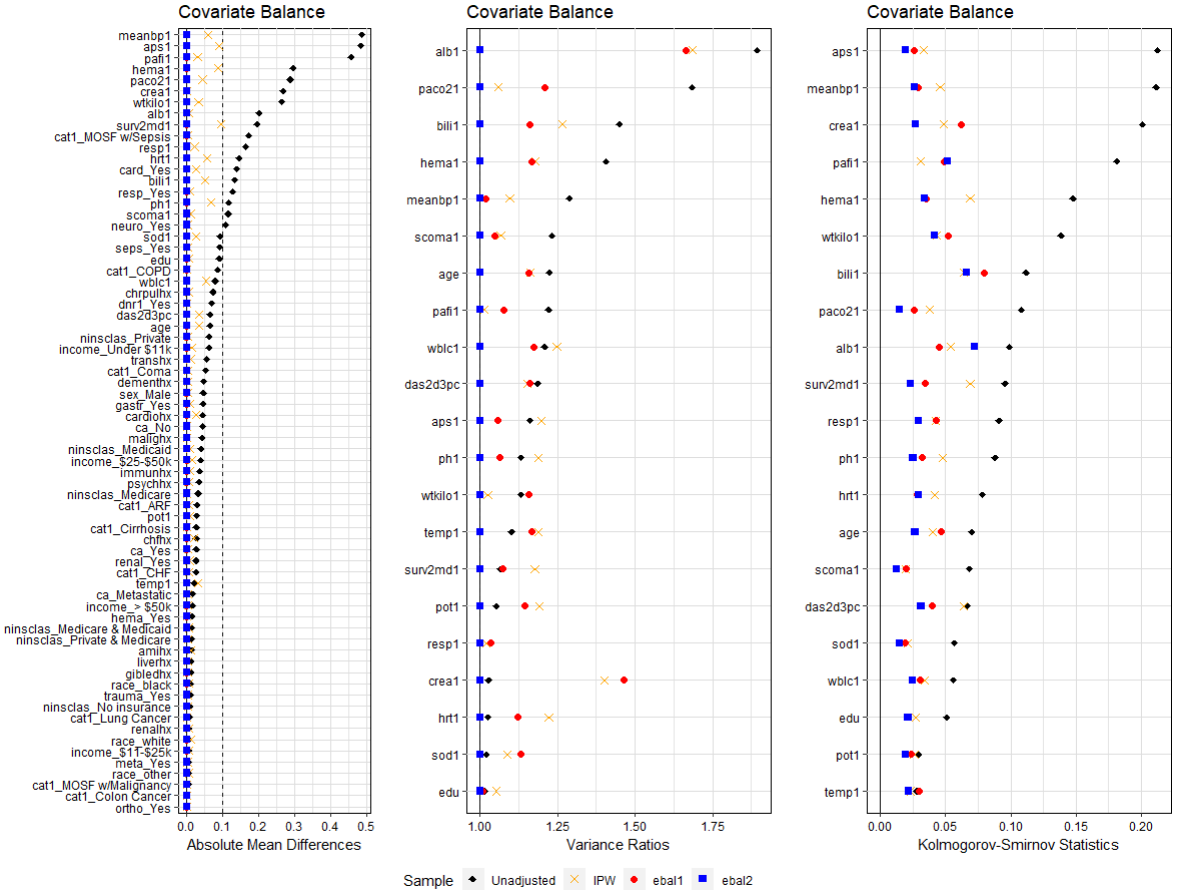

Figure 6.5: Assessment of covariate balance for the various methods. *Note: ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment*

6.6.1 Balance Diagnostics

We applied diagnostics for assessing the covariate balance in the data weighted by entropy balancing (EB), with its performance evaluated relative to IPW. We did not restrict balance to means only, but also investigted variance and the empirical distribution of continuous covariates. For balance on the means alone, we considered the ASMD. For balance on higher-order moments, we adopted variance ratios, which some authors (Rubin, 2001) recognized values close to one as acceptable, and the

Kolmogorov-Smirnov (KS) statistic, which when close to zero is satisfactory (Ali et al., 2015). For entropy balancing , we considered the first and second moments on which covariate balance is desired. We attempted to include constraints up to the third moments, as well as interactions between pairs of continuous covariates, but the EB algorithm did not converge.

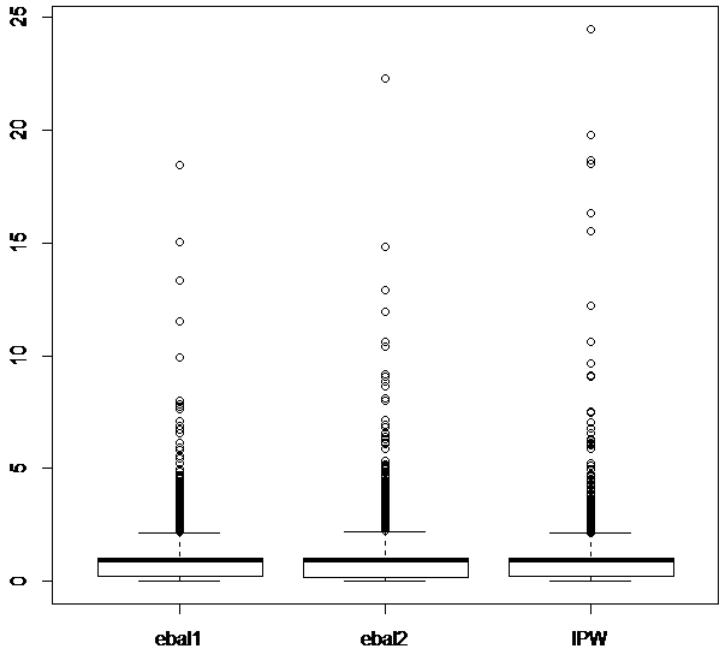

Figure 6.6: Distribution of weights for the control group units for the RHC dataset. *Note: ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment*

Figure 6.5 provides information on the balance on covariates. Before any weighting, there is a high degree of confounding in the original data. The weighting methods produced ASMD values that did not exceed the vertical line 0.1 threshold superimposed in Figure 6.5 . Entropy balancing achieved a perfect covariate balance (ASMD=0) on all the covariates, while there were still noticeably some non-zero ASMDs after applying IPW. As expected, variance ratios were all virtually 1 when moment constraints of entropy balancing included the second moment. Even when moment constraints included only the mean, entropy balancing still achieved vari-

ance ratios close to 1 about 71.4% of the time. IPW did poorly on variance ratios, as it increased the values for some covariates, thereby making things worse. As measured by the KS statistic, both methods performed remarkably well on the empirical distribution of the continuous covariates. However, about 67% of the time, entropy balancing produced KS values indicative of better balance, when moment constraints included the second moment; while, when it contained only the first moment, entropy balancing marginally outperformed IPW about 52% of the time.

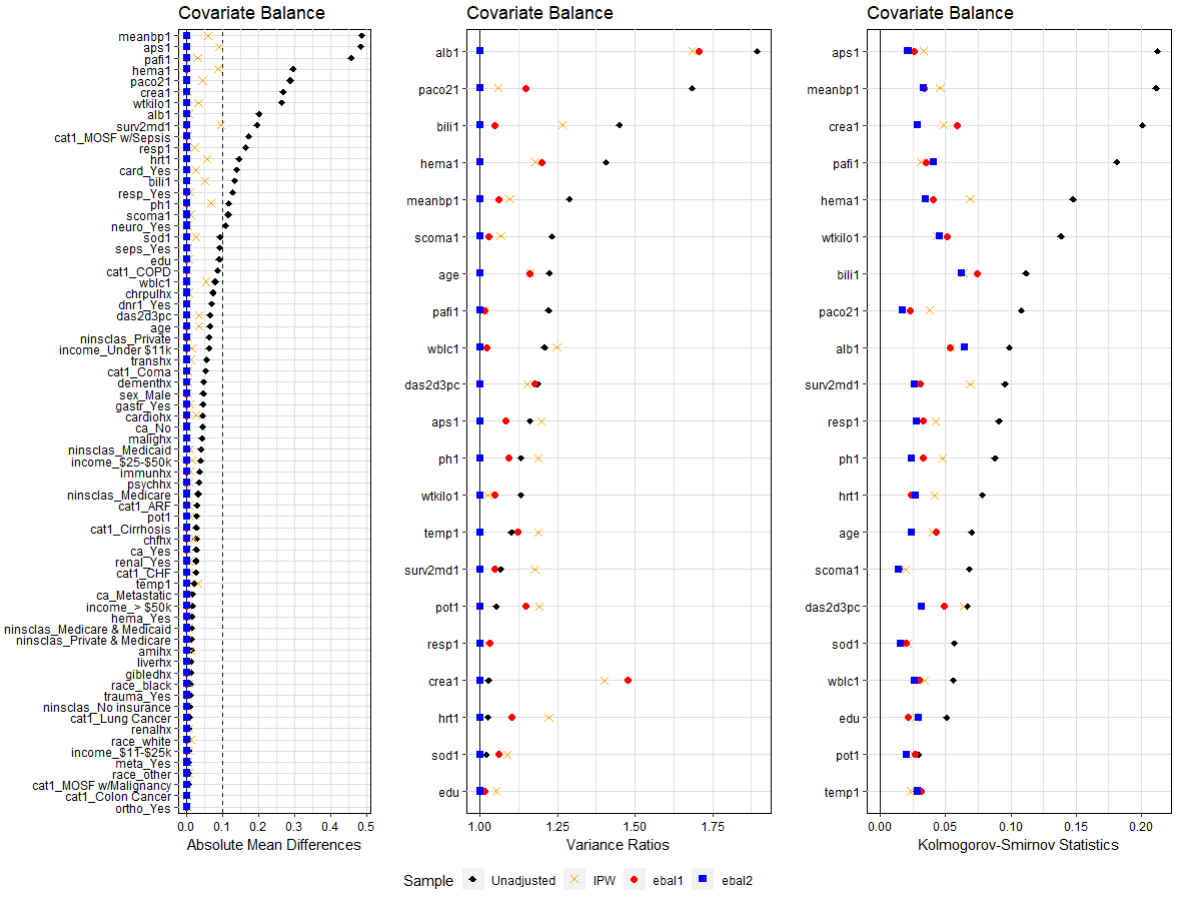

Figure 6.7: Assessment of covariate balance for the various methods after weights trimming. *Note: ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment*

6.6.2 Weight Diagnostics

Both methods produced many outlying and highly skewed weights, as shown in Figure 6.6. Entropy balancing, when moment constraints included the variance,

produced weights with mean, maximum, standard deviation and skewness equal to $0.76, 18.45, 0.79$ and 6.58, respectively. When moment constraints included only the means, it produced weights with mean, maximum, standard deviation and skewness equal to $0.76, 22.28, 0.87$ and 7.21, respectively. IPW produced weights with mean, maximum, standard deviation, and skewness equal to $0.77, 24.49, 0.92$, and 10.48, respectively. Based on these diagnostics, entropy balancing marginally produced less extreme and outlying weights, even though it might be of interest to trim them. Accordingly, we trimmed the entropy balancing weights to confirm if they still produce sufficient covariate balance, and they did. Results are shown in Figure 6.7.

6.6.3 Outcome Analyses

Next, we estimate the average treatment effect for those who received RHC. Alongside the IPW method, we applied weights to the outcome modelling from entropy balancing, when moment constraints included the means only, denoted by *ebal1*, as well as when it added the variance, denoted by *ebal2*. Using logistic regression to regress the occurrence of death at 30 days of admission, we adopted the risk difference, as suggested by clinical commentators (Cook & Sackett, 1995; Laupacis et al., 1988; Sackett et al., 1996; Schechtman, 2002), as the estimate of interest. The model incorporated the weights induced by entropy balancing and IPW. Standard errors of the weighted estimators were estimated using the sandwich-type variance estimators.

The causal treatment effects estimated using these stated methods are shown in Table 6.1. All the considered estimators produced qualitatively similar estimates that are statistically significant at the 0.01 level, which indicate that applying RHC leads to a higher mortality rate. These results agree with the substantive conclusions made in previous studies (Connors et al., 1996; Crump et al., 2009; Li et al., 2018). Both IPW and entropy balancing methods produced very close standard errors. Estimators

based on the entropy balancing had smaller confidence interval lengths (0.0731 for *ebal1* and 0.0767 for *ebal2*) than the corresponding ones based on the IPW (0.0787). As shown in Table 6.1, trimming the weights for the entropy balancing estimators, slightly reduced the corresponding length of confidence intervals.

Table 6.1: Causal effect estimation of RHC, using the various methods

Methods	Raw entropy balance weights			Trimmed entropy balance weights		
	Estimate	CI	P-value	Estimate	CI	P-value
Unweighted	0.051	0.025-0.076	< 0.0001	-	-	-
IPW	0.056	0.017-0.096	0.0051	-	-	-
ebal1	0.072	0.035-0.108	0.0001	0.057	0.023-0.091	0.001
ebal2	0.057	0.019-0.096	0.0034	0.043	0.006-0.079	0.023

Note: CI: Confidence Interval; ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment

6.7 Discussion and Conclusion

Propensity score weighting methods have conventionally been used to estimate treatment effects in the presence of confounding factors. We used simulations and an empirical example to highlight our experiences with using entropy balancing, and its performance relative to the inverse probability weighting method. We are motivated by the under-utilization of the entropy balancing technique in the biomedical sciences, despite its increased usage and successful application in the social sciences. We chose a simulation structure that mimics what is common in most biomedical studies. Our empirical data have also been analyzed by many previous studies representative of a clinical application.

Entropy balancing aims to achieve covariate balance between the treatment groups so that valid estimates of the treatment effect can be obtained. Though both entropy balancing and IPW methods provided adequate covariate balance, we found that entropy balancing outperformed IPW in terms of all the considered performance metrics. Relative to IPW, entropy balancing improved covariate balance as treatment

prevalence increased. Both methods improved covariate balance for larger sample sizes. There was also no evidence of treatment prevalence on the bias and MSE of estimated effects.

The empirical application showed that IPW worsened covariate balance on a few covariates. This could have been remedied by iteratively tweaking the PS model until the desired covariate balance is achieved. However, unlike the entropy balancing method, there is no guarantee that this tedious and exhausting process of PS model specification will ensure that IPW produces the desired covariate balance. Both methods produced extreme weights that may require trimming. Accordingly, it is recommended to extensively study the effect of trimming entropy balancing weights on its performance, as was done for PS weighting in a previous study (Lee et al., 2011).

We attempted the following situations which did not allow convergence of the entropy balancing algorithms: (i) smaller sample sizes (less than 300) for treatment prevalence rates higher than 33%, in the simulations; (ii) including the 3rd moments in the moment constraints for the case study; (iii) including pairs of interaction of continuous covariates for the case study. The above findings agree with the caution given by Hainmueller (2012), in light of potential situations, depending on the data, that may prevent convergence of the entropy balancing algorithm. Furthermore, even though previous studies like (Zagar et al., 2017) stated that the presence of interaction effects might improve the performance of the entropy balancing, the interaction effects are not always feasible for a large number of covariates as we have experienced with our case study in Scenario III above.

To our knowledge, no previous study had explored entropy balancing using Monte Carlo simulations with binary outcomes. As with any simulation, our simulation results might be limited to the factors associated with our simulation data, therefore,

the results cannot be generalized to settings that have not been evaluated. Another limitation of entropy balancing is that it does not address unmeasured confounding, which is still a vexing problem in observational studies.

Overall, we found the entropy balancing technique useful, with excellent performance, and one that is frequently less tedious than the inverse probability weighting approach. Entropy balancing merits more widespread adoption for estimating the effects of treatment, especially in the medical and health sciences, when using observational data. In this chapter, while we have examined entropy balancing with an example on binary outcomes, it is imperative to extend the examination to some other standard estimators of treatment effects in the next chapter.

Chapter 7

Extending the Examination of Entropy Balancing

There is an increasing interest in using entropy balancing to estimate marginal or average treatment effects of different types of outcome (Adhikary et al., 2016; Brettschneider et al., 2017; Grupp et al., 2017; Mattke et al., 2015; Parish et al., 2018; Pearson et al., 2014). Accordingly, in this chapter, we investigate the performance of entropy balancing in evaluating treatment effects on continuous, binary, count, and time-to-event outcomes. Using the IPW method as a benchmark, we used Monte Carlo simulations to examine the performance of entropy balancing in estimating some measures of treatment effects. We considered the estimation of difference in means, odds ratios, rate ratios, and hazard ratios for the continuous, binary, count and time-to-event outcomes, respectively. We also utilized the average treatment effect among the treated (ATT) as our estimand of interest. This chapter is based on (Amusa et al., 2019c).

7.1 Simulation Study

We conducted a broad range of Monte Carlo simulations to evaluate the performance of entropy balancing in estimating treatment effects while using the IPW method as a benchmark. We considered continuous, binary, count and time-to-event

outcomes.

7.1.1 Data Generation

For the covariates and treatment variable generation, we used the same data generation scheme of Section 6.4. For each of the units, we generated an outcome Y_i conditional on T_i, and the seven covariates $(X_1, X_2, X_3, X_4, X_8, X_9, X_{10})$ associated with Y_i. We generated Y_i separately for continuous, binary, count, and time-to-event outcomes.

Continuous Outcomes

While we fixed the true treatment effect at $\delta = 1$, the continuous outcome was generated as

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i + \varepsilon_i, \quad (7.1)$$

where $\varepsilon_i \sim N(0, \sigma^2)$

Binary Outcomes

We generated a binary outcome as $Y_i \sim Bernoulli(P_i)$ using a logistic model:

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i \,. \quad (7.2)$$

Count Outcomes

We generated a count outcome as $Y_i \sim Poisson(\eta_i)$ using a Poisson model (Amusa et al., 2019b):

$$log(\eta_i) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i \,. \quad (7.3)$$

Time-to-event Outcomes

For time-to-event outcomes, we used a data-generating process described by a previous study (Bender et al., 2005). Survival times t_i are generated as

$$t_i = \left(\frac{-log(U_i)}{\lambda e^{LP}} \right)^{\frac{1}{v}}, \tag{7.4}$$

where $U_i \sim Uniform(0, 1)$, and the linear predictor, $LP = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i$. We set v = 2 and $\lambda = 0.000001$. This process generates survival times from a Cox-Weibull distribution. We assumed that all event times are observed for the current analyses.

7.1.2 Parameter Values for Data Generation

The regression coefficients in the outcome data generation took the values, $\beta_1 = \beta_2 = \beta_3 = log(2), \beta_4 = \beta_5 = \beta_6 = log(1.75)$ and $\beta_7 = log(1.5)$ to reflect very high, high, and moderate effect sizes (Austin et al., 2007; Austin, 2014).

For continuous outcomes, the standard deviation values were fixed at σ = 1 and 0.5. The conditional treatment effect β_{trt} values were fixed at log (1.5) and log (0.5) for odds ratios, hazard ratios and rate ratios. The chosen values of β_{trt} = log (1.5) and log (0.5) were aimed at reflecting beneficial $(\beta_{trt} > 0)$ and adverse $(\beta_{trt} < 0)$ treatment effect, respectively. In generating dichotomous outcomes, the β_0 value was set to ensure that the prevalence of the event of interest occurred for approximately 70% of the units. Finally, the above data-generating process has randomly generated treatment variable, covariates and four different outcomes each of size n units, while inducing a conditional treatment effect.

A conditional treatment effect is the average effect, at the individual or unit level, of moving a unit from control to treated group. In contrast, a marginal effect is the average effect, at the population level, of moving the whole population from control

to treated group (Greenland, 1987). Since the difference in means is collapsible, the conditional treatment effect coincides with the true marginal treatment effect. However, the other three treatment effects are not collapsible (Austin, 2013; Gail et al., 1984). Thus, for each of the conditional treatment effects (log-odds ratios, log-hazard ratios, and log-rate ratios), we determined their corresponding true marginal treatment effects. Details of this process of obtaining the true marginal treatment effect have been explained elsewhere (Austin, 2013, 2014; Austin & Stuart, 2017). The obtained true marginal treatment effect in the treated population from this process is regarded as the true ATT, for each of the considered treatment effects.

7.1.3 Statistical Analyses in Simulated Datasets

For a given treatment effect associated with each of the type of outcome considered, we randomly generated 1000 datasets of size 500, using the earlier described data-generating scheme. Using each of the simulated datasets, we separately estimated the different treatment effects, while utilizing each of the ATT weights of entropy balancing and IPW methods. The treatment effects, δ, were estimated from the following generalized linear model

$$g(E(Y|T)) = \beta_0 + \delta T, \tag{7.5}$$

where g was considered as the canonical link function for the normal linear model, logistic model, Poisson model, and Cox survival model for estimating the difference in means, odds ratios, hazard ratios, and rate ratios, respectively. We adopted the robust sandwich-type estimator for estimating the standard errors (Austin & Stuart, 2015; Joffe et al., 2004). We utilized the R-package ebal (Hainmueller, 2014) for implementing entropy balancing.

Let δ_i denote the ith estimated treatment effect using a given method, whereas δ is the true ATT. We then determined the following: Bias $= \frac{1}{1000}\sum_{i=1}^{1000}(\delta_i - \delta)$, mean squared error (MSE) $= \frac{1}{1000}\sum_{i=1}^{1000}(\delta_i - \delta)^2$. We also examined precision by averaging

the model-based standard errors (SE) over the 1000 simulated datasets. Finally, we examined 95% coverage, which is the proportion of times δ is enclosed in the 95% confidence interval of δ over the simulated datasets.

7.2 Results

We present the simulation results according to each of the type of estimated treatment effects explained in the earlier section. We focus on the performance of entropy balancing method, using the IPW method as a threshold for evaluating the results. As a form of sensitivity analysis, we ran simulations for other sample sizes $(n = 300, 1000)$, but we do not present the results as no qualitative differences were observed in the relative performance of the methods. However, we present results of the two standard deviation values $(\sigma = 1, 0.5)$ assumed while estimating difference in means, as well as the two different true ATT values, varied each for odds ratios, hazard ratios, and rates ratios estimation. Altering these parameter values also did not change the conclusions in all the scenarios, except for when rate ratios were estimated.

Continuous Outcomes: Difference in Means

Results are summarized in Figures 7.1 and 7.2. In terms of bias, Figure 7.1 shows that both methods produced estimates with very low (near-zero) bias. However, EB produced slightly higher biases, except for when the prevalence rate was 10%. For the MSE, EB outperformed IPW across the board (Figure 7.1). Both methods yielded very similar SE estimates, with the values decreasing with increasing prevalence rates (Figure 7.2). Though EB produced superior CI coverages - near perfect in most cases, both methods achieved reasonably high 95% CI coverages (Figure 7.2).

Binary Outcomes: Odds Ratios

In terms of bias, Figure 7.3 shows that EB consistently produced higher biased estimates. For the MSE, EB outperformed IPW across the board, with the values de-

creasing with increasing prevalence rates (Figure 7.3). Both methods yielded very similar SE estimates, with the values decreasing with rising prevalence rates (Figure 7.4). Though EB produced superior CI coverages, both techniques achieved reasonably high 95% CI coverages (Figure 7.4).

Count Outcomes: Rate Ratios

Figure 7.5 shows that the MSE of both methods increased as the treatment prevalence increased from 10% to 40%. When the conditional rate ratio was positive ($\beta_{trt} > 0$), EB consistently produced estimates with higher bias, higher MSE, and lower 95% CI coverages. However, for $\beta_{trt} < 0$, EB produced higher bias and MSE estimates only when the treatment prevalence was 20% or lower (Figure 7.5). The SE estimates were very similar for both methods, with the values decreasing with increasing prevalence rates (Figure 7.6). Both methods achieved reasonably high 95% CI coverage. Though EB had lower CI coverages when the conditional rate ratio was positive, it is not clear which of them produced higher coverage when the conditional rate ratio was negative (Figure 7.6).

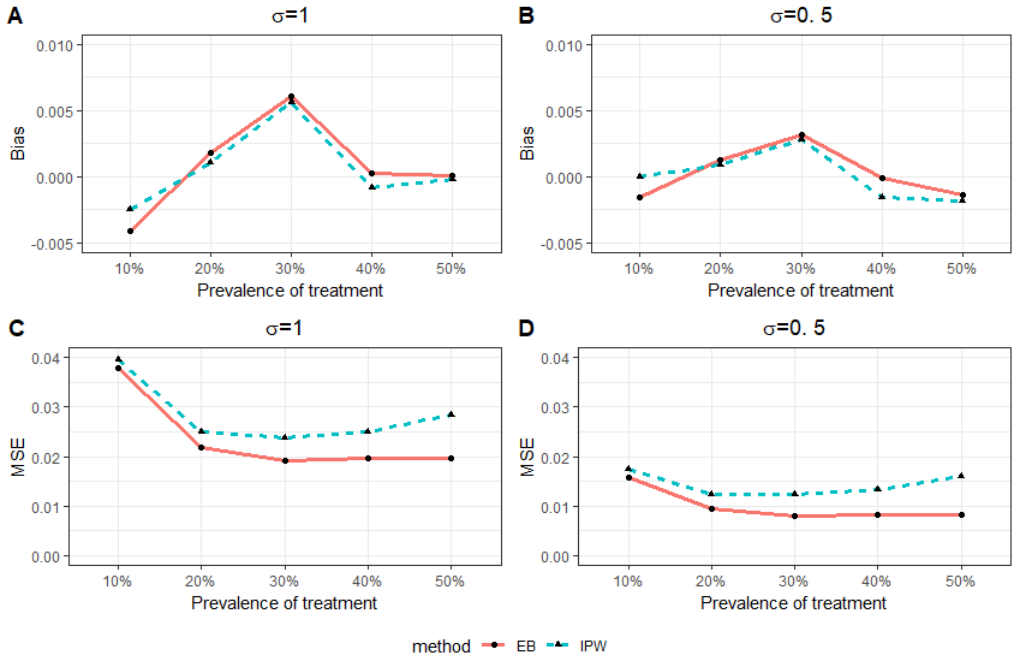

Figure 7.1: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating difference in means.

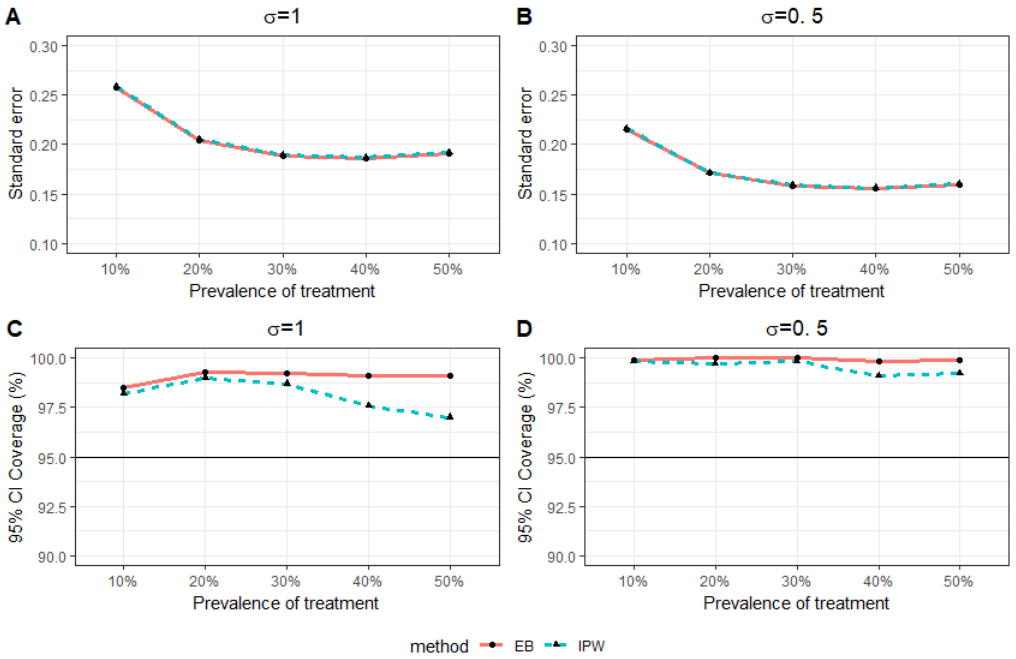

Figure 7.2: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating difference in means.

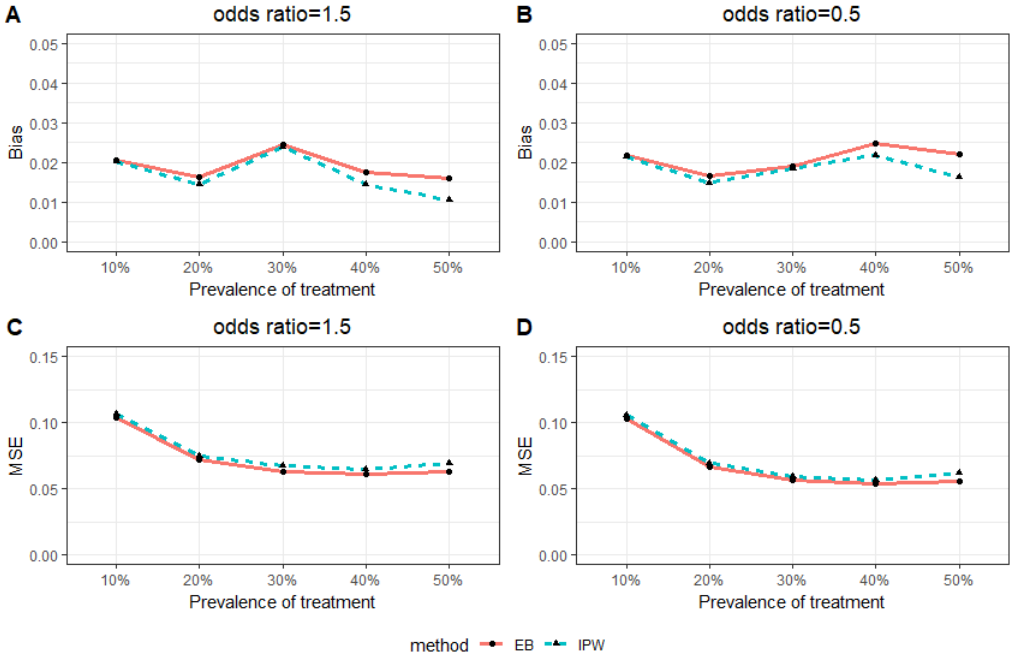

Figure 7.3: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating odds ratios.

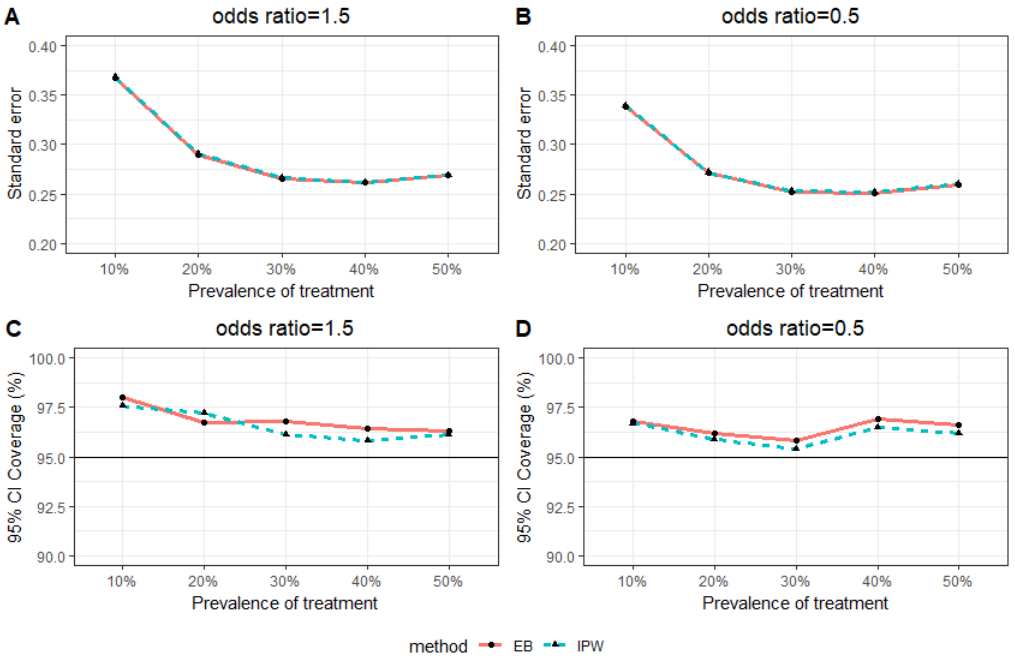

Figure 7.4: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating odds ratios.

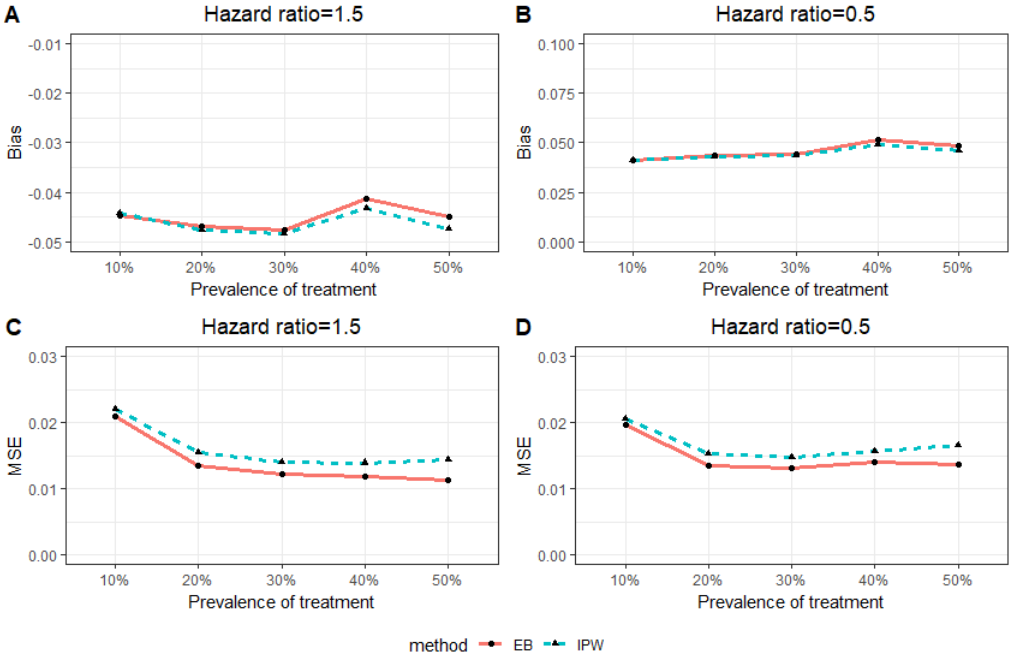

Figure 7.5: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating rate ratios.

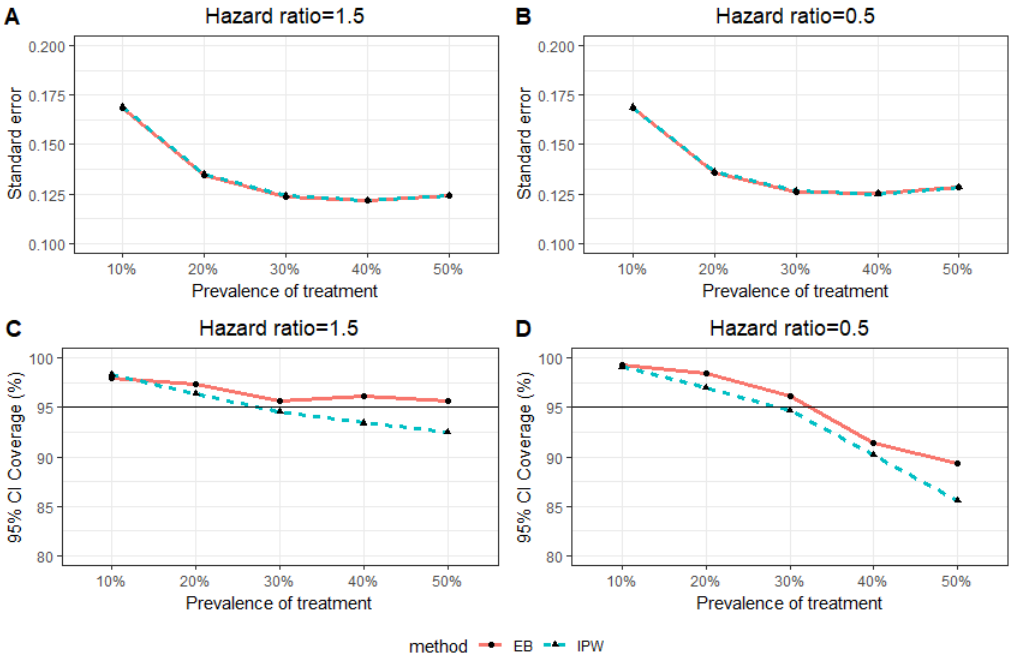

Figure 7.6: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating rate ratios.

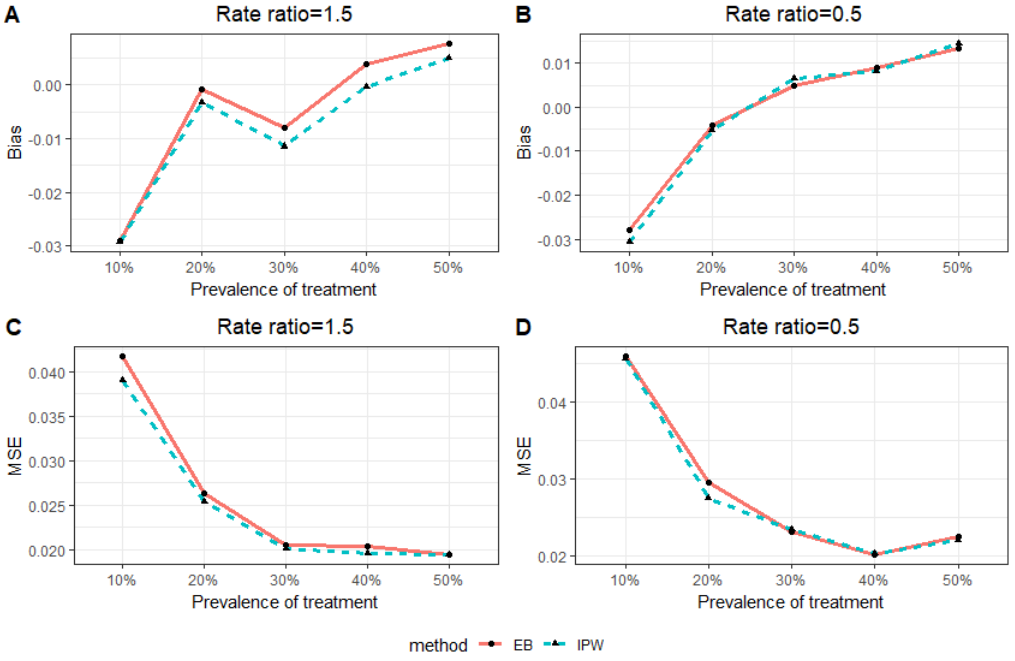

Figure 7.7: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating hazard ratios.

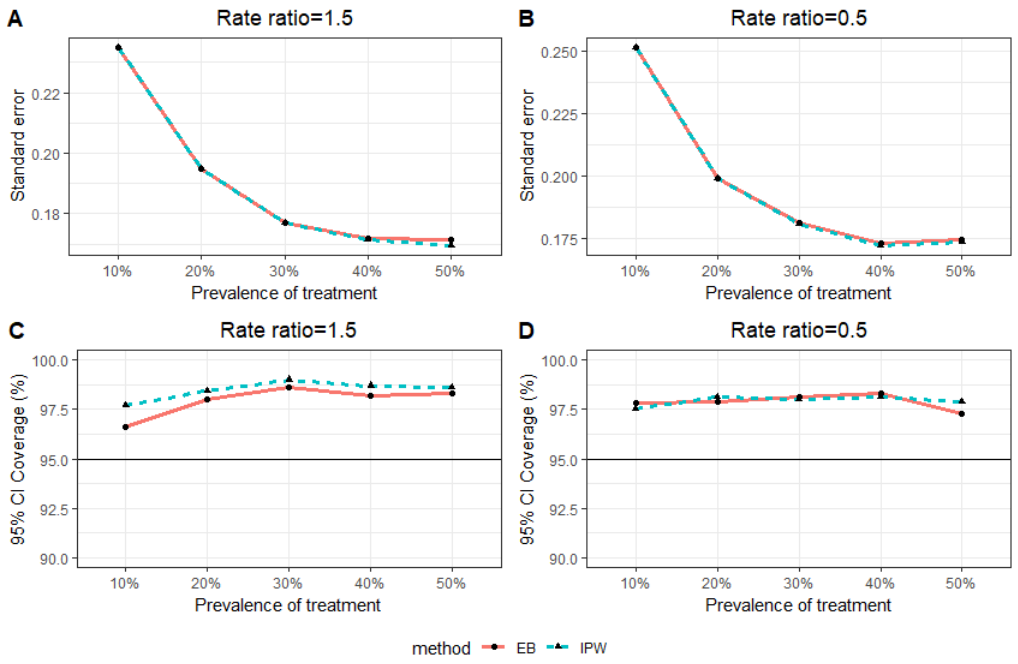

Figure 7.8: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating hazard ratios.

Time-to-event Outcomes: Hazard Ratios

Figure 7.7 shows that the bias of both methods are not substantially different, except for higher prevalence rates (40% and 50%), where EB produced higher bias estimates. For the MSE, EB consistently outperformed IPW (Figure 7.7). As shown in Figure 7.8, the SE estimates were again very similar between both methods, with the values decreasing with increasing prevalence rates. Figure 7.8 illustrates that when the actual hazard ratio = 0.5, EB produced 95% CIs slightly below the nominal coverage rate at prevalence rates higher than 30%. However, EB provided superior CI coverages overall.

7.3 Discussion and Conclusion

We utilized Monte Carlo simulations to evaluate the performance of entropy balancing, relative to the IPW method, in estimating some standard measures of treatment effect. While focusing on entropy balancing, we summarize our findings and where

necessary, place them in the context of existing literature.

Though both methods performed reasonably well in estimating the various treatment effects considered, we found that on average, entropy balancing outperformed IPW for all the considered situations. However, a few exceptions were found: (i) When rate ratios were estimated, entropy balancing tended to produce estimates with slightly higher biases and mean squared errors. Although they considered conditional and not marginal treatment effects, a previous study by Austin (2007) found that conditioning on the propensity score did not substantially introduce bias into the estimation of rate ratios. (ii) The model-based standard errors for both IPW and entropy balancing methods were consistently indistinguishable. (iii) In terms of bias, across all the estimated treatment effects, entropy balancing consistently produced more biased estimates. Hence, there is an interesting bias-variance trade-off of the two techniques. However, entropy balancing has the facility to optimize the bias-variance trade-off by tightening the pre-specified tolerance on covariate balance (Harvey et al., 2017). Previous studies (Austin, 2007, 2013; Austin & Stuart, 2017) also support our findings in favour of IPW producing an unbiased estimation of odds ratios and hazard ratios.

A significant strength of this study is in the use of an algorithm which determines the true marginal treatment effect corresponding to a particular conditional treatment effect. Many simulation studies estimated average, or marginal treatment effects, using a conditional model to relate the outcome with the treatment and associated covariates, even though the estimated effects are not collapsible (i.e. marginal and conditional treatment effects will not coincide) (Austin, 2013; Gail et al., 1984; Greenland, 1987). For binary outcomes, even though odds ratios are not collapsible and for other reasons (Newcombe, 2006), we chose to adopt odds ratios due to its frequent usage in biomedical research.

Like the simulation of time-to-event outcomes in Chapter 4, we did not include censoring due to computational simplicity. Allowing the degree of censoring to be another factor in the design of the Monte Carlo simulations would increase the computational burden of the simulations substantially and increase the number of results that would require reporting. However, this may warrant future investigations.

To our knowledge, no previous research had studied the performance of entropy balancing in estimating treatment effects of different types of outcomes, using Monte Carlo simulations. Like any simulation, our simulation results might be limited to the scenarios considered by our simulation data. Therefore, the results cannot be generalized to settings that have not been evaluated.

Overall, we found the entropy balancing technique useful and excellent in performance. Entropy balancing merits more widespread adoption for estimating treatment effects of different types, when using observational data.

Chapter 8

A Comparative Study of the Different Strategies for Estimating Causal Treatment Effects

As evident from the previous chapters, quite a number of studies have substantially added to the repository of strategies for controlling confounding in the estimation of treatment effects in observational studies. However, applied practitioners need guidance to decide on the optimal strategy for any given scenario. To address this gap, we conducted series of Monte Carlo simulations evaluating both well-established methods, including the IPW, entropy balancing and the more recently proposed CBRMD and iCBRMD methods.

8.1 Background

So far, we have introduced and studied several strategies to estimate causal treatment effects in observational studies. Some of these methods are matching methods, while others are weighting methods. Though matching methods are loosely referred to as weighting methods, since they produce discrete weights. Continuous weights are obtained from entropy balancing and inverse probability weighting

(IPW) methods, which belong to the empirical calibration weighting and propensity score weighting methods, respectively.

With the increasingly high number of statistical methods to be applied on a broader spectrum of available observational data, there is a need to evaluate the performance of these methods under different real-world scenarios. Without loss of generality, we conducted series of Monte Carlo simulations to compare the performance of IPW, entropy balancing, PSM, CBRMD, and iCBRMD methods. We believe that these five methods are representative of the different methodological approaches of estimating causal treatment effects, as mentioned earlier.

8.2 Simulation Study

We used the same data generation scheme and considered the different scenarios of Section 3.2. We varied the values of $\alpha_{0,trt}$ in Equations (3.1) and (3.2), to ensure that the percentage of subjects who received the treatment (subsequently referred to as treatment prevalence) was fixed at 20%, 25%, 33% and 50%. For Scenario 1 (S1) and Scenario 2 (S2), we varied the sample sizes between n = 300 and 500, while the sample size was fixed at 1000 for Scenario 3 (S3).

Overall, each scenario was repeated 1000 times, all analysis methods were applied to the datasets, and the ATT estimates were obtained as the coefficient of the ATT-weighted linear regressions of Y on T. We reported the absolute bias and root mean square error (RMSE) of the estimated treatment effects, to evaluate the performance of the adjustment techniques.

8.3 Results

In this section, results obtained from analyzing the simulated datasets in each of the considered scenarios are presented. We present the results for each scenario under

separate subsections. As a form of sensitivity analysis, we ran simulations for other sample sizes ($n = 200, 1000$), however, we do not present the results as no qualitative differences were observed in the relative performance of the methods.

For the two PS methods, IPW performed better than PSM in most cases. Furthermore, PSM always performed poorly when the treatment prevalence was as high as 50%. Both CBRMD and iCBRMD methods had relatively higher biases across all the situations.

Scenario 1

This section presents the results of a situation where there are no noise covariates, i.e. a covariate is either related to the treatment or the outcome or both. Results are summarized in Figure 8.1.

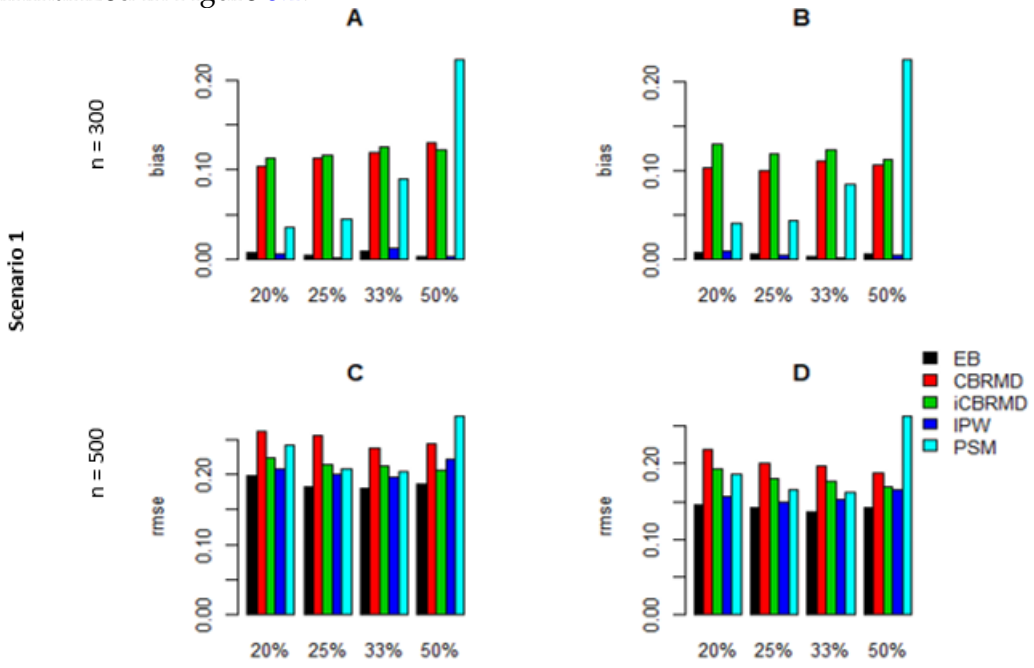

Figure 8.1: Absolute Bias (A and B) and RMSE (C and D) of the considered methods evaluated for Scenario 1

The RMSE estimates of CBRMD and iCBRMD methods decreased as the prevalence of treatment increased. IPW and EB methods produced indistinguishable estimates with minimal (near zero) bias and outperformed the other two techniques. For the

RMSE, EB outperformed IPW, followed by PSM (except when treatment prevalence equaled 50%), iCBRMD and CBRMD. Sample size did not change the pattern of results.

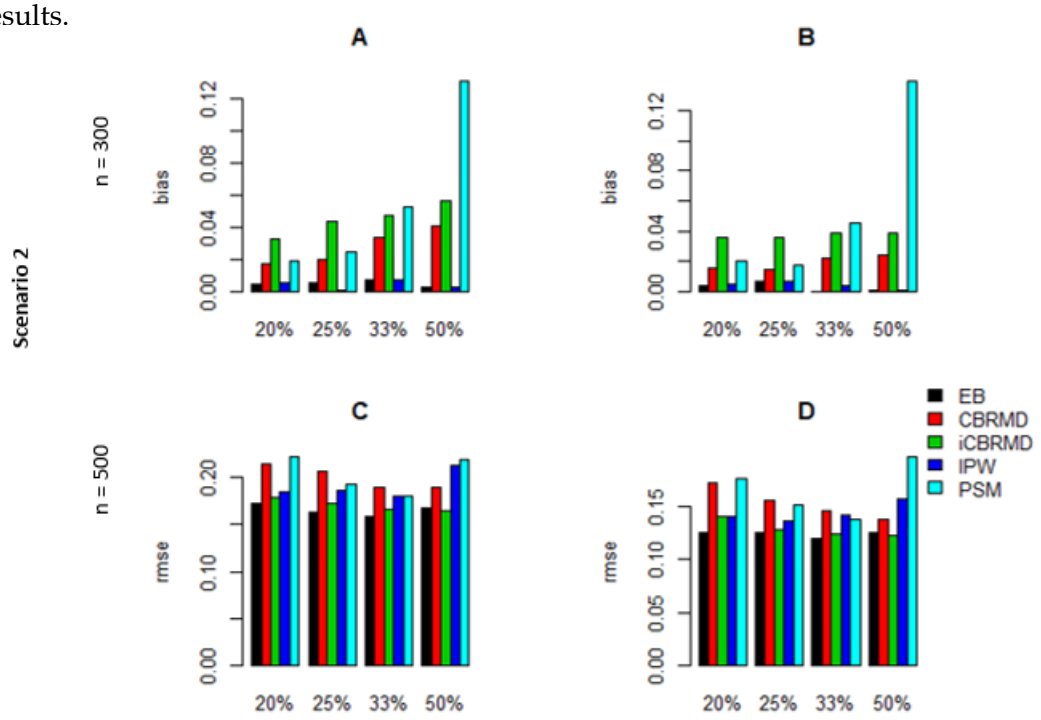

Figure 8.2: Absolute Bias (A and B) and RMSE (C and D) of the considered methods evaluated for Scenario 2

Scenario 2

In this section, we present the results of a situation that captures what is obtainable as a first resort in practice, where all covariates are included in the treatment assignment model and the outcome regression model in a linear fashion. Results are summarized in Figure 8.2. EB and IPW produced indistinguishable estimates, with both methods dominating the other two, in terms of bias. The RMSE estimates of CBRMD and iCBRMD methods decreased as the prevalence of treatment increased. For the RMSE, EB outperformed iCBRMD, followed by IPW and CBRMD. Increase in sample size, did not change the pattern of results.

Scenario 3

Here, we present the results of a situation where interactions of some covariates, as well as non-linearity in the outcome model, are introduced. Results are summarized in Figure 8.3. The results show that the purpose of introducing model complexities was not achieved, as IPW and PSM produced the highest RMSE estimates across the board. However, the performance in terms of bias worsened relative only to EB, while it still maintained smaller bias compared to the other two methods.

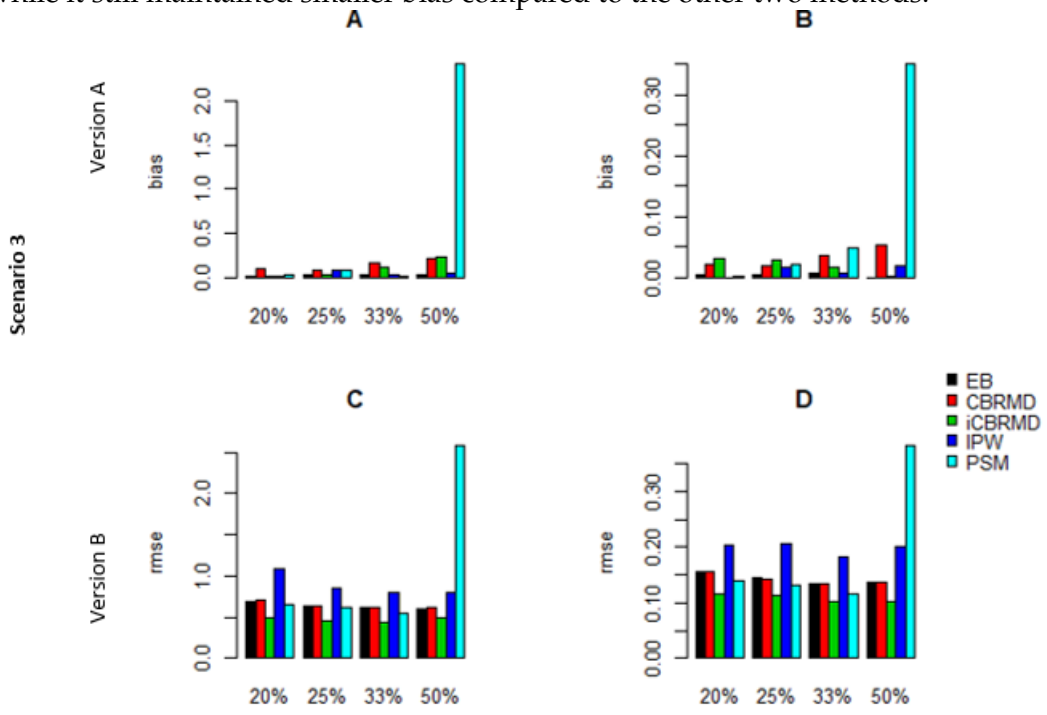

Figure 8.3: Absolute Bias (A and B) and RMSE (C and D) of the considered methods evaluated for Scenario 3

8.4 Discussion and Conclusion

We presented a simulation study of some strategies for estimating causal treatment effects, namely, entropy balancing, IPW, PSM, CBRMD, and iCBRMD. We evaluated the performance of these methods under different scenarios, based on the performance of treatment effect estimates in terms of bias and RMSE.

Relative to the other scenarios, IPW and PSM performed better in terms of RMSE,

for Scenario 1. This is expected, since the PS model is most likely to be correct, in Scenario 1. However, entropy balancing produced the lowest estimates, for Scenario 1. In Scenario 2, where the PS model is most likely to be false, IPW performed relatively weaker, with its accuracy in terms of RMSE, being better than only the CBRMD method. Entropy balancing produced the best results for this scenario. Results of Scenario 3 suggest that the inclusion of interactions and squared terms in both the treatment and outcome model, does not necessarily improve the performance of the PS methods (IPW and PSM), unless the right level of model complexity is chosen. Better estimation techniques for the treatment assignment, like the generalized boosted models and covariate-balancing propensity scores methods, are capable of circumventing the shortcomings of the logistic regression in estimating PS models (Imai et al., 2008; McCaffrey et al., 2004).

Though our findings suggest that no technique was overall superior to others, entropy balancing produced the best estimates in most cases. The reason why PSM always performed poorly when the treatment prevalence was as high as 50%, is that some treated group units may not be matched, as expected of pair matching methods. Pair-matching typically requires a large pool of potential control units - much larger than the number of treated units (Austin, 2014; Stuart, 2010).

Findings from our simulations are reliable and generalizable, because they were based on traditional study designs that mimic practice reality. Further, we introduced collinearity in a few pairs of covariates, as against some previous studies, which unrealistically assume independence.

Certain limitations to the current study require noting. First, only the essential, off-the-shelf versions of each of the weighting methods were utilized, since that is what most applied practitioners would likely do. Second, the number of evaluated techniques is not exhaustive, since they were limited to those that had been used tra-

ditionally, had been proposed in recent literature, or were promising. Thirdly, like any simulation, our simulation results might be limited to the scenarios considered by our simulation data. Therefore, the results cannot be generalized to settings that have not been evaluated. It would be interesting to expand the simulation scenarios and to accommodate other estimands in future studies.

In summary, this simulation study has laid out criteria for when one method for estimating causal treatment effects is expected to perform at its best or otherwise, relative to the other methods. While there is more work, we trust this simulation study will assist to move us one step closer towards best practices in comparative effectiveness research, for efficient estimation of causal treatment effects.

Chapter 9

Evaluating Treatment Effects from an HIV Study

In this chapter, we further demonstrate the effectiveness of the proposed techniques developed in this thesis, relative to the other considered methods, by applying them to data from an HIV study. Though we had examined these methods in the previous chapters using published datasets in the literature, we extend our examination to yet another dataset from a real-life study - a novel application of the considered techniques to a crucial public health issue in South Africa. While it is clear that we are not comparing the adjustment techniques, as we have already done so in the previous chapters, we aim to use this study to further illustrate the methods, as well as solidifying previous findings.

9.1 HIPSS Study

South Africa has the highest number of HIV infected individuals (over seven million) in the world, and the KwaZulu-Natal province is the worst hit, with a prevalence of 27.9% as at the end of 2015 (Kharsany et al., 2018). Accordingly, numerous public health initiatives to better control the HIV epidemic, have been implemented. There is thus a need for studies that monitor, evaluate and inform the programmatic interventions and policies over time. One such study is the HIV Incidence Provincial

Surveillance System (HIPSS). HIPSS provides timely, detailed and robust surveillance data to monitor HIV prevalence and incidence trends in Kwazulu-Natal, South Africa.

The main objective of the HIPSS study was to assess the impact of HIV-related prevention and treatment programmes on HIV prevalence, uptake of antiretroviral therapy (ART), CD4 cell counts and viral suppression, in a real-world non-experimental setting. The study also aimed to assess household sociodemographic and individual biological and behavioural characteristics, in association with HIV infection.

The HIPSS study design, source population and recruitment procedures, have been described previously (Kharsany et al., 2015, 2018). Briefly, HIPSS was a household-based study conducted in the Vulindlela and the Greater Edendale areas, in the uMgungundlovu District of KwaZulu-Natal, South Africa. The study had two cross-sectional surveys of 10,000 randomly selected individuals, aged 15-49 years, conducted one year apart. For each survey, a multi-stage cluster sampling method was used to choose enumeration areas, households and individuals. All completed questionnaires had peripheral blood samples collected and were allocated a unique identification number, with a linking number to link the household, respondents questionnaire and laboratory data. The University of Kwazulu-Natal (UKZN) Biomedical Research Ethics Committee (BF 269/13), the Associate Director of Science of the Center for Global Health (CGH) and in collaboration with the Provincial Department of Health (KwaZulu-Natal; HRKM 08/14), approved the HIPSS study protocol and informed consent.

9.2 Description of Data Collected from the HIPSS Study

We extracted data from the HIPSS study database for the first household survey comprising a total of 9812 men and women (15-49 years old), enrolled between June

2014 and June 2015. We were provided with the participants' information on the following:

- Demographic and behavioural variables, including gender, current age, marital status, socioeconomic status, and educational status.
- HIV status variables, including HIV testing history, HIV test results, ART use, as well as laboratory outcome variables, including CD4 and CD8 cell counts (cells/mm3) and HIV-1 plasma viral loads (copies/ml).
- Exposure to and treatment for other sexually transmitted infections (STI), namely; *Chlamydia trachomatis, Neisseria gonorrhoea, Mycoplasma genitalium, Trichomoniasis vaginalis*, herpes simplex virus type 2 (HSV-2) antibodies, syphilis, hepatitis B, and human papillomavirus (HPV) infection.
- Medical male circumcision (MMC) status (males only).

9.3 Data Analysis

Male circumcision, according to some studies (Liu et al., 2013; Price et al., 2010), reduces the bacterial load on the penis, as well as decreasing the relative abundance of these anaerobic genera, associated with HIV infection. The world health organization reports (World Health Organization, July 2018), from three randomized trials, provided compelling evidence that MMC reduces the risk of heterosexual HIV-1 acquisition in men by approximately 60%. Other studies (Prodger & Kaul, 2017; Tobian et al., 2014), also produced similar findings.

We next considered the causal effect of male medical circumcision (MMC) on HIV status. Motivated by the availability of the variable, HIV test outcome, in the data collected, we illustrated the use of the considered adjustment techniques in estimating the causal effect mentioned above. The ultimate goal was to estimate the average treatment effect among those who actually had MMC - the average treatment

effect among the treated (ATT), using entropy balancing (EB), inverse probability weighting (IPW), propensity score matching (PSM), covariate balancing rank-based Mahalanobis distance (CBRMD) and improved CBRMD (iCBRMD) techniques.

We attempted to approximate a randomized experiment by applying the considered adjustment techniques to provide evidence of a causal effect of MMC on HIV infection. Limited by the number of variables in the available data, we selected matching variables, based on their theoretical association with MMC and HIV infection. From the original 9812 participants, 3507 were retained for subsequent analysis, after restricting the sample only to males, and participants with complete cases on MMC and HIV status.

Overall, 1237 (35.3%) of participants did MMC, while 2270 (64.7%) did not. A total of 28.5% of the participants tested HIV positive. The covariates are summarized for participants who did and did not do MMC, in Table 9.1. Categorical variables were binary-coded. Absolute standardized mean differences (ASMD), were used to compare the balance in the measured covariates between those who did and did not do MMC. Seven of the twelve measured covariates had ASMDs that exceeded 0.1, which some authors consider as a threshold, indicative of negligible imbalance. The largest observed ASMDs, were for *current age* (0.487) and *HSV2* (0.340).

For EB, PSM and IPW, the propensity score was estimated from a regression of an indicator variable denoting MMC status on the twelve covariates (main terms only), described in Table 9.1, using a logit model. For PSM, we utilized the conventional nearest available pair-matching algorithm. While we superimposed a vertical line on Figure 9.1 as a threshold, we computed ASMDs for each of the twelve covariates in the sample that incorporated the weights induced by the five adjustment techniques. Results show that the five methods resulted in substantial reductions in imbalance, i.e. all ASMDs were less than 0.10. Of the five techniques, entropy

Table 9.1: Characteristics of participants by MMC status in the original sample

Label	Variable Description	MMC: Yes (N = 1237)	MMC: No (N = 2270)	ASMD
current_age	Participant's current age (years)	26.1 (8.6)	30.3 (9.5)	0.487
HSV2	Participant has HSV-2	445 (35.9%)	1186 (52.3%)	0.340
chlamydia	Participant has *chlamydia*	82 (6.6%)	105 (4.6%)	0.081
gonorrhoeae	Participant has *gonorrhoeae*	21 (1.7%)	48 (2.1%)	0.032
M_genitalium	Participant has *Mycoplasma genitalium*	47 (3.8%)	156 (6.9%)	0.167
T_vaginalis	Participant has *Trichomoniasis vaginalis*	31 (2.5%)	124 (5.5%)	0.198
hepatitisB	Participant has *hepatitis B*	34 (2.8%)	130 (5.7%)	0.182
syphilis	Participant has *syphilis*	27 (2.2%)	62 (2.7%)	0.038
degree	have a diploma/degree	112 (9.1%)	93 (4.1%)	0.173
high_school	completed only high school	548 (44.3%)	843 (37.1%)	0.144
married_before	widowed/divorced/separated	38 (3.1%)	96 (4.2%)	0.067
currently_married	married/living together with spouse	84 (6.8%)	156 (6.9%)	0.003

Note: Continuous variables are represented as mean (standard deviation), while dichotomous variables are represented as N (%).

balancing resulted in the best balance (all ASMDs were perfect zeros).

We used a binary logit model to regress the HIV test outcome (+ve or ve) on an indicator variable denoting MMC status. The estimated crude marginal odds ratio was 0.383 (95% CI: [0.321, 0.453]). The model incorporated the weights induced by the five adjustment techniques. We performed bootstrapping (2000 samples) to produce 95% confidence intervals (CIs), which has been shown to account for uncertainty in the matching procedure (Stuart, 2010). As shown in Figure 9.2, all the five adjusted estimates, including the crude estimate, suggest that MMC significantly reduces the odds of HIV acquisition, among those who ultimately had MMC. When entropy balancing was used, the estimated marginal odds ratio was 0.667 (95% CI: 0.556 - 0.796), IPW produced a marginal odds ratio estimate of 0.665 (95% CI: 0.556 - 0.803), while PSM produced a marginal odds ratio estimate of 0.676 (95% CI: 0.555 - 0.825).

Relative to the unadjusted estimate, the estimated CIs were substantially wider for CBRMD: odds ratio = 0.637 (95% CI: 0.423 - 0.989). This may indicate that the CBRMD weights are subject to greater instability in this setting. iCBRMD, with a

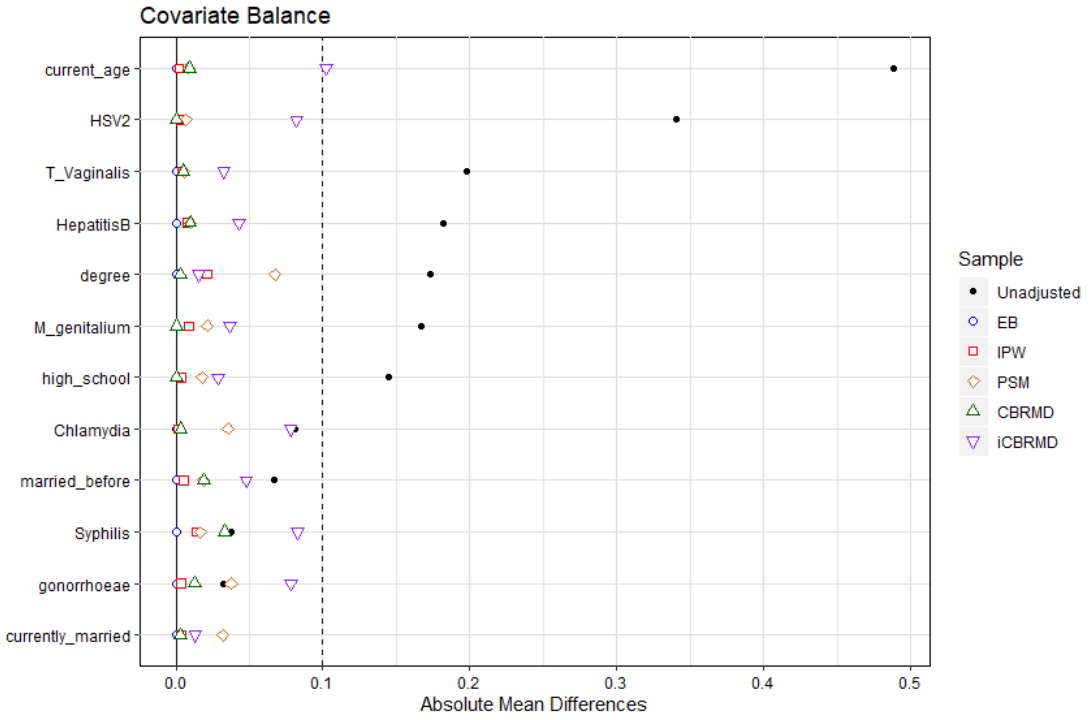

Figure 9.1: Covariate balance assessment

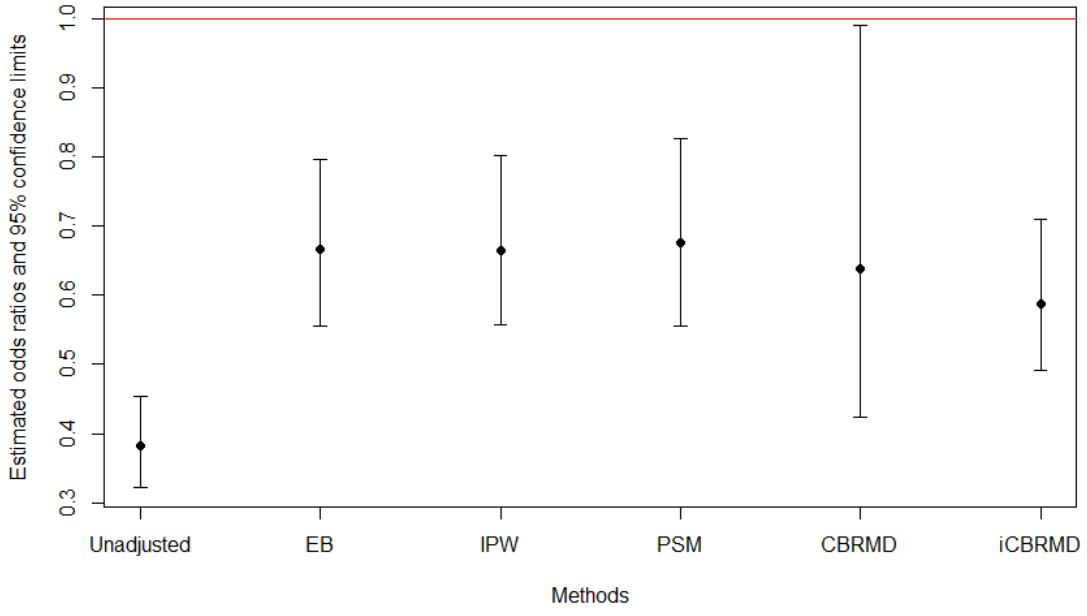

Figure 9.2: Estimated odds ratios of HIV test outcome and associated 95% CI

marginal odds ratio of 0.588, produced an estimated confidence interval with the smallest width (95% CI: 0.492 - 0.709), relative to the crude estimate.

Chapter 10

Discussion and Conclusion

The prime goal of this research work is to explore several topics relating to the strategies for inferring causal effects in observational studies. The study comprised both simulations and real-life applications. These ventures are mainly articulated in Chapters 3 - 9 of this thesis. They are precluded by Chapter 2, which provides a review of matching and weighting methods. This encompassed formal definitions and equations or formulas, for the various forms matching and weighting methods. We used many of the introduced concepts in the later chapters.

A major finding from this study, is that there is no overall best technique for estimating causal treatment effects, for observational data. Different techniques performed better in different scenarios. Though entropy balancing has excellent statistical properties and in many cases was found to outperform the other methods. Our proposed methods show promise and compete favourably with the other methods.

Our simulation findings were reliable and generalizable, as the simulation studies in each chapter were either based on traditional study designs that mimic practice reality, or based on notable existing real-life studies. In the empirical data examples, three famous datasets were used, namely, the Lalonde-PSID, Lindner, and RHC datasets. These datasets have been used in previous studies for evaluating methods

in the causal inference literature. Additionally, we utilized a dataset from a recently conducted study, which allows us the privilege of a novel application of the considered techniques to a crucial public health issue in South Africa. Each of these datasets has its unique properties in terms of size, degree of imbalance and the distribution of covariates.

Certain limitations require noting: This thesis focused only on the average treatment effects among the treated (ATT) and did not consider variants of aggregate causal effects. Our simulation results might be limited to the scenarios considered by our simulation data. Therefore, the simulation results obtained in each chapter cannot be generalized to settings that have not been evaluated. The number of assessed techniques was limited to those that had been used traditionally, had been proposed in the recent literature, or were, in our opinion, promising.

We acknowledge that we did not provide analytic proofs of our proposed methods, instead we utilized the law of large numbers to study the methods by Monte Carlo simulations. Furthermore, our simulations and empirical applications were based on homogeneous treatment effects. Our findings cannot be extended to heterogeneous treatment effects, where the goal of inference is a well defined average treatment effect, or average treatment effect on the treated. Finally, this thesis did not address unmeasured confounding, which is still a vexing problem in observational studies. We do acknowledge the profound influence that unmeasured confounding can have on estimators based on measured confounding.

The algorithms that execute the proposed methods, as well as the simulation studies, were developed using R statistical software (R Core Team, 2019). We have attached some of the R codes in the Appendix (10.1). The implementation of the proposed methods shall be incorporated into the main R library within a short period to facilitate its availability to any interested users. In conclusion, estimating causal effects

in observational studies is a much broader topic than what we have studied. We hope that this thesis has embroidered its concept and provided guidelines for moving forward. We aim to inspire new insights into issues involving causal inference in observational studies and provide a stimulus for further explorations in future research.

10.1 Recommendation for Future Studies

The current form of the newly proposed methods in this thesis, like any other methods, presents several opportunities for further improvements to enhance its general usage. However, whatever modifications or extensions intended at this stage, shall be addressed in future research works. For the iCBRMD method, one such task that comes to mind for the benefit of future studies, is the automation of selection of the optimal value of parameter λ that produces stable weights that maximizes covariate balance, while reducing bias and increase efficiency. Work is ongoing in that regard.

For future studies, evaluating the estimators in the presence of heterogeneous treatment effect should be considered. Heterogeneous treatment effects are usually of interest when assessing the efficacy of social programs and medical treatments. Treatment effect heterogeneity examines the degree to which different treatments have a differential causal impact on each unit. For instance, ascertaining subpopulations for which treatment is most beneficial (or harmful), is an essential goal of many clinical trials.

Amongst other things, the major contribution of this thesis was invested on aspects of the Mahalanobis distance, with the ultimate goal of circumventing its sensitivity to the various aspects of departure from the underlying assumption of Gaussianity. Though nonparametric-type approaches like the rank-based variant of the Mahalanobis distance utilized in this thesis, is a robust extension, more robust distance measures beyond the Mahalanobis may also be considered in future studies.

In sensu stricto, similarity measures can be conveniently thought of as inverses to dissimilarity measures. In fact, most kernels (similarity measures) are readily derived as inverses of distances (dissimilarity measures). With the popularity of kernels getting re-ignite since the rise to fame of Deep Neural Networks, it would be of interest to shift the focus from robust distances in causal inference to robust kernels. The compendium of kernels for several applications never ceases to increase in size, and it would be nice to revisit the topics explored in this thesis with some of the kernels, especially those that are robust to various assumptions.

Another emerging and very active research area is the targeted maximum likelihood estimation (Van Der Laan & Rubin, 2006). TMLE is a state-of-the-art technique for making causal inferences, which has gained tremendous popularity of late, especially in epidemiology and public health. TMLE is a semiparametric doubly-robust and locally efficient technique that improves the chances of correct model specification by allowing for flexible estimation using (nonparametric) machine-learning methods. It therefore requires weaker assumptions than its competitors. The statistical properties of TMLE make it a suitable tool for applied researchers aiming to estimate causal effects.

Observational studies tend to be replete with covariates that are highly correlated or noisy in the sense of being useless with respect to the response. A natural sequel to this thesis could concentrate on exploratory variable selection and feature learning concurrently with matching and causal inference. This would provide a powerful mechanism for policy making of the highest kind, if realized.

While the methods considered in this thesis can adjust for observed confounding, unobserved confounding is the Achilles heel of most observational studies. Future studies can extend our investigations to control for unobserved confounding. The

implementation of methods, such as sensitivity analysis (Pan & Bai, 2016; Rosenbaum & Rubin, 1983), can help increase confidence in results from observational studies.

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, *59*(5), 967–985.

Adhikary, S. D., Liu, W.-M., Memtsoudis, S. G., Davis III, C. M., & Liu, J. (2016). Body mass index more than 45 kg/m2 as a cutoff point is associated with dramatically increased postoperative complications in total knee arthroplasty and total hip arthroplasty. *The Journal of arthroplasty*, *31*(4), 749–753.

Ali, M. S., Groenwold, R. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C., de Boer, A., & Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of clinical epidemiology*, *68*(2), 122–131.

Amusa, L., Zewotir, T., & North, D. (2019a). Evaluation of subset matching methods: Evidence from a monte carlo simulation study. *American Journal of Applied Sciences*, *16*(3), 92–100.

Amusa, L., Zewotir, T., & North, D. (2019b). A weighted covariate balancing method for estimating causal effects in case-control studies. *Modern applied science*, *13*(4), 40–50.

Amusa, L. B., Zewotir, T., & North, D. (2019c). Examination of entropy balancing technique for estimating some standard measures of treatment effects: A simulation study. *Electronic Journal of Applied Statistical Analysis*, *12*(2), 491–35.

Amusa, L. B., Zewotir, T., & North, D. (2019d). A simulation study of some modern weighting methods for estimating treatment effects in observational studies. *Journal of Modern Applied Statistical Methods*, *In press*.

Aria, M., Capaldo, G., Iorio, C., Orefice, C. I., Riccardi, M., & Siciliano, R. (2018). Pls path modeling for causal detection of project management skills: a research field in national research council in italy. *Electronic Journal of Applied Statistical Analysis*, *11*(2), 516–545.

Austin, P. (2007). The performance of different propensity score methods for estimating marginal odd ratios. *Stat. Med.*, *26*, 3078–3094.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, *27*(12), 2037–2049.

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

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, *32*(16), 2837–2849.

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

Austin, P. C., Grootendorst, P., Normand, S. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study. *Statistics in medicine*, *26*(4), 754–768.

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N., & Stanbrook, M. B. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of clinical epidemiology*, *63*(2), 142–153.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, *34*(28), 3661–3679.

Austin, P. C., & Stuart, E. A. (2017). Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, *26*(6), 2505–2525.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, *24*(11), 1713–1723.

Brettschneider, C., Bleibler, F., Hiller, T. S., Konnopka, A., Breitbart, J., Margraf, J., Gensichen, J., Koenig, H. H., & Jena, P. S.-G. (2017). Excess costs of panic disorder with or without agoraphobia in germany - the application of entropy balancing to multiple imputed datasets. *Journal of Mental Health Policy and Economics*, *20*, S3–S3. URL <GotoISI>://WOS:000400956000007

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, *25*(24), 4279–4292.

Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(3), 673–700.

Ciavolino, E., & Carpita, M. (2015). The gme estimator for the regression model with a composite indicator as explanatory variable. *Quality & Quantity*, *49*(3), 955–965.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., & Califf, R. M. (1996). The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, *276*(11), 889–897.

Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: a clinically useful measure of treatment effect. *Bmj*, *310*(6977), 452–454.

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

d'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, *17*(19), 2265–2281.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407–424.

de los Angeles Resa, M., & Zubizarreta, J. R. (2016). Evaluation of subset matching methods and forms of covariate balance. *Statistics in medicine*, *35*(27), 4961–4979.

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

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376–382.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945.

Drabik, A., Büscher, G., Thomas, K., Graf, C., Müller, D., & Stock, S. (2012). Patients with type 2 diabetes benefit from primary care-based disease management: a propensity score matched survival time analysis. *Population health management*, *15*(4), 241–247.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Fullerton, B., Phlmann, B., Krohn, R., Adams, J. L., Gerlach, F. M., & Erler, A. (2016). The comparison of matching methods using different measures of balance: Benefits and risks exemplified within a study to evaluate the effects of german disease management programs on longterm outcomes of patients with type 2 diabetes. *Health services research*, *51*(5), 1960–1980.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, *71*(3), 431–444.

Golan, A. (2018). *Foundations of info-metrics: Modeling, inference, and imperfect information*. Oxford University Press.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology*, *125*(5), 761–768.

Grupp, H., Kaufmann, C., Knig, H.-H., Bleibler, F., Wild, B., Szecsenyi, J., Herzog, W., Schellberg, D., Schfert, R., & Konnopka, A. (2017). Excess costs from functional somatic syndromes in germanyan analysis using entropy balancing. *Journal of psychosomatic research*, *97*, 52–57.

Gu, X., & Rosenbaum, P. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*, 405–20.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46.

Hainmueller, J. (2014). *ebal: Entropy reweighting to create balanced samples*. R package version 0.1-6.
URL https://CRAN.R-project.org/package=ebal

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, *99*(467), 609–618.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234.

Harvey, R. A., Hayden, J. D., Kamble, P. S., Bouchard, J. R., & Huang, J. C. (2017). A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: a simulation and empirical example. *Pharmacoepidemiology and Drug Safety*, *26*(4), 368–377.
URL <GotoISI>://WOS:000398540200002

He, H., Wu, P., & Chen, D.-G. (2016). *Statistical causal inferences and their applications in public health research*. Springer.

Heckman, J. (1997). Instrumental variables. *Journal of human resources*, *32*(3), 441–462.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, *2*(3-4), 259–278.

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

Ho, D., Imai, K., King, G., & Stuart, E. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8).

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199–236.

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

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663–685.

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

Imai, K., King, G., & Stuart, E. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *Roy. Statist. Soc. A*, *171*, 481–502.

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

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jacovidis, J. N. (2017). *Evaluating the Performance of Propensity Score Matching Methods: A Simulation Study*. Thesis.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, *58*(4), 272–279.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, *22*(4), 523–539.

Kereiakes, D. J., Obenchain, R. L., Barber, B. L., Smith, A., McDonald, M., Broderick, T. M., Runyon, J. P., Shimshak, T. M., Schneider, J. F., & Hattemer, C. R. (2000). Abciximab provides cost-effective survival advantage in high-volume interventional practice. *American heart journal*, *140*(4), 603–610.

Kharsany, A. B., Cawood, C., Khanyile, D., Grobler, A., Mckinnon, L. R., Samsunder, N., Frohlich, J. A., Karim, Q. A., Puren, A., Welte, A., et al. (2015). Strengthening hiv surveillance in the antiretroviral therapy era: rationale and design of a longitudinal study to monitor hiv prevalence and incidence in the umgungundlovu district, kwazulu-natal, south africa. *BMC public health*, *15*(1), 1149.

Kharsany, A. B., Cawood, C., Khanyile, D., Lewis, L., Grobler, A., Puren, A., Govender, K., George, G., Beckett, S., Samsunder, N., et al. (2018). Community-based hiv prevalence in kwazulu-natal, south africa: results of a cross-sectional household survey. *The Lancet HIV*, *5*(8), e427–e437.

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript*, *15*.

Kullback, S. (1959). *Information theory and statistics*. Wiley, New York.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, (pp. 604–620).

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England journal of medicine*, *318*(26), 1728–1733.

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

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, *6*(3), e18174.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400.

Liu, C. M., Hungate, B. A., Tobian, A. A., Serwadda, D., Ravel, J., Lester, R., Kigozi, G., Aziz, M., Galiwango, R. M., Nalugoda, F., et al. (2013). Male circumcision

significantly reduces prevalence and load of genital anaerobic bacteria. *MBio*, *4*(2), e00076–13.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., Rochon, P. A., & Anderson, G. M. (2005). Reader's guide to critical appraisal of cohort studies: 2. assessing potential for confounding. *Bmj*, *330*(7497), 960–962.

Mattke, S., Han, D., Wilks, A., & Sloss, E. (2015). Medicare home visit program associated with fewer hospital and nursing home admissions, increased office visits. *Health Affairs*, *34*(12), 2138–2146.

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

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

Miksch, A., Laux, G., Ose, D., Joos, S., Campbell, S., Riens, B., & Szecsenyi, J. (2010). Is there a survival benefit within a german primary care-based disease management program? *The American journal of managed care*, *16*(1), 49–54.

Murphy, D. J., & Cluff, L. E. (1990). The support study. *Journal of Clinical Epidemiology*, *43*, V–X.

Newcombe, R. G. (2006). A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, *25*(24), 4235–4240.

Neyman, J. (1923). edited and translated by dorota m. dabrowska and terrence p. speed (1990). on the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, *5*(4), 465–472.

Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryan, T., Cleary, P., & McNeil, B. (2001). Validating recommendations for coronary angiography following

an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*, 387–398.

Pan, W., & Bai, H. (2016). A robustness index of propensity score estimation to uncontrolled confounders. In *Statistical Causal Inferences and Their Applications in Public Health Research*, (pp. 91–100). Springer.

Parish, W. J., Keyes, V., Beadles, C., & Kandilov, A. (2018). Using entropy balancing to strengthen an observational cohort study design: lessons learned from an evaluation of a complex multi-state federal demonstration. *Health Services and Outcomes Research Methodology*, *18*(1), 17–46.

Pearl, J., et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96–146.

Pearson, J. L., Stanton, C. A., Cha, S., Niaura, R. S., Luta, G., & Graham, A. L. (2014). E-cigarettes and smoking cessation: insights and cautions from a secondary analysis of data from a study of online treatment-seeking smokers. *Nicotine & Tobacco Research*, *17*(10), 1219–1227.

Price, L. B., Liu, C. M., Johnson, K. E., Aziz, M., Lau, M. K., Bowers, J., Ravel, J., Keim, P. S., Serwadda, D., Wawer, M. J., et al. (2010). The effects of circumcision on the penis microbiome. *PloS one*, *5*(1), e8422.

Prodger, J. L., & Kaul, R. (2017). The biology of how circumcision reduces hiv susceptibility: broader implications for the prevention field. *AIDS research and therapy*, *14*(1), 49.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Rosenbaum, P. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. (2002). *Observational Studies*. New York: Springer-Verlag., 2nd ed.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.*, *79*, 51624.

Rosenbaum, P., & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–8.

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., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*(2), 212–218.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, *2*(1), 1–26.

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, (pp. 293–298).

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3-4), 169–188.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios! *Evidence-based medicine*, *1*(6), 164.

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treatwhich of these should we use? *Value in health*, *5*(5), 431–436.

Scott, D. W. (2009). Sturges' rule. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(3), 303–306.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for r.

Setodji, C. M., McCaffrey, D. F., Burgette, L. F., Almirall, D., & Griffin, B. A. (2017). The right tool for the job: Choosing between covariate balancing and generalized boosted model propensity scores. *Epidemiology (Cambridge, Mass.)*, *28*(6), 802.

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., & Cook, E. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*, 546–555.

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

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, (pp. 465–472).

Stock, S., Drabik, A., Büscher, G., Graf, C., Ullrich, W., Gerber, A., Lauterbach, K. W., & Lüngen, M. (2010). German diabetes management programs improve quality of care and curb costs. *Health Affairs*, *29*(12), 2197–2205.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research & Evaluation*, *18*(13).

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Stürmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distributiona simulation study. *American journal of epidemiology*, *172*(7), 843–854.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, *97*(3), 661–682.

Tobian, A. A., Kacker, S., & Quinn, T. C. (2014). Male circumcision: a globally relevant but under-utilized method for the prevention of hiv and other sexually transmitted infections. *Annual review of medicine*, *65*, 293–306.

Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1).

Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012). Propensity scores: from naive enthusiasm to intuitive understanding. *Statistical methods in medical research*, *21*(3), 273–293.

Windt, R., & Glaeske, G. (2010). Effects of a german asthma disease management program using sickness fund claims data. *Journal of Asthma*, *47*(6), 674–679.

World Health Organization (July 2018). Voluntary medical male circumcision for HIV prevention. [Online; accessed 20-September-2019].
URL https://www.who.int/hiv/pub/malecircumcision/vmmc-progress-brief-2018/en/

Zagar, A. J., Kadziola, Z., Lipkovich, I., & Faries, D. E. (2017). Evaluating different strategies for estimating treatment effects in observational studies. *Journal of biopharmaceutical statistics*, *27*(3), 535–553.

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.

Appendix A: Published papers

Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v12n2p491

Examination of Entropy balancing technique for estimating some standard measures of treatment effects: a simulation study

By Amusa, Zewotir, North

Published: 14 October 2019

Electronic Journal of Applied Statistical Analysis
Vol. 12, Issue 02, October 2019, 491-35
DOI: 10.1285/i20705948v12n2p491

Examination of Entropy balancing technique for estimating some standard measures of treatment effects: a simulation study

Lateef Amusa*, Temesgen Zewotir, and Delia North

Department of Statistics, University of Kwazulu-Natal, Westville campus, South Africa

Published: 14 October 2019

In observational studies, propensity score weighting methods are regarded as the conventional standard for estimating the effects of treatments on outcomes. We consider entropy balancing, which despite its excellent conceptual properties, has been under-utilized in the applied studies. Using an extensive series of Monte Carlo simulations, we evaluated the performance of entropy balancing, in estimating difference in means, marginal odds ratios, rate ratios, and hazard ratios. The performance of entropy balancing was relatively compared with that of inverse probability of treatment weighting using the propensity score. We found that entropy balancing outperformed the IPW method in estimating difference in means, marginal odds ratios, and hazard ratios, but when estimating marginal rate ratios, IPW performed better. Entropy balancing produced more biased estimates in many cases. However, the entropy balancing algorithm is capable of controlling bias by loosening the tightening of the pre-specified tolerance on covariate balance. We report findings as to when one technique is better than the other with no proclamation on whether one method is in every case superior to the other. Entropy balancing merits more widespread adoption in applied studies.

keywords: Entropy balancing, Monte Carlo simulation, Observational studies, Propensity score weighting, Treatment effect, odds ratios, hazard ratios, rate ratios.

*Corresponding author: amusasuxes@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

The evaluation of a treatment or intervention is particularly straightforward in experiments but very complicated in observational studies where treatment assignment is not random. In observational studies, treatment selection is usually related to the background covariates and can confound estimated treatment effects.

Estimation of treatment effects in observational studies has conventionally been done using propensity score (PS) methods (Austin, 2014; Dehejia and Wahba, 2002; Guo et al., 2006; Guo and Fraser, 2010; Hirshberg and Zubizarreta, 2017). Among the PS methods, PS weighting (Hirano and Imbens, 2001) received more attention. In particular, the inverse probability of treatment weighting (IPW) is the most commonly used weighting method by applied researchers and practitioners, especially in the medical and health sciences (Austin and Stuart, 2015).

Much recently, entropy balancing – an optimization-based method, has gained the attention of applied researchers (Adhikary et al., 2016; Brettschneider et al., 2017; Grupp et al., 2017; Mattke et al., 2015; Pearson et al., 2014). Entropy balancing (Hainmueller, 2012) performs excellently in achieving covariate balance and efficient estimation of treatment effects. Additionally, entropy balancing (EB) is straightforward to implement. EB calibrates weights using the control group's distribution moments as constraints while optimizing the covariate balance apriori. Consequently, EB obviates the need for continually specifying the PS model until the desired covariate balance is achieved.

There is an increasing interest in using entropy balancing to estimate marginal or average treatment effects on outcomes of different types (Adhikary et al., 2016; Brettschneider et al., 2017; Grupp et al., 2017; Mattke et al., 2015; Parish et al., 2018; Pearson et al., 2014). Accordingly, we investigate the performance of entropy balancing in estimating treatment effects on continuous, binary, count, and time-to-event outcomes.

Using the IPW method as a benchmark, the current study used Monte Carlo simulations to examine the performance of entropy balancing in estimating some measures of treatment effects. We considered the estimation of difference in means, odds ratios, rate ratios, and hazard ratios for the continuous, binary, count and time-to-event outcomes, respectively. We also utilized the average treatment effect among the treated (ATT) as our estimand of interest.

This paper is structured as follows: In the next Section, we describe briefly the methodology of the entropy balancing, as well as the inverse probability of treatment weighting. In Section 3, we describe the Monte Carlo simulation scheme that were used to examine the performance of the two considered techniques. In particular, we report on bias, mean squared error (MSE), model-based standard errors, and 95% confidence interval coverage. The simulation results are presented in Section 4. Finally, in Section 5, we summarize our findings and gave final remarks.

2 Methods

We briefly describe the weighting methods that were included in the simulation study. We consider units or subjects indexed i $(i = 1, ..., n)$. We assume that there is a binary treatment variable, T_i, and the size of the treated and control group units, respectively, n_1, n_0, are known, while $\mathbf{X}_k$ denote a K-dimensional column vector of the observed background covariates.

2.1 Entropy balancing

Entropy balancing is a preprocessing method that can guarantee covariates balance, via a reweighting scheme that assigns a scalar weight to each sample unit such that the reweighted groups satisfy a set of balance constraints that are imposed on the sample moments of the covariate distributions (Hainmueller, 2012). The reweighting scheme belongs to the family of Maximum Entropy methods, which has roots in information theory and applied statistics (Kullback, 1959; Golan, 2018; Aria et al., 2018; Ciavolino and Carpita, 2015; Carpita and Ciavolino, 2017). The weights w_i are selected to minimize the relative entropy:

$$\min_{w_i} H(w) = \min_{w_i} \sum_{i|T=0} w_i \; log(w_i/q_i) \tag{1}$$

subject to the constraints:

$$\sum_{i|T=0} w_i \; c_{ri} \; (X_k) = m_r, for \; r = 1, \ldots, R \tag{2}$$

$$\sum_{i|T=0} w_i = 1 \tag{3}$$

$$w_i \geq 0, \forall i, \tag{4}$$

where $q_i = \frac{1}{n_0}$ is a vector of the base weights, and m_r describes a set of R balance constraints imposed on the covariate moments of the reweighted control group. m_r is the formulation containing the rth order moment of a given variable X_k from the treated group, while the moment functions are specified for the control group as $c_{ri} \; (X_k) = X_k^r$ or $c_{ri} \; (X_k) = (X_k - \mu_k)^r$, with mean μ_k. Equation (2) is the balance constraint specified in terms of the rth moment to be achieved on all covariates; (3) is the normalization constraint, while (4) is the non-negativity constraint.

The minimization problem described above is computed from an unconstrained dual problem and reduced to a system of non-linear equations with R Lagrange multipliers (Hainmueller, 2012) of the form:

$$\min_z L^d = log(q' \; e^{-C'z}) + m'z \tag{5}$$

where $\mathbf{z} = (\lambda_1, \ldots, \lambda_R)'$ is a vector $(\mathbf{z}^*)$ of Lagrange multipliers for the balance constraints, rewritten in matrix form as $\mathbf{CW} = \mathbf{M}$, with the $(R \times n_0)$ constraint matrix,

$\mathbf{C} = (c_1(\mathbf{X_k}), \ldots, c_R(\mathbf{X_k}))'$, and the vector of moments $\mathbf{m} = (m_1, \ldots, m_R)'$. The corresponding solution of (5) is:

$$W^* = \frac{Q \ . \ exp(-C'Z)}{Q' \ exp(-C'Z)} \tag{6}$$

An iterative Levenberg-Marquardt algorithm exploits the 2nd order information to solve the dual problem:

$$z^{new} = z^{old} - l \ \nabla_z^2 \ L^{d-1} \ \nabla_z \ L^d \tag{7}$$

Where l is the step length, ∇_z and ∇_z^2 is the gradient and Hessian, respectively. The optimal step length is selected for each iteration.

We utilized the ATT weights (Parish et al., 2018), which are defined for the entropy balancing as fixing treated units' weight at unity and reweighting the control group units using the algorithm described above.

2.2 Inverse Probability of Treatment Weighting

The propensity score, defined by $e(x) = P(T = 1|X), 0 < e < 1$, is the probability of a subject or unit receiving the treatment of interest given the observed baseline covariates (Rosenbaum, 1983). In IPW, each unit's weight equals the reciprocal of the probability of receiving the treatment that the unit received. We utilized the ATT weights (Austin and Small, 2014; Austin and Stuart, 2017), which are defined for the IPW as fixing the treated units' weight at unity, and the control units as $\frac{\hat{e}(x)}{1-\hat{e}(x)}$ (Imbens, 2004). We estimated $\hat{e}(x)$ by using a logistic regression model to regress treatment status on the covariates associated with the treatment.

3 Simulation study

We conducted a series of Monte Carlo simulations to examine the performance of entropy balancing in estimating treatment effects while using the IPW method as a benchmark. We considered continuous, binary, count and time-to-event outcomes. All simulations were done using the R statistical package (R Core Team, 2019).

3.1 Data-generating process

We used a data generation scheme derived from previous studies (Lee et al., 2010; Setoguchi et al., 2008). We randomly generated ten baseline covariates, where each of them $(X_1 - X_{10}) \sim N(0, 1)$. Some pair of covariates were induced with specified levels of dependence. $X_1, X_3, X_5, X_6, X_8, X_9$ were dichotomized. Figure 1 describes the simulation design in terms of the causal relationship of the variables.

As shown in Figure 1, the simulation study aligns with practice reality: X_1, X_2, X_3, X_4 are associated with both treatment and outcome, X_5, X_6, X_7 are predictors of the treatment variable only, while X_8, X_9, X_{10} are predictors of the outcome variable only. The

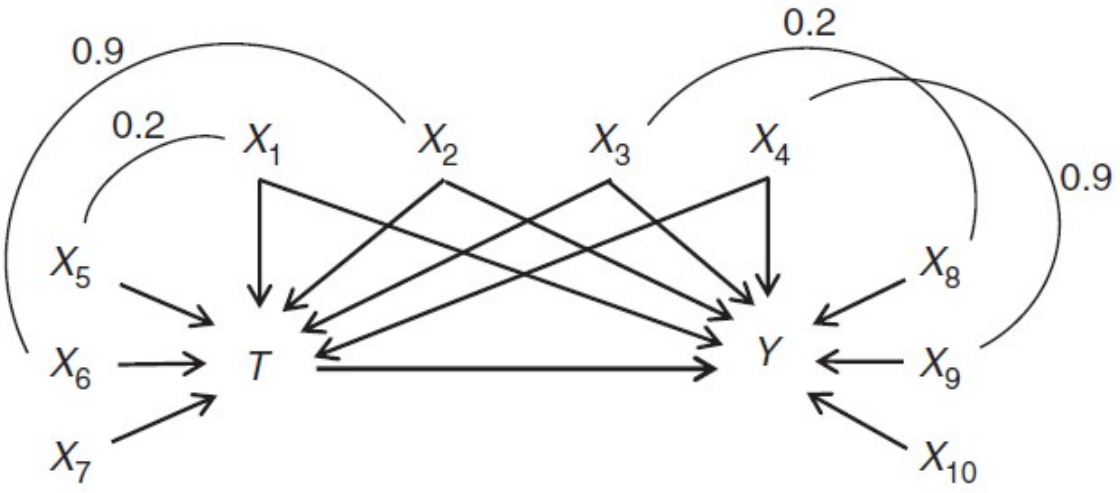

Figure 1: Data structure of the simulation study

treatment status was generated from a Bernoulli distribution: $T_i \sim Ber(p_{i,trt})$ where the probability of treatment selection $p_{i,trt}$ was determined from:

$$log\left(\frac{p_{i,trt}}{1-p_{i,trt}}\right) = \alpha_{0,trt} + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7 \quad (8)$$

The coefficients, $\alpha_1, \ldots, \alpha_7$ were based on real-life data utilized in a previous study (Setoguchi et al., 2008), while $\alpha_{0,trt}$ was selected so that the proportion of units who received the treatment (subsequently referred to as prevalence of treatment) was fixed at $\pi = 10\%$, 20%, 30%, 40%, and 50%. We developed the following iterative algorithm which was used to determine the value of $\alpha_{0,trt}$ that induced targeted prevalence π (Amusa et al., 2019a):

(i) We varied values of $\alpha_{0,trt}$ within reason (-3 to 3 in this case), and simulated n units.

(ii) For all the considered $\alpha_{0,trt}$ values, the corresponding individual $p_{i,trt}$ values were computed using (5), while the treatment variables $T_i \sim Ber(p_{i,trt})$ were generated, and the mean of each T_i correspond to π.

(iii) Based on the principle of the law of large numbers: the average of the results obtained from a large number of trials should be close to the expected value, Steps (i) and (ii) were repeated 1000 times to increase the precision of the estimation, and the value of $\alpha_{0,trt}$ which correspond to the desired π is chosen.

For each of the units, we generated an outcome Y_i conditional on T_i, and the seven covariates $(X_1, X_2, X_3, X_4, X_8, X_9, X_{10})$ associated with the outcome. Y_i was generated separately for continuous, binary, count, and time-to-event outcomes.

3.1.1 Continuous outcomes

While we fixed the true treatment effect at $\gamma = 1$, the continuous outcome was generated as

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i + \varepsilon_i \quad (9)$$

where $\varepsilon_i \sim N(0, \sigma^2)$

3.1.2 Binary outcomes

We generated a binary outcome as $Y_i \sim Bernoulli(p_i)$ using a logistic model:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i \quad (10)$$

3.1.3 Count outcomes

We generated a count outcome as $Y_i \sim Poisson(\eta_i)$ using a Poisson model (Amusa et al., 2019b):

$$log(\eta_i) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i \quad (11)$$

3.1.4 Time-to-event outcomes

For time-to-event outcomes, we used a data-generating process described by a previous study (Bender et al., 2005). Survival times t_i are generated as

$$t_i = \left(\frac{-log(U_i)}{\lambda e^{LP}}\right)^{\frac{1}{v}} \quad (12)$$

Where $U_i \sim Uniform(0, 1)$, and the linear predictor, $LP = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i$ This process generates survival times from a Cox-Weibull distribution. We assumed that all event times are observed for the current analyses.

3.2 Parameter values for data generation

The regression coefficients in the outcome data generation took the values: $\beta_1 = \beta_2 = \beta_3 = log(2), \beta_4 = \beta_5 = \beta_6 = log(1.75)$, and $\beta_7 = log(1.5)$ to reflect very high, high, and moderate effect sizes (Austin et al., 2007; Austin, 2014).

For continuous outcomes, the standard deviation values were fixed at $\sigma = 1$ and 0.5. The conditional treatment effect β_{trt} values were fixed at log (1.5) and log (0.5) for odds ratios, hazard ratios and rate ratios. The chosen values of β_{trt} = log (1.5) and log (0.5) were aimed at reflecting beneficial ($\beta_{trt} > 0$) and adverse ($\beta_{trt} < 0$) treatment effect, respectively. In generating dichotomous outcomes, the β_0 value was set to ensure that the prevalence of the event of interest occurred for approximately 70% of the units. We set v = 2 and $\lambda = 0.000001$ when generating time-to-event outcomes.

Finally, the above data-generating process has randomly generated treatment variable, covariates, and four different outcomes each of size n units, while inducing a conditional treatment effect.

A conditional treatment effect is the average effect, at the individual or unit level, of moving a unit from control to treated group. In contrast, a marginal effect is the average effect, at the population level, of moving the whole population from control to treated group (Greenland, 1987). Since the difference in means is collapsible, the conditional treatment effect coincides with the true marginal treatment effect. However, the other three treatment effects are not collapsible (Austin, 2013; Gail et al., 1984). Thus, for each of the conditional treatment effects (log-odds ratios, log-hazard ratios, and log-rate ratios), we determined their corresponding true marginal treatment effects. Details of this process of obtaining the true marginal treatment effect have been explained elsewhere (Austin, 2013, 2014; Austin and Stuart, 2017). The obtained true marginal treatment effect in the treated population from this process is regarded as the true ATT, for each of the considered effects.

3.3 Statistical analyses in simulated datasets

For a given treatment effect associated with each of the type of outcome considered, we randomly generated 1000 data sets of size 500 using the earlier described data-generating scheme. Using each of the simulated datasets, we separately estimated the different treatment effects, while utilizing each of the ATT weights of entropy balancing and IPW. The treatment effects, γ, were estimated from the following generalized linear model:

$$g(E(Y|T)) = \beta_0 + \gamma T, \tag{13}$$

Where g was considered as the canonical link function for the normal linear model, logistic model, Poisson model, and Cox survival model for estimating the difference in means, odds ratios, hazard ratios, and rate ratios, respectively. We adopted the robust sandwich estimator for estimating the standard errors (Austin and Stuart, 2015; Joffe et al., 2004). We utilized the R-package ebal (Hainmueller, 2014) for implementing entropy balancing.

Let γ_i denote the ith estimated treatment effect using a given method, whereas γ is the true ATT. We then determined the following: Bias $=\frac{1}{1000}\sum_{i=1}^{1000}(\gamma_i - \gamma)$, mean squared error (MSE) $= \frac{1}{1000}\sum_{i=1}^{1000}(\gamma_i - \gamma)^2$. We also examined precision by averaging the model-based standard errors (SE) over the 1000 simulated datasets. Finally, we examined 95% coverage - the proportion of times γ is enclosed in the 95% confidence interval of γ over the simulated datasets.

4 Results

We present the simulation results according to each of the type of estimated treatment effects explained in the earlier Section. We focus on the performance of entropy balancing method, using the IPW method as a threshold for evaluating the results. As a form of

sensitivity analysis, we ran simulations for other sample sizes (n = 300, 1000), but we do not present the results as no qualitative differences were observed in the relative performance of the methods. However, we present results of the two standard deviation values ($\sigma = 1, 0.5$) assumed while estimating difference in means, as well as the two different true ATT values, varied each for odds ratios, hazard ratios, and rates ratios estimation. Altering these parameter values also did not change the conclusions in all the scenarios, except for when rate ratios were estimated.

4.1 Continuous outcomes: Difference in means

Results are summarized in Figures 2 and 3. In terms of bias, Figure 2 shows that both methods produced estimates with very low (near zero) bias. However, EB produced slightly higher biases, except for when the prevalence rate was 10%. For the MSE, EB outperformed IPW across the board (Figure 2). Both methods yielded very similar SE estimates, with the values decreasing with increasing prevalence rates (Figure 3). Though EB produced superior CI coverages - near perfect in most cases, both methods achieved reasonably high 95% CI coverage (Figure 3).

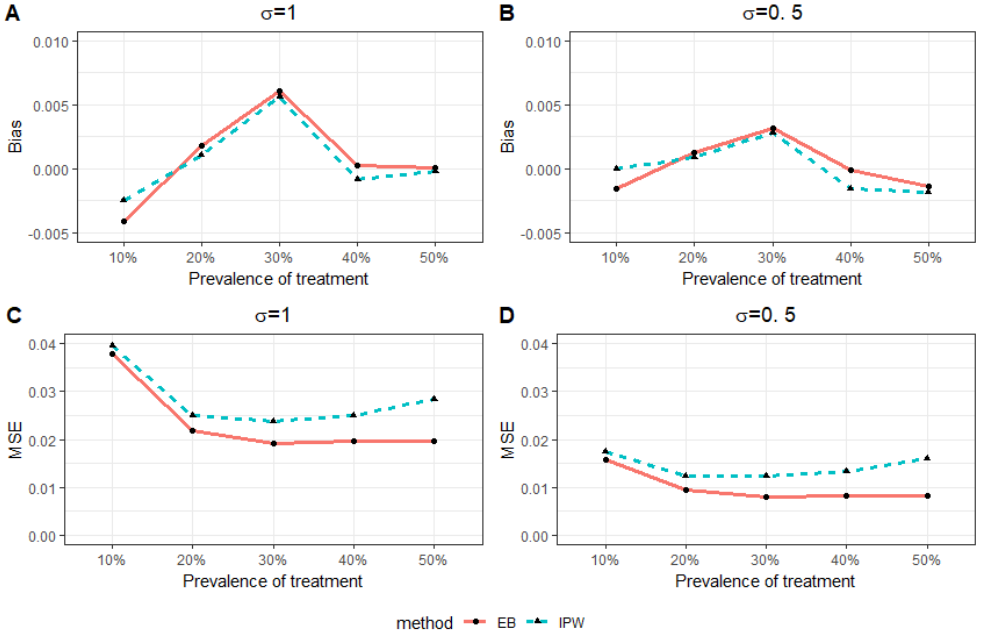

Figure 2: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating difference in means.

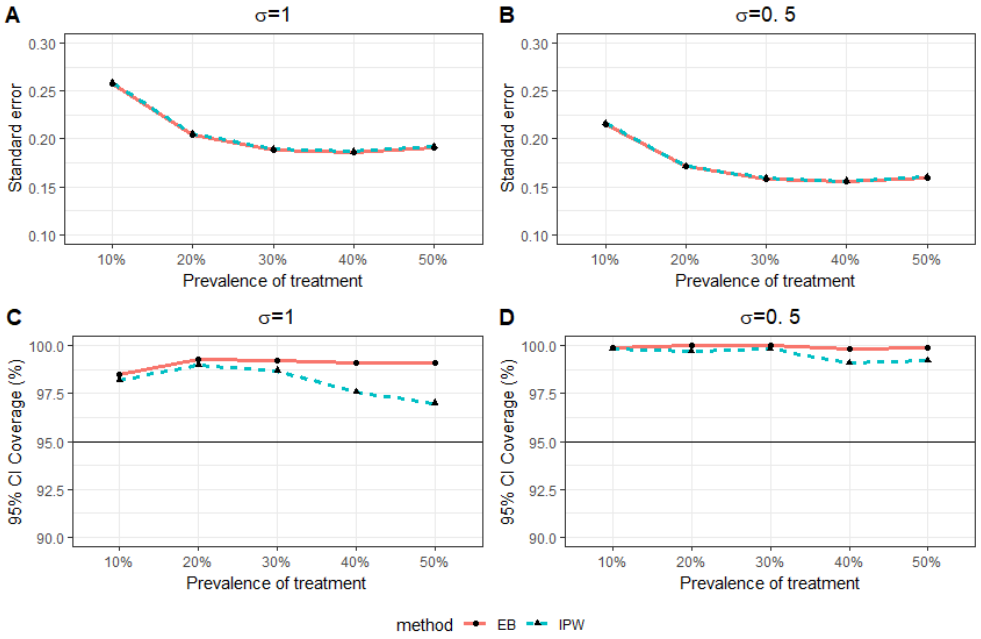

Figure 3: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating difference in means.

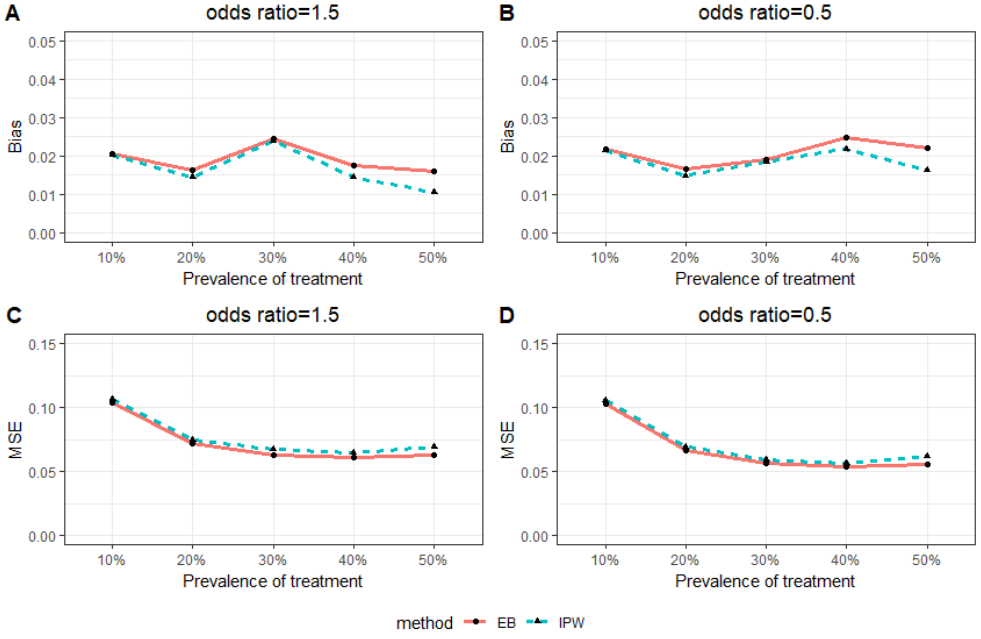

Figure 4: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating odds ratios.

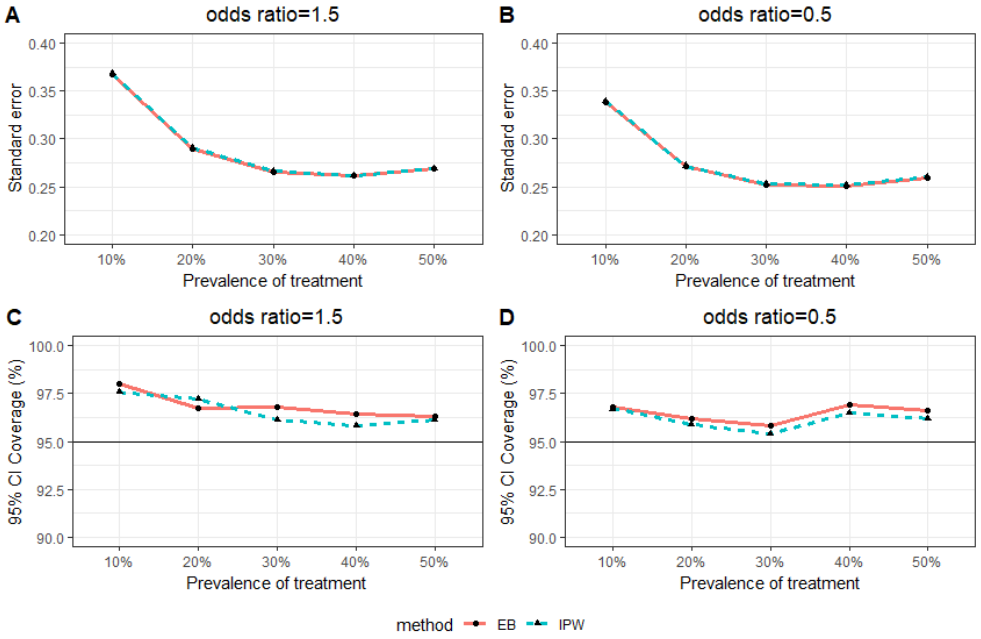

Figure 5: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating odds ratios.

4.2 Binary outcomes: Odds ratios

In terms of bias, Figure 4 shows that EB consistently produced higher biased estimates. For the MSE, EB outperformed IPW across the board, with the values decreasing with increasing prevalence rates (Figure 4). Both methods yielded very similar SE estimates, with the values decreasing with rising prevalence rates (Figure 5). Though EB produced superior CI coverages, both techniques achieved reasonably high 95% CI coverage (Figure 5).

4.3 Count outcomes: Rate ratios

Figure 6 shows that the MSE of both methods increased as the treatment prevalence increased from 10% to 40%. When the conditional rate ratio was positive ($\beta_{trt} > 0$), EB consistently produced estimates with higher bias, higher MSE, and lower 95% CI coverage. However, for ($\beta_{trt} < 0$), EB produced higher bias and MSE estimates only when the treatment prevalence was 20% or lower (Figure 6). The SE estimates were very similar between both methods, with the values decreasing with increasing prevalence rates (Figure 7). Both methods achieved reasonably high 95% CI coverage. Though EB had lower CI coverages when the conditional rate ratio was positive, it is not clear which of them produced higher coverage when the conditional rate ratio was negative (Figure 7).

4.4 Time-to-event outcomes: Hazard ratios

Figure 8 shows that the bias of both methods are not substantially different, except for higher prevalence rates (40% and 50%) where EB produced higher bias estimates. For the MSE, EB consistently outperformed IPW (Figure 8). As shown in Figure 9, the SE estimates were again very similar between both methods, with the values decreasing with increasing prevalence rates. Figure 9 illustrates that when the actual hazard ratio = 0.5, EB produced 95% CIs slightly below the nominal coverage rate at prevalence rates higher than 30%. However, EB provided superior CI coverages overall.

5 Discussion

Propensity score (PS) methods are the most widely used in estimating average treatment effects in observational studies. While the inverse probability of treatment weighting (IPW) method appears to be the most common implementation of PS methods, we introduce entropy balancing – a relatively new, but under-utilized weighting method, despite having nice conceptual properties. This study aims to use Monte Carlo simulations to evaluate the performance of entropy balancing, relative to the traditional IPW, in estimating some standard measures of treatment effect. While focusing on entropy balancing, we summarize our findings, and where necessary, place them in the context of existing literature.

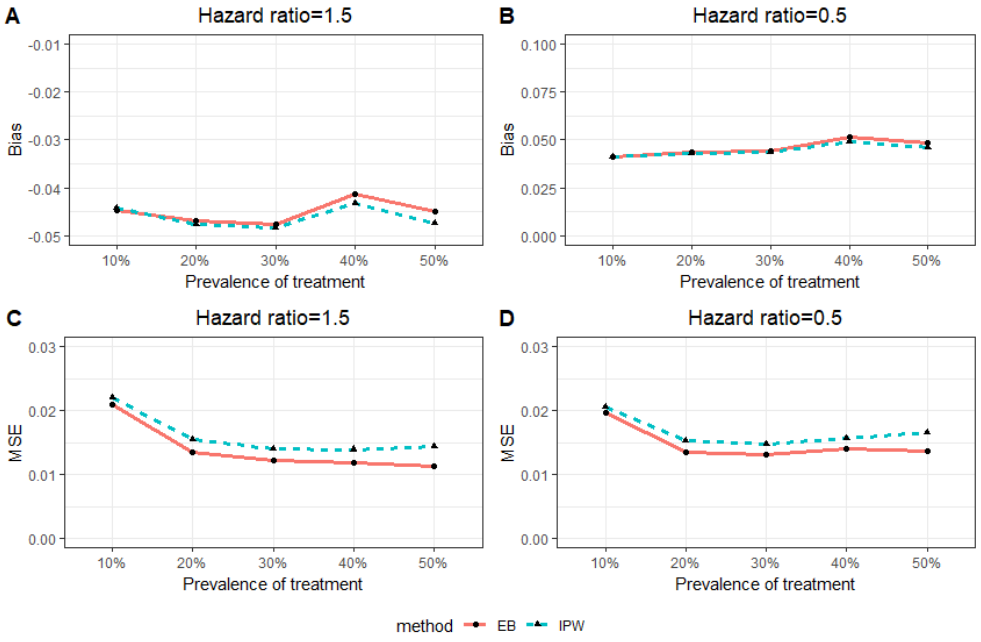

Figure 6: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating hazard ratios.

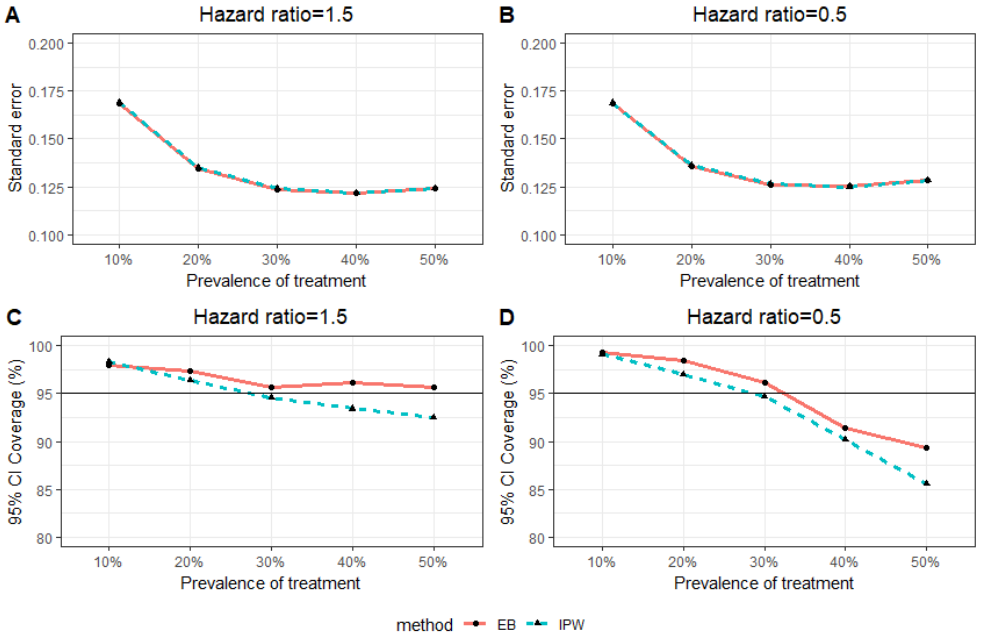

Figure 7: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating hazard ratios.

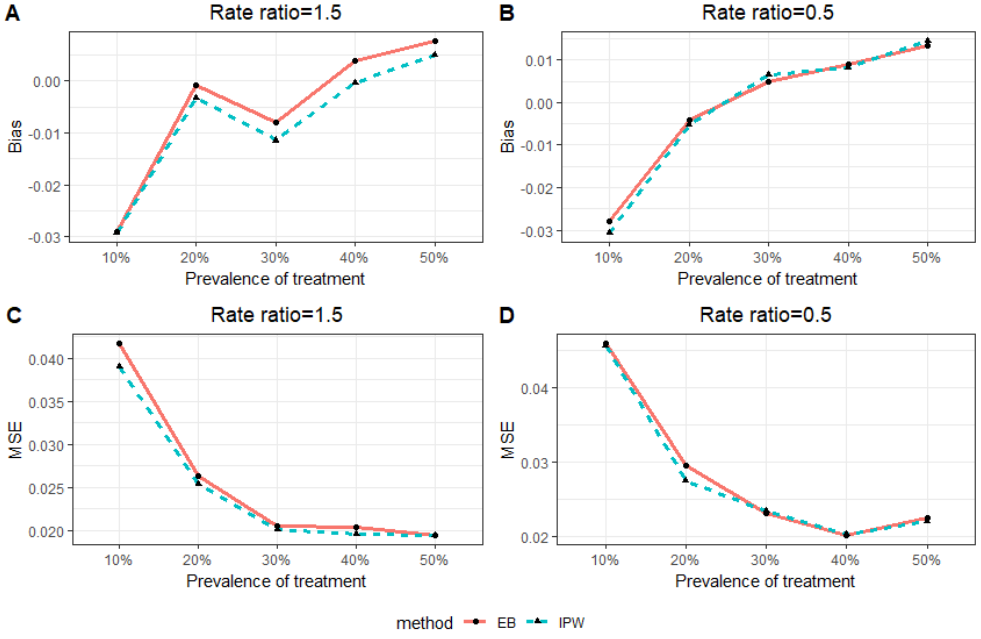

Figure 8: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating rate ratios.

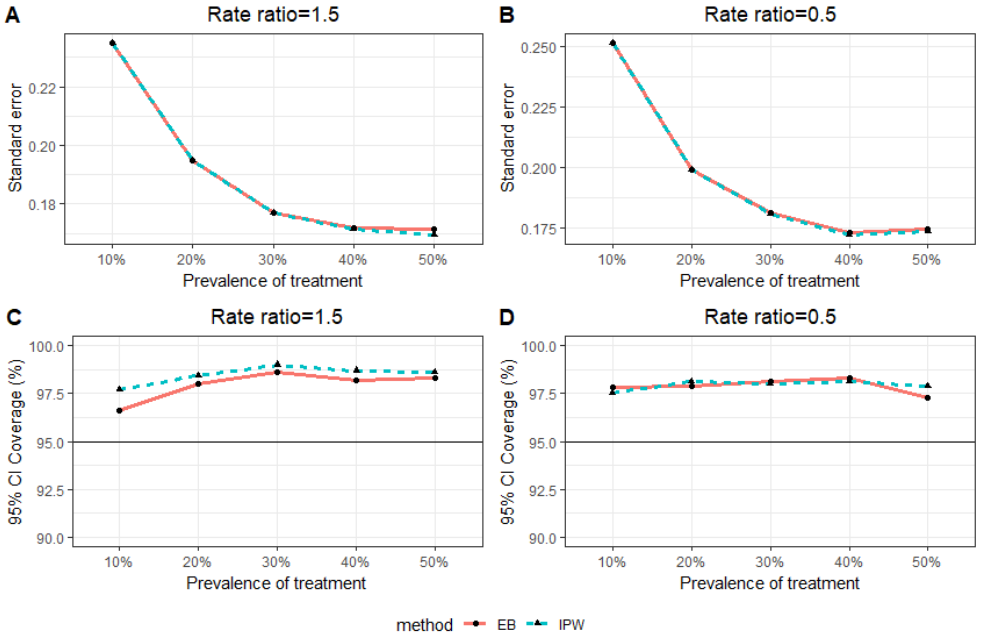

Figure 9: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating rate ratios.

Though both methods performed reasonably well in estimating the various treatment effects considered, we found on average that entropy balancing outperformed IPW for all the considered situations. However, a few exceptions were found: (i) When rate ratios were estimated, entropy balancing tended to produce estimates with slightly higher biases and mean squared errors. Although they considered conditional and not marginal treatment effects, a previous study by Austin (2007) found that conditioning on the propensity score did not substantially introduce bias into the estimation of rate ratios. (ii) The model-based standard errors for both methods were consistently indistinguishable. (iii) In terms of bias, across all the estimated treatment effects, entropy balancing consistently produced more biased estimates. Hence, there is an interesting bias-variance trade-off of the two techniques. However, entropy balancing has the facility to optimize the bias-variance trade-off by tightening the pre-specified tolerance on covariate balance (Harvey et al., 2017). Previous studies (Austin, 2007, 2013; Austin and Stuart, 2017) also support our findings in favour of IPW producing an unbiased estimation of odds ratios and hazard ratios.

A significant strength of this study is in the use of an algorithm which determines the true marginal treatment effect corresponding to a particular conditional treatment effect. Many simulation studies estimated average or marginal treatment effects using a conditional model to relate the outcome with the treatment and associated covariates, even though the estimated effects are not collapsible (i.e. marginal and conditional treatment effects will not coincide) (Austin, 2013; Gail et al., 1984; Greenland, 1987).

For binary outcomes, even though odds ratios are not collapsible and other reasons (Newcombe, 2006), we chose to adopt odds ratios due to its frequent usage in biomedical research.

The limitation of this study is that we did not include censoring in our simulation of time-to-event outcomes. The reason is due to computational simplicity. Allowing the degree of censoring to be another factor in the design of the Monte Carlo simulations would increase the computational burden of the simulations substantially and increase the number of results that would require reporting. However, this may warrant future investigations.

To our knowledge, no previous research had studied the performance of entropy balancing in estimating treatment effects of different types of outcomes, using Monte Carlo simulations. Like any simulation, our simulation results might be limited to the scenarios considered by our simulation data. Therefore, the results cannot be generalized to settings that have not been evaluated.

6 Conclusion

Overall, we found the entropy balancing technique useful and excellent in performance. Entropy balancing merits more widespread adoption for estimating treatment effects of different types when using observational data.

Availability of data and materials

The R codes for a complete reproduction of the results in this study are available from the corresponding author upon request.

Competing interests

The authors declare that they have no competing interests.

References

Adhikary, S. D., Liu, W.-M., Memtsoudis, S. G., Davis III, C. M., and Liu, J. (2016). Body mass index more than 45 kg/m2 as a cutoff point is associated with dramatically increased postoperative complications in total knee arthroplasty and total hip arthroplasty. *The Journal of arthroplasty*, 31(4):749–753.

Amusa, L., Zewotir, T., and North, D. (2019a). Evaluation of subset matching methods: Evidence from a monte carlo simulation study. *American Journal of Applied Sciences*, 16(3):92–100.

Amusa, L., Zewotir, T., and North, D. (2019b). A weighted covariate balancing method for estimating causal effects in case-control studies. *Modern applied science*, 13(4):40–50.

Aria, M., Capaldo, G., Iorio, C., Orefice, C. I., Riccardi, M., and Siciliano, R. (2018). Pls path modeling for causal detection of project management skills: a research field in national research council in italy. *Electronic Journal of Applied Statistical Analysis*, 11(2):516–545.

Austin, P. (2007). The performance of different propensity score methods for estimating marginal odd ratios. *Stat. Med.*, 26:3078–3094.

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, 32(16):2837–2849.

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

Austin, P. C., Grootendorst, P., Normand, S. T., and Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study. *Statistics in medicine*, 26(4):754–768.

Austin, P. C. and Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679.

Austin, P. C. and Stuart, E. A. (2017). Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, 26(6):2505–2525.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Brettschneider, C., Bleibler, F., Hiller, T. S., Konnopka, A., Breitbart, J., Margraf, J., Gensichen, J., Koenig, H. H., and Jena, P. S.-G. (2017). Excess costs of panic disorder with or without agoraphobia in germany - the application of entropy balancing to multiple imputed datasets. *Journal of Mental Health Policy and Economics*, 20:S3–S3.

Carpita, M. and Ciavolino, E. (2017). A generalized maximum entropy estimator to simple linear measurement error model with a composite indicator. *Advances in Data Analysis and Classification*, 11(1):139–158.

Ciavolino, E. and Carpita, M. (2015). The gme estimator for the regression model with a composite indicator as explanatory variable. *Quality & Quantity*, 49(3):955–965.

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

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3):431–444.

Golan, A. (2018). *Foundations of info-metrics: Modeling, inference, and imperfect information*. Oxford University Press.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology*, 125(5):761–768.

Grupp, H., Kaufmann, C., König, H.-H., Bleibler, F., Wild, B., Szecsenyi, J., Herzog, W., Schellberg, D., Schäfert, R., and Konnopka, A. (2017). Excess costs from functional somatic syndromes in germany—an analysis using entropy balancing. *Journal of psychosomatic research*, 97:52–57.

Guo, S., Barth, R., and Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28:357–83.

Guo, S. and Fraser, M. (2010). *Propensity score analysis; Statistical methods and applications*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hainmueller, J. (2014). *ebal: Entropy reweighting to create balanced samples*. R package version 0.1-6.

Harvey, R. A., Hayden, J. D., Kamble, P. S., Bouchard, J. R., and Huang, J. C. (2017). A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: a simulation and empirical example. *Pharmacoepidemiology and Drug Safety*, 26(4):368–377.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278.

Hirshberg, D. A. and Zubizarreta, J. R. (2017). On two approaches to weighting in causal inference. *Epidemiology*, 28(6):812–816.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279.

Kullback, S. (1959). *Information theory and statistics*. Wiley, New York.

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

Mattke, S., Han, D., Wilks, A., and Sloss, E. (2015). Medicare home visit program associated with fewer hospital and nursing home admissions, increased office visits. *Health Affairs*, 34(12):2138–2146.

Newcombe, R. G. (2006). A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, 25(24):4235–4240.

Parish, W. J., Keyes, V., Beadles, C., and Kandilov, A. (2018). Using entropy balancing to strengthen an observational cohort study design: lessons learned from an evaluation of a complex multi-state federal demonstration. *Health Services and Outcomes*

Research Methodology, 18(1):17–46.

Pearson, J. L., Stanton, C. A., Cha, S., Niaura, R. S., Luta, G., and Graham, A. L. (2014). E-cigarettes and smoking cessation: insights and cautions from a secondary analysis of data from a study of online treatment-seeking smokers. *Nicotine & Tobacco Research*, 17(10):1219–1227.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., and Cook, E. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, 17:546–555.

American Journal of Applied Sciences

Original Research Paper

Evaluation of Subset Matching Methods: Evidence from a Monte Carlo Simulation Study

Lateef Amusa, Temesgen Zewotir and Delia North

Department of Statistics, School of Mathematics,
Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Article history
Received: 20-01-2019
Revised: 13-02-2019
Accepted: 11-04-2019

Corresponding Author:
Lateef Amusa
Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa
Email: amusasuxes@gmail.com

Abstract: In the absence or infeasibility of experiments, matching methods have increasingly been used in making causal claims using observational data. This paper conducts a Monte Carlo simulation study, based on a household panel survey, to compare the performance of some widely used subset matching methods. The methods include the propensity score caliper matching, Mahalanobis distance matching, and coarsened exact matching. Comparisons were made in terms of the ability to reduce covariate imbalances, as well as effective recovery of the real treatment effect. Numerical results from our simulations provided evidence of coarsened exact matching outperforming the other methods. Our results also showed that, except for the Mahalanobis distance matching method, the efficiency of treatment effect estimates decreases with an increasing proportion of treated units.

Keywords: Matching, Balance, Monte Carlo Simulation, Observational Studies, Propensity Score

Introduction

Randomized experiments are the gold standard for estimating causal effects: They guarantee that the treated and control groups are only randomly different from one another with respect to the background covariates. Many matching methods have been proposed for replicating this scenario as much as possible for observed covariates with observational data.

Several methods serve as alternatives to matching, including adjusting for background variables in a regression model, instrumental variables, structural equation models and regression discontinuity designs. However, matching methods have been paid more attention and widely used because of its intuitiveness and more importantly, straightforward diagnostics, by which the performance is evaluated.

Matching is a nonparametric method for taking control of the confounding influence of background covariates or pretreatment control variables in observational or non-experimental data. The main aim of matching is to selectively prune observations from the data so that a better balance between the treated and control groups is achieved with the remaining data, which in other words means that the empirical distributions of the covariates in the two groups are then more similar. Statistical modelling assumptions handle any residual imbalance. The primary merit of matching is that it significantly reduces model dependence (King *et al*., 2011).

There are several matching methods existing in the literature, and they employ different distance measures, algorithms and rules for selecting control group members. Each technique could potentially choose different control group members from the overall control pool to create the matched group. The matched control group composition could, therefore, vary considerably depending on the particular matching algorithm used (Jacovidis, 2017). Matching techniques have been applied either using covariate (Miksch *et al*., 2010) or propensity score matching (Stock *et al*., 2010; Windt and Glaeske, 2010; Drabik *et al*., 2012) with some authors providing evidence for the superiority of propensity score matching (Drabik *et al*., 2012). The literature has shown that propensity score matching is not necessarily the gold standard (Fullerton *et al*., 2016). Depending on the scenario, other matching techniques can induce a better balance on the covariates and furthermore, the performance of propensity score matching highly depends on the correct specification of the propensity score model, choice of covariates and the matching algorithm used (Dehejia and Wahba, 2002; King *et al*.,

2011; Rosenbaum and Rubin, 1984). While many simulation studies have compared the performance of different matching methods, it cannot be taken for granted that their findings are transferrable to another data situation (Franklin *et al*., 2014). Consequently, there is a need for extensive research on identifying which matching methods perform best in several scenarios. Even though there have been a few notable studies that have examined the performance of matching techniques in terms of how well they balance the groups on the covariates, only a few of them have extended the evaluation of the matching techniques to the outcome analyses (Jacovidis, 2017; Austin, 2014; Stone and Tang, 2013).

Accordingly, this study aims to compare the performance of three (3) matching methods that are widely used in applied studies, under systematically manipulated conditions. The performance of each matching method was evaluated in terms of the ability to balance covariates between treated and control groups and efficient recovery of the real treatment effect. The abundance of subset matching methods and their variations is too large to be all compared in one study; without loss of generality, we studied the Propensity Score Caliper Matching (PSCM), Mahalanobis Distance Matching (MDM) and Coarsened Exact Matching (CEM).

Materials and Methods

Matching Methods

In this section, we briefly describe the matching methods we focused on in this study, each of which is commonly used in the applied literature. For the Mahalanobis distance and propensity score matching methods, we assumed a 1-1 matching without replacement, with the greedy matching algorithm being used to define the matched pairs. In matching without replacement, an already matched control unit is no longer available as a potential match for other treated units. In the case of greedy matching, a treated unit is chosen randomly, and the nearest control unit is then selected for matching to this treated unit (Austin, 2009).

Consider the unit i (i = 1,..., n), where T_i denotes a treatment variable coded 1 and 0 for the treated and control groups respectively. Let $\{Y_i(t): t \in (0,1)\}$ be the potential outcome variable value, also known as a counterfactual outcome (Rubin, 1974). This implies that $Y_i = T_iY_i(1) + (1\text{-}T_i)Y_i(0)$ is observed. Let X_i be a vector of pretreatment covariates; while, let m_T and m_C be the number of matched treated and control units respectively, for the methods. In estimating the average treatment effects, the Sample Average Treatment effect on the Treated units (SATT) was utilized. SATT = $\frac{1}{n_t}\sum_{i\in T} TE_i$, where $TE_i = Y_i(1)\text{-}Y_i(0)$.

Propensity Score Caliper Matching

Propensity score caliper matching is by far the most widely used matching method in the applied literature (Amusa, 2018). As the name of this method implies, it matches treated and control groups, based on the corresponding propensity scores, which weight covariates by how well they predict group membership. The propensity score was defined by Rosenbaum (1983) as the probability of treatment assignment, given the observed baseline covariates, stated mathematically as:

$$e_i = P\left(T_i = 1 \mid X_i\right), \tag{1}$$

where, it is assumed that, given the X's, the T_i's are independent:

$$P\left(T_1 = t_1 ... T_N = t_n\right) = \prod_{i=1}^{N} e_i^{t_i} \left\{1 - e_i\right\}^{1-t_i} \tag{2}$$

Let π_t and π_c and be the propensity scores for the treated and control group respectively, I_1 be the set of units in the treated group and I_0 be the set of units in the control group. A neighbourhood C (π_c) is defined to contain the c units control group ($c \in I_o$) as a match for the treated group t (t $\in$ I_1), where the absolute difference of propensity scores is the smallest among all possible pairs of propensity scores between t and c, i.e.:

$$C\left(\pi_t\right) = \min \| \pi_t - \pi_c \|, c \in I_o \tag{3}$$

Once a particular value for c is found to match t, c is removed from I_o, without replacement. There is a further restriction imposed on the distance between π_t and π_c, and as such, c is selected as a match for t, only if the absolute difference of propensity scores between the two groups meets the following condition:

$$\| \pi_t - \pi_c \| < \xi, k \in I_o, \tag{4}$$

where, ξ is a caliper or a pre-specified tolerance for matching.

This procedure is known as propensity score caliper matching. A caliper size of a quarter of the estimated propensity scores' standard deviation has been suggested in the literature (Rosenbaum and Rubin, 1985).

Mahalanobis Distance Matching

Similar to PSCM, the Mahalanobis distance matching method is built on specific notions of between observations of pretreatment covariates. MDM is unlike PSM which matches are made based on a scalar "Propensity Score", known as a balancing score; MDM matches on covariates by a specified distance,

which consequently ensures that covariates have equal weights. MDM measures the distance between two units, X_t and X_c as:

$$M\left(X_t, X_c\right) = \sqrt{\left(X_t - X_c\right)' S^{-1}\left(X_t - X_c\right)} \quad (5)$$

Where X_t, X_c denote the treated group and control group covariates respectively; S is the sample covariance matrix of X. Once the distance metric d is selected, a matching algorithm can then be applied. The procedure is known as the Mahalanobis distance matching.

Coarsened Exact Matching

The earlier mentioned methods are known as Equal Percent Bias Reduction (EBPR) methods, where improvements in the bound of balance for one covariate will affect each of the other covariates. To avoid this and other shortcomings of the EPBR methods, a new generalized class of matching methods known as Monotonic Imbalance Bounding (MIB), which has Coarsened Exact Matching (CEM) as a particular case, was introduced (Iacus *et al*., 2011; 2012). The strength of this method lies in the fact that, unlike other matching methods where balance is being continually checked until it is improved, CEM inverts the process and thus guarantees that the covariate imbalances between the matched treated and control groups will not be more than the user's pre-chosen level. MIB methods, therefore, improve bounds in the balance of one covariate in isolation as it will not affect the maximum imbalance of each of the other covariates (Iacus *et al*., 2012).

The essential thought of CEM is to coarsen each variable as reasonably as possible temporarily, through automated choices of coarsening using the Sturges rule (Scott, 2009), or any user-defined coarsening could be used. The automated approach was adopted for this study because of its ease and intuition. The exact matching algorithm is then applied to the coarsened data to determine the matches and to prune unmatched units. Finally, the coarsened data are left out, and the original values of the matched data are retained. In other words, after coarsening, the CEM algorithm creates a set of strata, say sЄS, each with same coarsened values of X. Units in strata containing at least one treated and one control unit are retained, while units in the remaining strata are then removed from this sample.

We denote by T_s and C_s, the treated and control units, respectively in stratum s; m_T^s as the number of matched units in T_s; m_C^s is the number of matched units in C_s. The number of matched units are, respectively, $m_T = U_{s\epsilon S} m_T^s$ and $m_C = U_{s\epsilon S} m_C^s$, for the treated and control units. Unmatched units receive zero weight, while to each matched unit i in stratum s, CEM assigns the weights:

$$W_i = \begin{cases} 1, & i \epsilon T^s \\ \dfrac{m_C m_T^s}{m_T m_C^s}, & i \epsilon C^S \end{cases} \quad (6)$$

Simulation Scheme

In this section, we describe the design of the Monte Carlo simulations which were used for data generation and to compare the performance of the considered matching methods. The performance was assessed using the following criteria: (a) Quality of matches: The ability to induce balance on measured background covariates; (b) Absolute bias of estimated treatment effects; (c) Root Mean Squared Error (RMSE) of estimated treatment effects.

The data-generating process and analyses were conducted with R packages, "MatchIt" (Ho *et al*., 2011) and "Matching" (Sekhon, 2011), in the environment of R version 3.4.1 (R Core Team, 2016).

We replicate previous simulation designs that had been used to evaluate matching methods (Iacus *et al*., 2012; Jacovidis, 2017; Austin, 2011), with slight modifications – the proportion of treated units where varied. Data were generated to mimic the Lalonde non-experimental data described in the next section.

Data Generation – Covariates Balance

Data were generated to mimic the structure and properties of the famous non-experimental Lalonde-PSID data. A small portion of the data is a U.S. job training program provided to participants for 12-18 months to help them find a job (Lalonde, 1986). The dataset comprises the original Lalonde's experimental treated units and non-experimental control units from the Panel Study of Income Dynamics (PSID), which includes 185 treated and 2490 control units. The choice of this dataset is driven by its importance in the evaluation literature since there has been considerable knowledge accumulated on evaluating non-experimental estimators, using this data.

The dataset comprises ten covariates: Four continuous covariates including age (age), years of education (education), real earnings in 1974 (re74) and 1975 (re75); as well as six binary covariates including marital status (married), black race (black), Hispanic race (Hispanic), lack of a high school diploma (nodegree) and indicator variables for unemployment in 1974 (u74) and 1975 (u75).

Using the idea of Austin (2011), we related the ten covariates with the probability of treatment selection via the following logistic regression model:

$$\begin{aligned} Logit\left(\pi_{i,t}\right) &= \alpha_{0,t} + \alpha_1 age + \alpha_2 education + \alpha_3 re74 \\ &+ \alpha_4 re75 + \alpha_5 married + \alpha_6 black + \alpha_7 hispanic \\ &\alpha_8 nodegree + \alpha_9 u74 + \alpha_{10} u75 \end{aligned} \quad (7)$$

The treatment group membership was regressed on the covariates for the study data and was used as coefficients (α_1, α_2,…, α_{10}) above. The intercept $\alpha_{0,t}$ was modified such that the proportion of treated units is varied at four different levels: 0.17, 0.20, 0.25, 0.33. For each unit *i*, in each of 1000 replications from this process, treatment status (denoted by *T*) was generated from a Bernoulli distribution with parameter $\pi_{i,t}$, i.e., $T_i \sim$ Ber ($\pi_{i,t}$), so that the number of pre-match treated and control units in the sample varies over replications.

Data Generation – Recovery of the True Treatment Effect

Next, outcome scores (*Y*) were generated as follows:

$$\begin{aligned} Y &= 1000T + \beta_1 age + \beta_2 education + \beta_3 re74 \\ &+ \beta_4 re75 + \beta_5 married + \beta_6 black + \beta_7 hispanic \\ &+ \beta_8 nodegree + \beta_9 u74 + \beta_{10} u75 + \varepsilon \end{aligned} \tag{8}$$

ATT was fixed at 1000 and $\varepsilon \sim N(0,10)$ as assumed by Iacus *et al*. (2012). Also, like Jacovidis (2017), the covariances between the covariates and outcome variable were obtained for the study data and were used to calculate the coefficients (β_1, β_1,…, β_{10}) above. A total of 1000 replications of each dataset were generated and matched with each method.

Performance Assessment

As stated in Section 1, the performance of the matching methods were evaluated relative to the unmatched data, under two criteria: (i) quality of matches and (ii) recovery of the true treatment effect. For each criterion, we varied the proportion of units who received the treatment (subsequently referred to as proportion of treated) at 17%, 20%, 25% and 33% levels, which corresponds to treatment-to-control ratios of 1:5, 1:4, 1:3 and 1:2, respectively.

Quality of Matches

In terms of the quality of matches, the methods were compared in terms of their ability to induce covariates balance between treated and control groups. This was achieved using the absolute standardized mean difference and percent bias reduction for all the covariates. The Absolute Standardized Mean Difference (ASMD), according to Rosenbaum and Rubin (1985), is defined as:

$$ASMD_k = \begin{cases} \left| \dfrac{\bar{x}_t - \bar{x}_c}{\sqrt{\dfrac{s^2t + s^2c}{2}}} \right| * 100\%, \text{ for continuous covariates} \\ \left| \dfrac{\hat{p}_t - \hat{p}_c}{\sqrt{\dfrac{\hat{p}_t(1-\hat{p}_c) + \hat{p}_c(1-\hat{p}_c)}{2}}} \right| * 100\%, \text{ for dichotonous covariates}, \end{cases} \tag{9}$$

where, $\bar{x}_t$ and $\bar{x}_c$ denote the sample mean of the treated and control units, respectively for the *kth* covariate; s^2t and s^2c denote the sample variance of the treated and control units, respectively for the *kth* covariate; $\hat{p}_t$ and $\hat{p}_c$ denote the mean of the *kth* dichotomous variable in the treated and control units, respectively.

It has been suggested that a standardized mean difference of at most 10% is quite sufficient at balancing a given covariate between the treatment groups (Austin, 2007; Normand *et al*., 2001).

Following the convention, the Percent Bias Reduction (*PBR*) for each covariate was also utilized. A threshold value of at least 80% is acceptable for judging the effectiveness of a matching method in reducing covariate imbalances (Cochran and Rubin, 1973; Pan and Bai, 2015). The percent bias reduction is defined as follows:

$$PBR_k = \frac{|B_{k,before\ matching}| - |B_{k,after\ matching}|}{|B_{k,before\ matching}|} * 100\%, \tag{10}$$

where, $B_{k,befor\ matching}$ and $B_{k,after\ matching}$ denote the mean difference in the *kth* covariate between the treated and control units, before and matching respectively.

For each of the ten covariates, the absolute standardized mean difference and percent bias reduction values were averaged across the 1000 simulated datasets.

Recovery of the True Treatment Effect

In each of the matched sets, we estimated the SATT estimators based on the difference in means between the observed outcome in the treated units and the control units. The performance of estimated treatment effects was assessed by its absolute bias, calculated as $|\bar{\hat{\gamma}} - \gamma|$ and root mean square error (RMSE), calculated as $\sqrt{\left(\bar{\hat{\gamma}} - \gamma\right)^2 + \text{var}\left(\hat{\gamma}\right)}$, where $\bar{\hat{\gamma}}$ is the mean of the 1000 estimated treatment effects.

Results

In this section, we present results from the simulation study. We compared the matching methods in terms of covariates balance and the performance of treatment effect estimates.

Covariates Balance

The covariates balance assessment was varied at 17%, 20%, 25% and 33% proportions of treated units, as shown respectively in Table 1 to 4. As confirmed by the balance metrics, the raw data which we simulated from, is highly imbalanced - all the covariates have high standardized mean difference values - more substantial than the recommended 10% threshold value (Austin, 2007; Normand *et al*., 2001; Stuart, 2010).

Table 1: Balance assessment of covariates for 33% proportion of treated units (treatment-control ratio of 1:2)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.56	0.06	0.45	0.01	89.84	18.91	97.81
Education	0.57	0.04	0.24	0.01	92.42	57.78	98.22
re74	2.01	0.07	1.40	0.06	96.57	30.63	97.17
re75	2.86	0.09	2.00	0.34	97.00	30.05	88.14
Black	0.86	0.05	0.55	0.00	94.61	36.49	100.00
Hispanic	0.14	0.04	0.03	0.00	73.62	76.90	100.00
Married	0.64	0.04	0.49	0.00	93.16	24.50	100.00
Nodegree	0.53	0.04	0.25	0.00	92.02	53.35	100.00
u74	0.66	0.03	0.61	0.00	95.07	8.22	100.00
u75	0.67	0.03	0.60	0.00	95.56	10.16	100.00

Note: The presented values are averages from each of the 1000 replications

Table 2: Balance assessment of covariates for 25% proportion of treated units (treatment-control ratio of 1:3)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.59	0.06	0.45	0.01	90.52	23.75	97.90
Education	0.58	0.04	0.22	0.01	92.36	61.95	98.09
re74	2.16	0.06	1.26	0.05	97.23	41.35	97.57
re75	3.17	0.08	1.91	0.36	97.62	39.56	88.49
Black	0.92	0.05	0.45	0.00	94.90	50.89	100.00
Hispanic	0.14	0.04	0.00	0.00	72.57	99.79	100.00
Married	0.73	0.04	0.46	0.00	94.35	37.44	100.00
Nodegree	0.57	0.04	0.26	0.00	92.82	53.79	100.00
u74	0.73	0.03	0.59	0.00	95.89	19.51	100.00
u75	0.71	0.03	0.55	0.00	95.37	21.83	100.00

Note: The presented values are averages from each of the 1000 replications

Table 3: Balance assessment of covariates for 20% proportion of treated units (treatment-control ratio of 1:4)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.63	0.06	0.48	0.01	91.01	24.85	97.96
Education	0.59	0.05	0.22	0.01	91.57	63.41	97.89
re74	2.29	0.06	1.16	0.05	97.19	49.53	97.71
re75	3.43	0.08	1.84	0.40	97.55	46.15	88.30
Black	0.98	0.05	0.48	0.00	94.84	50.69	100.00
Hispanic	0.14	0.04	0.00	0.00	71.86	100.00	100.00
Married	0.81	0.04	0.47	0.00	95.03	42.18	100.00
Nodegree	0.61	0.04	0.27	0.00	92.58	56.00	100.00
u74	0.79	0.04	0.55	0.00	95.49	30.67	100.00
u75	0.74	0.04	0.51	0.00	94.92	30.20	100.00

Note: The presented values are averages from each of the 1000 replications

Table 4: Balance assessment of covariates for 17% proportion of treated units (treatment-control ratio of 1:5)

	ASMD				PBR (%)		
Covariates	RAW	PSCM	MAH	CEM	PSCM	MAH	CEM
Age	0.69	0.06	0.55	0.01	91.51	20.13	98.13
Education	0.61	0.05	0.21	0.01	91.58	65.77	97.67
re74	2.46	0.07	1.01	0.06	97.08	58.82	97.51
re75	3.75	0.10	1.73	0.44	97.45	53.88	88.23
Black	1.06	0.05	0.53	0.00	95.04	50.36	100.00
Hispanic	0.14	0.04	0.00	0.00	66.70	100.00	100.00
Married	0.91	0.04	0.53	0.00	95.25	42.28	100.00
Nodegree	0.65	0.05	0.26	0.00	92.76	60.16	100.00
u74	0.85	0.04	0.46	0.00	95.44	45.43	100.00
u75	0.77	0.04	0.44	0.00	94.77	43.08	100.00

Note: The presented values are averages from each of the 1000 replications

In terms of the absolute standardized mean difference, except for the Mahalanobis distance matching method - which resulted in values extremely above the 10% threshold for almost all covariates, propensity score and coarsened exact matching methods had qualitatively comparable balance in the measured covariates. This pattern was consistent across the proportion of treated units. It is however worthy of note that coarsened exact matching had absolute standardized mean difference values of zero for the six continuous covariates across all treatment-control ratios considered.

In terms of the PBR, the performance of coarsened exact matching was excellent - all ten covariates had the Cochran and Rubin's acceptable threshold value of at least 80% PBR value. Mahalanobis distance matching had the worst performance. Propensity score caliper matching also had close to such an excellent performance, barring one covariate which consistently had PBR values below the 80% threshold. CEM further consistently had quantitatively higher PBR values. This pattern was consistent across the considered proportion of treated units. Overall, coarsened exact matching performed best in balancing covariates between the treated and control groups.

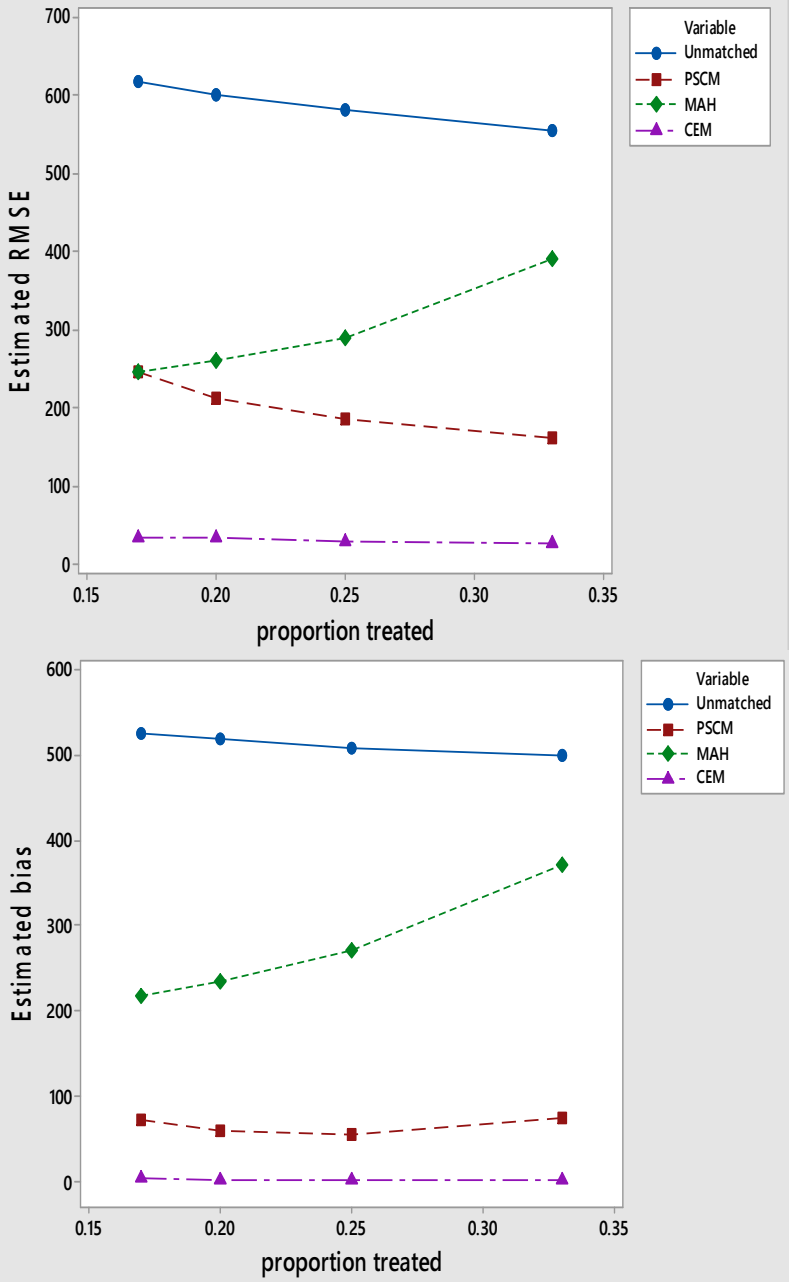

Fig. 1: Top panel: Root mean square error of estimated treatment effects; Bottom panel: Absolute bias of estimated treatment effects

Table 5: Absolute Bias and Root mean square of the matching methods relative to the unmatched data

	Proportion of treated: 17%		Proportion of treated: 20%		Proportion of treated: 25%		Proportion of treated: 33%	
Method	AB	RMSE	AB	RMSE	AB	RMSE	AB	RMSE
Unmatched	524.49	618.07	518.24	600.94	507.63	581.37	498.95	555.44
PSCM	72.60	246.62	59.46	211.75	55.33	186.13	75.98	161.54
MAH	218.59	246.14	235.35	260.59	271.01	290.81	371.49	392.15
CEM	4.33	35.19	2.30	33.16	2.22	29.28	1.80	27.91

AB means absolute bias
Note: values were averaged over 1000 Monte Carlo replications

Performance of Treatment Effect Estimates

The absolute bias (AB) and root mean square error (RMSE) of mean difference in outcomes between treated and control units of the matched data, across the considered proportion of treated units, are reported in Table 5 and Fig. 1.

Relative to the unmatched data, all the three matching methods had lower absolute bias and RMSE values. Regardless of the proportion of treated units, coarsened exact matching (CEM) produced the least absolute bias and RMSE values - absolute bias ranged from 1.80 to 4.33; RMSE ranged from 27.91 to 35.19. Also, the absolute bias and RMSE values of CEM reduced as the proportion of treated units increased from 17% to 33%. The same pattern was, however, not observed for the other two methods, while only the RMSE values for propensity score caliper matching (PSCM) followed the same pattern.

Unlike the other methods, Mahalanobis Distance Matching (MDM) produced absolute bias and RMSE values, which increased as the proportion of treated units increased from 17% to 33%.

Discussion

In this study, we presented a Monte Carlo simulation study of three subset matching methods, namely; propensity score caliper matching, Mahalanobis distance matching and coarsened exact matching. We evaluated the performance of these methods based on the ability to induce balance on measured background covariates, as well as the performance of treatment effect estimates via the assessment of their absolute biases and root mean square errors.

This study revealed that coarsened exact matching is the most effective in balancing covariates. As effective as CEM appears to be, the choice of coarsening can make or mar its performance: If the elements of the coarsening values are too small, then too many observations may be discarded. It may then lead to inefficient solutions in the analysis stage: if they are set too high, more observations will be retained, but more covariate imbalances, model dependence and statistical bias, will be introduced (Iacus *et al*., 2012). It is fine if there is a constant treatment effect (discarding units will not change the estimand of interest) but discarding units in the case of heterogeneous treatment effects may dramatically shift the estimand being estimated.

In assessing the recovery of the true treatment effect, Mahalanobis distance matching was the most biased. Mahalanobis distance matching also resulted in the highest RMSE across all considered proportions of treated units. Overall, coarsened exact matching had the least absolute bias and RMSE across all considered proportions of treated units.

Matching based on propensity score methods is by far the most widely used in applied studies to date. Previous research findings reveal that propensity score caliper matching was the best PSM technique (Bai, 2011). However, it is worthy of note that when the sample size is small or violates the statistical assumptions, caliper matching will possibly become problematic, because it usually ignores the cases when they do not have matched pairs or do not meet the caliper's criterion. Thus, it requires larger sample sizes to be very effective. Also worthy of note about matching on propensity scores is the correct specification of the propensity score model. In practice, an excellent alternative to 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 (Lee *et al*., 2010).

A significant strength of this study is the utilization of a real data set that has been used to evaluate the performance of matching methods and to provide a suitable structure for simulating the 1,000 data sets. It has the advantage of simplifying data generation procedures and avoiding making arbitrary choices. This study has a few limitations: Firstly, we have not exhausted all possible matching methods that have been described in the literature. Secondly, we assumed a one-to-one pair matching and therefore did not consider the many-to-one or many-to-many matching methods. Thirdly, we only assumed matching without replacement. Lastly, Optimal matching (Rosenbaum, 1989) - another alternative to the utilized greedy, nearest neighbour matching method, was not considered in this study. The results of our simulation study are limited to scenarios represented by the simulated data, which are typical in the applied social sciences. Parameters of the data generation model were based on model coefficients of a widely used panel study of income dynamics survey.

Conclusion

In comparison to the other subset matching methods, the utilized simulation study has provided sufficient evidence for the outperformance of coarsened exact matching method to the other considered methods, in terms of balancing covariates and efficiency in estimation of treatment effects. Future studies should include more matching methods; simulations should be expanded to consider a broader range of settings, including a non-linear model and heterogeneous treatment effects.

Acknowledgment

We appreciate the anonymous reviewers for their valuable comments which improved this manuscript. The corresponding author would also like to appreciate the University of Ilorin, Nigeria, for granting him study leave to pursue his Ph.D. studies in South Africa.

Author's Contributions

Lateef Amusa designed the study, wrote the simulation codes, and analyzed the data. Temesgen Zewotir and Delia North critically reviewed the manuscript and gave constructive which improved the manuscript.

Ethics

This article is original and contains unpublished material. All authors declare and attest to no conflicts of interest in relation to this study.

References

Amusa, L.B., 2018. Reducing bias in observational studies: An empirical comparison of propensity score matching methods. Turkiye Klinikleri J. Biostatist., 10: 13-26. DOI: 10.5336/biostatic.2017-58633

Austin, P.C., 2007. The performance of different propensity score methods for estimating marginal odd ratios. Stat. Med., 26: 3078-94. DOI: 10.1002/sim.2781

Austin, P.C., 2009. Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. Biometr. J., 51: 171-84. DOI: 10.1002/bimj.200810488

Austin, P.C., 2011. Optimal caliper widths for propensity□score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Stat., 10: 150-61. DOI: 10.1002/pst.433

Austin, P.C., 2014. A comparison of 12 algorithms for matching on the propensity score. Stat. Med., 33: 1057-69. DOI: 10.1002/sim.6004

Bai, H., 2011. A comparison of propensity score matching methods for reducing selection bias. Int. J. Res. Meth. Educ., 34: 81-107. DOI: 10.1080/1743727X.2011.552338

Cochran, W.G. and D.B. Rubin, 1973. Controlling bias in observational studies: A review. Sankhya Serial A, 35: 417-46. DOI: 10.1017/CBO9780511810725.005

Dehejia, R.H. and S. Wahba, 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econom. Stat., 84: 151-161. DOI: 10.1162/003465302317331982

Drabik, A., G. Büscher, K. Thomas, C. Graf and D. Müller *et al*., 2012. Patients with type 2 diabetes benefit from primary care-based disease management: A propensity score matched survival time analysis. Populat. Health Manage., 15: 241-247. DOI: 10.1089/pop.2011.0063

Franklin, J.M., J.A. Rassen, D. Ackermann, D.B. Bartels and S. Schneeweiss, 2014. Metrics for covariate balance in cohort studies of causal effects. Stat. Med., 33: 1685-1699.

Fullerton, B., B. Pöhlmann, R. Krohn, J.L. Adams and F.M. Gerlach *et al*., 2016. The comparison of matching methods using different measures of balance: Benefits and risks exemplified within a study to evaluate the effects of German disease management programs on long□term outcomes of patients with type 2 diabetes. Health Services Res., 51: 1960-1980. DOI: 10.1111/1475-6773.12452

Ho, D.E., K. Imai, G. King and E.A. Stuart, 2011. MatchIt: Nonparametric preprocessing for parametric causal inference. J. Stat. Software, 42: 1-28.

Iacus, S.M., G. King and G. Porro, 2011. Multivariate matching methods that are monotonic imbalance bounding. J. Am. Stat. Assoc., 106: 345-61. DOI: 10.1198/jasa.2011.tm09599

Iacus, S.M., G. King and G. Porro, 2012. Causal inference without balance checking: Coarsened exact matching. Political Anal., 20: 1-24. DOI: 10.1093/pan/mpr013

Jacovidis, J.N., 2017. Evaluating the performance of propensity score matching methods: A simulation study. James Madison University.

King, G., R. Nielsen, C. Coberley, J.E. Pope and A. Wells, 2011. Comparative effectiveness of matching methods for causal inference. Unpublished Manuscript.

LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. DOI: 10.1257/aer.p20151025

Lee, B.K., J. Lessler and E.A. Stuart, 2010. Improving propensity score weighting using machine learning. Stat. Med., 29: 337-46. DOI: 10.1002/sim.3782

Miksch, A., G. Laux, D. Ose, S. Joos and S. Campbell *et al*., 2010. Is there a survival benefit within a German primary care-based disease management program? Am. J. Managed Care, 16: 49-54. DOI: 10.1186/s13098-015-0065-9

Normand, S.L.T., M.B. Landrum, E. Guadagnoli, J.Z. Ayanian and T.J. Ryan *et al*., 2001. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. J. Clin. Epidemiol., 54: 387-98. DOI: 10.1016/S0895-4356(00)00321-8

Pan, W. and H. Bai, 2015. Propensity score analysis: Concepts and issues. Propensity Score Anal.: Fundamentals Dev.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P.R., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70: 41-55. DOI: 10.1093/biomet/70.1.41

Rosenbaum, P.R. and D.B. Rubin, 1984. Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc., 79: 516-524. DOI: 10.2307/2288398

Rosenbaum, P.R. and D.P. Rubin, 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat., 39: 33-38. DOI: 10.1017/CBO9780511810725.019

Rosenbaum, P.R., 1989. Optimal matching for observational studies. J. Am. Stat. Assoc., 84: 1024-1032. DOI: 10.1080/01621459.1989.10478868

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol., 66: 688-688. DOI: 10.1037/h0037350

Scott, D.W., 2009. Sturges' rule. Wiley Interdisciplinary Reviews: Comput. Stat., 1: 303-306. DOI: 10.1002/wics.35

Sekhon, J.S., 2011. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. J. Stat. Software, 42: 1-52.

Stock, S., A. Drabik, G. Büscher, C. Graf and W. Ullrich *et al*., 2010. German diabetes management programs improve quality of care and curb costs. Health Affairs, 29: 2197-2205. DOI: 10.1377/hlthaff.2009.0799

Stone, C.A. and Y. Tang, 2013. Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. Practical Assess. Res. Evaluat.

Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. J. Inst. Math. Stat. DOI: 10.1214/09-STS313.

Windt, R. and G. Glaeske, 2010. Effects of a German asthma disease management program using sickness fund claims data. J. Asthma, 47: 674-679. DOI: 10.3109/02770900903556421

Modern Applied Science; Vol. 13, No. 4; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

A Weighted Covariate Balancing Method of Estimating Causal Effects in Case-Control Studies

Lateef B. Amusa[1], Temesgen Zewotir[1] & Delia North[1]

[1] Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Correspondence: Lateef B. Amusa, Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa. Tel: 27-659-904-686. E-mail: amusasuxes@gmail.com

Received: October 7, 2018 Accepted: November 18, 2018 Online Published: March 13, 2019

doi:10.5539/mas.v13n4p40 URL: https://doi.org/10.5539/mas.v13n4p40

Abstract

Propensity score methods have dominated the estimation of treatment effects based on observational data and particularly in the health and medical sciences. We propose a weighting method based on rank-based Mahalanobis distance, namely the covariate balancing rank-based Mahalanobis distance method, to estimate causal effects for observational data. Using Monte Carlo simulations, under different data structures and type of outcome variables, the proposed method is shown to have better performance, in terms of bias reduction and treatment effect estimation. Specifically, under the generalized linear model framework, we simulated datasets based on the Lalonde-PSID study, for linear link function; while datasets were simulated based on the Lindner study, for non-linear link functions. We further apply the proposed method to data extracted from the Nigeria Demographic Health Survey (2013), to investigate the effect of educational exposure on ideal family size among married couples in Nigeria. The proposed method is a viable alternative method that can improve covariates balance, bias reduction, and efficient estimation of treatment effects.

Keywords: weighting, covariate balance, generalized linear models, monte carlo simulation, treatment effect

1. Introduction

A principal objective of health outcomes research is to estimate the causal effect of a treatment or intervention on an outcome variable. Inferences made in observational studies, because the treatment assignment is devoid of randomization, are not regularly clear and straightforward.

In recent times, weighting methods have taken centre stage as a pre-processing procedure, which aims at improving the balance of background covariates, and efficiently estimating treatment effects. Weighting is a nonparametric balancing procedure, which applies weights to sample units to equal the distribution of a target population.

The literature on weighting methods has been dominated by the inverse probability of treatment weights (IPW), which originates from survey research (Crump, Hotz, Imbens, & Mitnik, 2009; Hirano & Imbens, 2001; Hirano, Imbens, & Ridder, 2003; Imbens, 2004). The idea of IPW was formed from the Horvitz-Thompson weight (Horvitz & Thompson, 1952), which for each sample unit is the inverse of the probability of such unit being assigned to the observed group. Despite their popularity, propensity score methods, with specific reference to IPW, rely heavily on the correct specification of the propensity score model - slight misspecification of the propensity score model will result in a substantial bias of estimated treatment effects (Kang & Schafer, 2007).

In this paper, we introduce a rank-based Mahalanobis distance weighting approach, namely, the covariate balancing rank-based Mahalanobis distance (CBRMD) method, to efficiently estimate treatment effects, in the presence of confounding factors. We show how to use a modified Mahalanobis distance, the rank-based Mahalanobis distance, proposed by Rosenbaum (Rosenbaum, 2002), as weights that can reduce covariates imbalance between treated and control groups, which are used to estimate treatment effects efficiently. In brief, we fix weights for the treated group sample units at unity, while those for control group units are obtained as the number of times a control unit has the smallest rank based Mahalanobis distance from the individual treated units.

We illustrate the general framework of the proposed method in the Methodology section. The performance of the proposed method is evaluated through a series of Monte Carlo simulations and a case study of data on the effect

of educational exposure on the desired family size among married couples in Nigeria. Using the IPW method as a benchmark, we study the effectiveness of the proposed technique in balancing covariates, and efficient estimation of treatment effects. Our choice of the IPW as a benchmark for evaluating the performance of the proposed method is due to its simplicity and familiarity.

2. Methodology

Consider a random sample of n= n_t + n_c units, with each i (i =1, . . ., n), belonging to only one of two groups for which estimation of causal effects are of interest, denoted by T_i. The ith unit received the treatment of interest, if T_i = 1, and T_i = 0, if it was not received (control group). Let X_i' denote a K-dimensional vector of observed pre-treatment covariates associated with unit i. Adopting the potential outcomes framework, we let $Y_i(1)$ be the potential outcome that unit i attains under treated group and $Y_i(0)$ the potential outcome under control group (Rubin, 1974). The observed outcome can then be represented as $Y_i = T_iY_i(1) + (1 - T_i) Y_i(0)$. We estimate the Sample Average Treatment effect on the Treated (SATT) as, SATT $= \frac{1}{n_t}\sum_{i\epsilon T} TE_i$, where $TE_i = Y_i(1) - Y_i(0)$.

The Mahalanobis distance between covariates of the treated unit, X_t and covariates of the control unit, X_c can be obtained from:

$$D^2 (X_t,\ X_c) = (\boldsymbol{X}_t - \boldsymbol{X}_c)^T \widehat{\sum}^{-1} (\boldsymbol{X}_t - \boldsymbol{X}_c) \quad (1)$$

where $\widehat{\sum}$ is the estimated sample covariance of **X**.

For multivariate normal covariates, the Mahalanobis distance works fine; but exhibit some rather odd behaviour with non-normal data and outliers-present data. Consequently, we replace the Mahalanobis with a rank-based Mahalanobis distance, defined by Rosenbaum (Rosenbaum, 2002) as follows:

$$rD^2 (X_t,\ X_c) = \left(r(\boldsymbol{X}_t) - r(\boldsymbol{X}_c)\right)^T adj\, \widehat{\sum}^{-1} \left(r(\boldsymbol{X}_t) - r(\boldsymbol{X}_c)\right) \quad (2)$$

where, $r(\boldsymbol{X}_t)\ and\ r(\boldsymbol{X}_c)$ are the ranks of each of the covariates belonging to the treated and control groups, respectively. Average ranks are used for ties.

Further, note that $adj\,\widehat{\sum}$ denotes adjusted covariance matrix, which adjusts the $\widehat{\sum}$ (variance-covariance matrix of the ranked covariates) by pre-multiplying and post-multiplying the covariance matrix of the ranks by a diagonal matrix whose diagonal values are the ratios of untied ranks' standard deviation, to the tied ranks' standard deviations of the covariates. In other words, adj $\widehat{\sum}$ is defined as:

$$\text{adj } \widehat{\sum} = \mathbf{D} \sum \mathbf{D} \quad (3)$$

where,

$$\mathbf{D} = \begin{bmatrix} \frac{S_u}{S_{t1}} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \frac{S_u}{StK} \end{bmatrix}$$

S_u is the standard deviation of untied ranks, and S_{tK} is the standard deviation of tied ranks for the kth covariate.

From the matrix $\mathbf{rD^2}$ with dimension t x c, where t is the number of treated units and control units, respectively, the proposed algorithm extracts the control units and its corresponding rank-based Mahalanobis distance on each row of the matrix.

Finally, sample weights for treated units are fixed at unity, while those for control group units are given as the number of times a control unit has the smallest rank-based Mahalanobis distance from the individual treated units.

If any control unit does not have the smallest rank-based Mahalanobis distance from any treated unit, the CBRMD procedure does not give it a weight of zero. Instead, it only down-weights them. When there are ties in the control units that have the least rank-based Mahalanobis distance from any treated unit, the weight is approximately equally distributed among them so that every sample unit contributes to the estimation, which in turn improves balance, reduces bias, and maximizes efficiency.

The proposed algorithm is described in the following steps:

Step 1: Sort the data in order of the treatment indicator, with the corresponding unit identification number.

Step 2: Compute the rank-based Mahalanobis distances of each treated units with the control group units, using Equation (2), and store the distances in a matrix with t rows and c columns.

Step 3: Create a vector which stores the column number of the control unit that has the smallest distance with the treated units in each row.

Step 4: Extract a frequency distribution based on Step 3, to identify the number of times each control unit had the smallest distance. Control units with zero frequencies are down-weighted approximately equally.

Step 5: Treated units have weights that are fixed at 1; while control units have weights based on step 4.

3. Simulations

3.1 Monte Carlo Simulations - Overview

In this section, we study the numerical performance of the proposed methodology. We conducted extensive Monte Carlo simulations to examine the performance of the proposed method, and compare its performance to that of IPW. Performance of the methods was assessed using absolute standardized bias of covariates; absolute biases and root mean squared errors (RMSE) of the estimated treatment effects. The data generating process and analyses were conducted in the R environment of version 3.4.3.

Two different phases of simulations were conducted overall. In the first phase, the simulation was made to be as realistic as possible by simulating from real-life data, which was achieved by explicitly focusing on two distinct scenarios:

In the first scenario, subsequently referred to as Scenario I, we generated treatment and outcome variables from covariates of the Lalonde-PSID data. This data has a reputation of being used as a benchmark in the causal inference literature. The data is a hybrid of program participants (treated units) from Lalonde's (LaLonde, 1986)experimental data and control group drawn from the Panel Study of Income Dynamics (PSID) data. The dataset comprises of ten covariates including age (*age*), indicator variables for unemployment in 1974 (*u74*) and 1975 (*u75*), marital status (*married*), lack of a high school diploma (*nodegree*), number of years of education (*education*), hispanic race (*hispanic*), black race (*black*), and real earnings in 1974 (*re74*) and 1975 (*re75*). The outcome was the real earnings in 1978. Choice of this data will enable us to evaluate how well our proposed method can recover the treatment effect estimates from the experimental data.

For the second scenario, subsequently referred to as Scenario II, we extend our evaluation to non-normal responses. We specifically consider three types of outcomes: binary outcomes (Binomial distribution), counts (Poisson distribution), and skewed continuous outcomes (Gamma distribution). The idea is to mirror some outcome variables that are mostly encountered in medical and health sciences. For example, presence or absence of diseases, number of antenatal care visits by pregnant women, and health care costs are usually described by the Binomial, Poisson, and Gamma distributions, respectively.

The simulations were based on the Lindner dataset. Details of this data have been published elsewhere (Abdia, Kulasekera, Datta, Boakye, & Kong, 2017). In brief, the Lindner dataset comprises information on 996 patients who were receiving an initial Percutaneous Coronary Intervention (PCI) at the Ohio Heart Health, Lindner Christ Hospital in 1997. The treatment indicator (*abcix)*, equals 1 when the patient was in PCI treatment with additional treatment abciximab (an expensive, high-molecular-weight IIb/IIIa cascade blocker), and 0 when the patient was in PCI group. Covariates include, indicator for recent acute myocardial infarction (*acutemi*); indicator for coronary stent insertion (*stent*); gender (*female*); height; left ventricle ejection fraction (*ejecfrac*); number of vessels involved in initial PCI (*ves1proc*); diabetic indicator (*diabetic*); and an indicator for survival at six months (*sixMonthSurvive*).

3.2 Monte Carlo Simulations – Data Generation

For scenarios I and II, like (Austin, 2011; Austin & Stuart, 2017), we assume a linear relationship between log-odds of treatment assignment and covariates from their respective real data, as shown in Equations (4) and (5).

$$\text{Logit}(\pi_i) = \alpha_0 + \alpha_1\ \text{age} + \alpha_2\ \text{education} + \alpha_3\text{re74} + \alpha_4\,\text{re75} + \alpha_5\ \text{married} + \alpha_6\ \text{black} + \alpha_7\,\text{hispanic} + \alpha_8\ \text{nodegree} + \alpha_9\ \text{u74} + \alpha_{10}\,\text{u75} \quad (4)$$

$$\text{Logit}(\pi_i) = \alpha_0 + \alpha_1\ \text{stent} + \alpha_2\ \text{height} + \alpha_3\text{female} + \alpha_4\,\text{diabetic} + \alpha_5\ \text{acutemi} + \alpha_6\ \text{ejecfrac} + \alpha_7\,\text{ves1proc} + \alpha_8\ \text{sixMonthSurvive} \quad (5)$$

To ensure varied number of treated and control units, we then generate the treatment variable for individual *i,* in 1000 separate runs as $T_i \sim \text{Bernoulli}(\pi_i)$.

In assessing covariates balance, we average the Absolute Standardized Bias (ASB) for each covariate, from the 1000 runs of the above data generation. ASB is given as:

$$\text{ASB} = \frac{1}{\sqrt{\frac{s^2t + s^2c}{2}}} \left| \frac{\sum_{i=1}^{n} x_i T_i w_i}{\sum_{i=1}^{n} T_i w_i} - \frac{\sum_{i=1}^{n} x_i (1-T_i) w_i}{\sum_{i=1}^{n} (1-T_i) w_i} \right| \quad (6)$$

where s^2t and s^2c are the sample variances of the covariate in the treated and control group respectively. For weighted data, $s^2 = \frac{\sum w_i}{(\sum w_i)^2} \sum w_i (x_i - \bar{x})^2$, where $\bar{x} = \frac{\sum x_i}{n}$.

We generate outcome variables differently for the two scenarios. For scenario I, we assume the following linear model:

$$Y = \beta_0 + \gamma T_i + \beta_1 age + \beta_2 education + \beta_3 re74 + \beta_4 re75 + \beta_5 married + \beta_6 black + \beta_7 hispanic + \beta_8 nodegree + \beta_9 u74 + \beta_{10} u75 + \varepsilon_i, \quad \varepsilon_i \sim N(0,10) \tag{7}$$

Following (Diamond & Sekhon, 2013), we set $\gamma = 1000$, and $\beta_0, \beta_1, \ldots, \beta_{10}$ are coefficients from linearly regressing the outcome on the covariates from the real data.

For scenario II, data are generated from the following generalized linear model

$$g(E(Y/X, Z)) = \beta_0 + \sum_{j=1}^{8} \beta_j X_j + \gamma T, \tag{8}$$

Where g is chosen to be the canonical link function for Binomial, Poisson, and Gamma distribution, respectively. The X_j's are the 8 covariates from the Lindner dataset.

Following (Austin, 2011), model coefficients for Equation (8), are set as $\beta_0 = 0$, $\beta_1 = \beta_2 = \log(1.1)$, $\beta_3 = \beta_4 = \log(1.25)$, $\beta_5 = \beta_6 = \log(1.5)$, and $\beta_7 = \beta_8 = \log(2)$. The non-zero coefficients are chosen to reflect low, medium, high and very high effect sizes.

3.2.1 Data Generation – Binary Outcomes (Binomial Distribution)

For binary outcomes, (8) becomes:

$$\text{Logit}(P(Y_i=1)) = \beta_0 + \sum_{j=1}^{8} \beta_j X_j + \gamma T_i \tag{9}$$

Equation (9) is the conventional logistic regression model that is usually encountered in clinical and epidemiological research. The regression parameter γ is the log-odds ratio for the treatment effect. The model assumes that the logit of the probability of outcomes changes with a subject's change in treatment status. The odds ratio is exp (γ) and has been described as a conditional or adjusted treatment effect (Austin, 2010). The logarithmic link function in Equation (9) does not require covariates from a distribution with support over the real line. For this reason, covariates from (9) are reduced to only the five binary covariates. Therefore, Equation (9) reduces to:

$$\text{Logit}(P(Y_i=1)) = \beta_0 + \sum_{j=1}^{5} \beta_j X_j + \gamma T_i \tag{10}$$

Coefficient γ is set to 1, while the estimated treatment effects were transformed to be on the odds ratio scale.

3.2.2 Data Generation – Count Outcomes (Poisson Distribution)

For count outcomes, Equation (8) becomes:

$$\text{Log}(\eta_i) = \beta_0 + \sum_{j=1}^{8} \beta_j X_j + \gamma T_i \tag{11}$$

The regression parameter γ is the expected change in log count, as treatment status changes from treated to control. The continuous covariates were scaled to have zero mean and unit variance, to permit the possible values of the log link function. Coefficient γ is set to 1, while the estimated treatment effects are on the log-rate ratio scale.

3.2.3 Data Generation – Skewed Continuous Outcomes (Gamma distribution)

For skewed continuous outcomes, Equation (8) becomes:

$$1/(\eta_i) = \beta_0 + \sum_{j=1}^{8} \beta_j X_j + \gamma T_i \tag{12}$$

The regression parameter γ is the expected change in the inverse of outcomes, as treatment status changes from treated to control. Coefficient γ is set to 1500, while the estimated treatment effects are on a natural scale.

3.3 Kang and Schafer Design

This second phase of simulation follows the Kang and Schafer (Kang & Schafer, 2007) design, which showed that misspecification of a propensity score model could adversely affect weighting methods that depend on the propensity score. This design has been used in the literature to evaluate the performance of propensity score methods when the true propensity score is known, and when it is unknown. Note that the true propensity score was unknown in the earlier phase of simulations. Also, this phase of simulations is only introduced for the estimation of treatment effects. Though this simulation phase may achieve other objectives, the main aim is to compare the proposed method and IPW method, under the case where IPW is expected to perform optimally.

We replicate the Kang and Schafer simulation study, using 1000 Monte Carlo simulation runs, for sample sizes, 200, 1000, and 5000. In brief, the design's data generation is as follows:

$Y_i = 210 + 27.4\ X_{i1} + 13.7(X_{i2} + X_{i3} + X_{i4}) + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$, the X_i's are independently standard normally distributed, and the true propensity scores are

$$\pi_i = \frac{1}{1+e^{(-X_{i1}+0.5X_{i2}-0.25X_{i3}-0.1X_{i4})}}$$

3.4 Assessing Performance of Treatment Effects

For each phase, scenario and model, 1000 datasets were simulated. The performance of estimated treatment effects was assessed by calculating the mean $\bar{\gamma}$ of the 1000 regression coefficients. The bias was calculated as $\bar{\hat{\gamma}} - \gamma$, and the root mean square error (RMSE) as $\sqrt{(\bar{\hat{\gamma}} - \gamma)^2 + var(\hat{\gamma})}$.

4. Results

4.1 Monte Carlo simulations - Results

In this subsection, we present and explain the results obtained from analysing the simulated datasets. Figures 1 and 2 visualises balance of each of the ten covariates after applying the proposed method and IPW. We superimposed horizontal lines on each panel to denote ASB of 0.25, as some authors have suggested that ASB values that exceed this threshold may indicate significant imbalance(Ho, Imai, King, & Stuart, 2007; Imai, King, & Stuart, 2008; McCaffrey et al., 2013). Simulations from the heavily imbalanced Lalonde data, produced datasets whose average ASB values ranged from 0.125 to 1.850. Our proposed method substantially improved the balance on the ten covariates, with average ASB values ranging from 0.023 to 0.219, while the IPW adjusted data have average ASB values ranging from 0.125 to 1.850 and 0.146 to 0.887, respectively. The Lindner data is moderately imbalanced, as datasets simulated from it, had average ASB values ranging from 0.052 to 0.428. Both set of weights substantially improved the balance, as average ASB values ranging from 0.007 to 0.176 for the proposed method, and 0.019 to 0.049 for the IPW adjusted data.

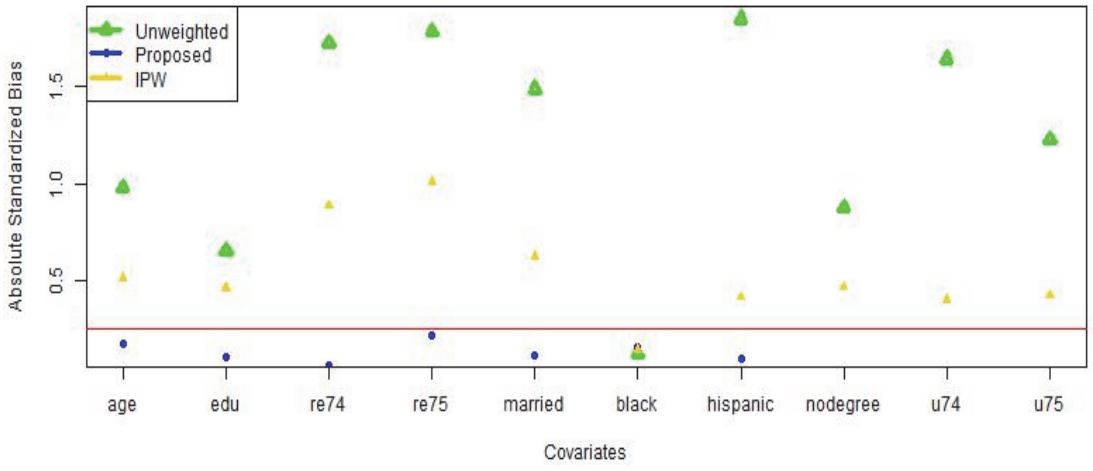

Figure 1. Plot of mean absolute standardized bias of covariates in the Lalonde data, under each weighting method

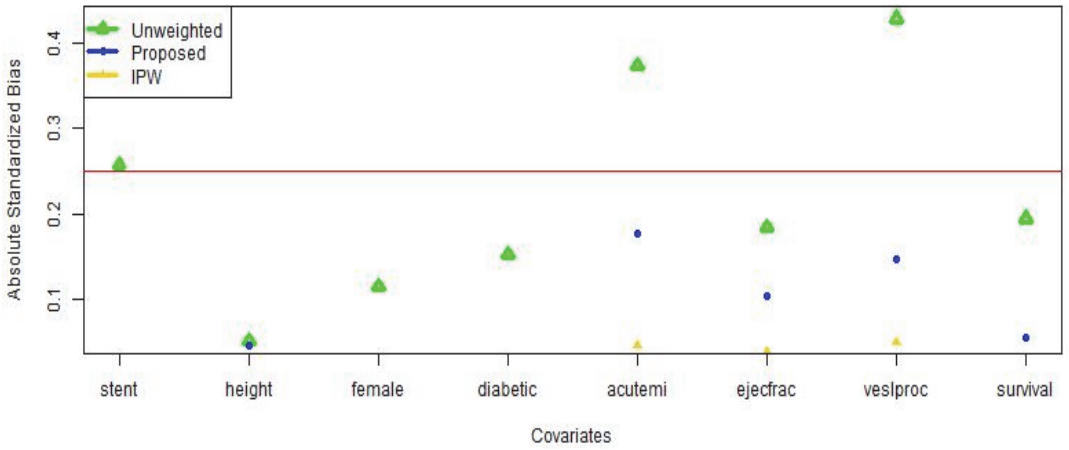

Figure 2. Plot of mean absolute standardized bias of covariates in the Lindner data, under each weighting method

In the first scenario, where the datasets were simulated from the Lalonde-PSID data and outcome variable is assumed normally distributed, there is an excellent performance of the proposed method, in terms of absolute bias and RMSE of estimated treatment effects. The proposed technique, as compared to others, has the least absolute bias and RMSE, resulting in a 65% reduction in these performance metrics. The extremely high values of the bias and RMSE is not surprising, given the high variance of the outcome variable. The average absolute biases and RMSEs of the weighting methods are shown in Table 1.

Table 1. Relative performance of the weighting methods under Scenario I

Method		
	Absolute bias	RMSE
Unweighted	67265.76	67274.98
Proposed	22908.59	23320.30
IPW	28584.00	31722.93

Note: values were averaged over 1000 Monte Carlo replications.

The estimated treatment effects for the weighting methods, under each type of outcome distribution, are reported in Figure 3. The lower left panel of the plot shown the treatments for the normally distributed outcomes, lower right panel for the binomially distributed outcomes, upper left panel for the Poisson distributed outcomes, and the upper right panel for the Gamma distributed outcomes.

Except for binomially distributed outcomes, the proposed method dominates the others both in terms of absolute bias and RMSE, as shown in Table 2. The most substantial outperformance for the proposed method was observed for Poisson distributed outcomes, where approximately 75% reduction in both absolute bias and RMSE was achieved. IPW method had the least absolute bias and RMSE for binomial outcomes, even though the difference, as compared to the other methods was not of considerable importance. For Gamma distributed outcomes, IPW method had no reduction in the absolute bias and RMSE, but slightly increased it instead; while the proposed method reduced the performance metrics by approximately 16%.

Table 2. Relative performance of the weighting methods under Scenario II

Method	Gamma		Poisson		Binomial	
	Absolute bias	RMSE	Absolute bias	RMSE	Absolute bias	RMSE
Unweighted	1491.92	1494.21	15.95	15.96	1.88	1.95
Proposed	1255.06	1276.16	3.94	4.09	1.80	1.93
IPW	1492.27	1494.72	13.15	13.18	1.73	1.82

Note: values were averaged over 1000 Monte Carlo replications.

Having established the performance of the proposed method under situations where the true propensity is unknown, Table 3 shows the result from an unusual situation where the correct propensity score model is specified. There is a bias-variance trade-off, as the proposed method consistently (over the considered sample sizes) had the least absolute bias at the expense of some increase in RMSE. Both weighting methods show an overall reduction in bias and RMSE, as compared to the unadjusted data. Also, there is a pattern of improved performance of the proposed method and the IPW method, when the sample size increases. This experiment has shown that the performance of IPW method (when the correct propensity score model is known) will only be better than the proposed method in terms of efficiency and not bias reduction.

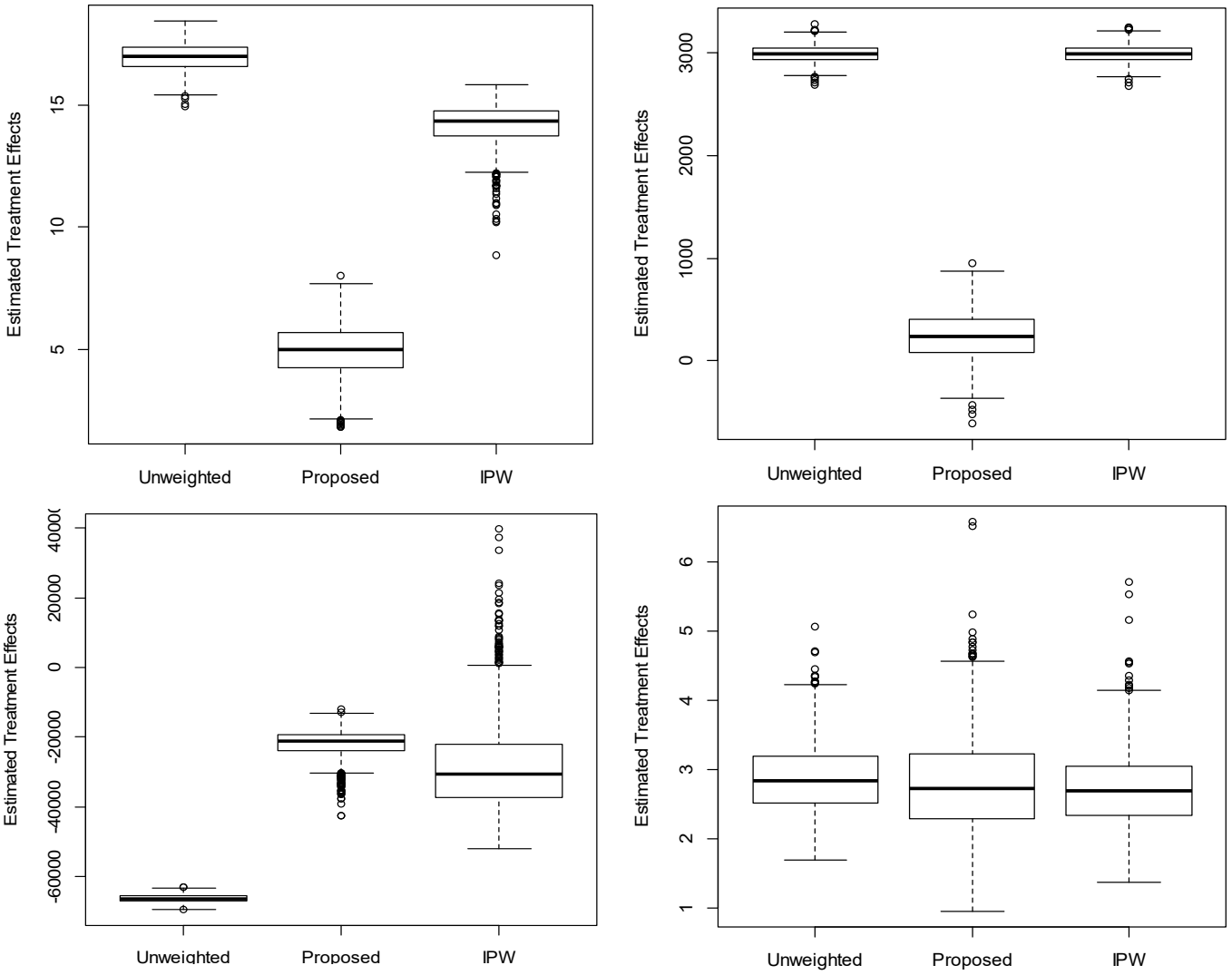

Figure 3. Boxplot of estimated treatment effects for the weighting methods.

Table 3. Relative performance of the weighting methods under a correct propensity score, based on Kang and Schafer (2007)

Sample size	Method	Bias	RMSE
200	*Unweighted*	19.958	20.523
	Proposed	0.101	7.468
	IPW	0.421	4.815
1000	*Unweighted*	19.979	20.100
	Proposed	0.061	3.354
	IPW	0.069	2.344
5000	*Unweighted*	20.026	20.051

	Proposed	0.007	1.071
	IPW	0.007	1.565
10000	*Unweighted*	19.995	20.007
	Proposed	0.002	1.122
	IPW	0.006	0.722

Note: values were averaged over 1000 Monte Carlo replications

5. Case Study

In this section, we apply the proposed method to extracted data from the Nigeria Demographic Health Survey of 2013. The Demographic and Health Survey (DHS) provides cross-sectional data on demographic and health indicators, including information on fertility and family planning, knowledge and current use of contraception methods, as well as sexually transmitted diseases (NDHS, 2013). Further details of this data can be found elsewhere (Amusa, 2018).

The sample consists of 18842 married respondents aged 15-49, 6373 of whom had at least a secondary school education, subsequently regarded as 'treated' group, and others 12469 had primary school or no formal education, regarded as 'control' group. The research question of interest is whether educational exposure causes a higher desired number of children (outcome variable). The data set includes information on ten covariates that potentially confound the treatment-outcome relationship. Table 4 presents the description of data variables and summary statistics, including averages and standard deviations for continuous variables, and percentages for categorical variables, as well as the absolute standardized bias as the balance metric.

Table 4. Summary statistics of baseline covariates of the two treatment groups in the case study. For continuous variables, the mean (standard deviation) is presented; for binary variables, the frequency (percentage) is presented.

Label	Variable Description	Treated	Control	
		N = 6373	N = 12469	ASB
Bmi	Body Mass Index	25.30 (5.00)	22.64 (4.14)	0.58
Age	Age of respondent	32.09 (7.81)	31.99 (8.74)	0.01
Agebirth	Age at first birth	28.87 (6.09)	28.94 (7.33)	0.01
Mbirth	Interval of marriage to birth (months)	18.41 (19.39)	27.98 (27.57)	0.40
Siblings	Number of siblings of respondent	5.39 (2.44)	5.42 (2.80)	0.01
Knowledge	Knowledge of any birth control method	6272 (38.20%)	10145 (61.90%)	0.59
Wealth	Wealth index (poor = 1)	980 (27.83%)	2541 (72.17%)	1.58
Res	Residence type (rural = 1)	3909 (57.74%)	2861 (42.26%)	0.84
Sexhead	Sex of household head (male = 1)	5436 (31.82%)	11649 (68.18%)	0.27
Working	Respondent is working	5099 (37.39%)	8538 (62.61%)	0.27

Note: ASB denotes absolute standardized bias.

The ASB values, using the >0.25 threshold (as used in the simulation study), suggest that the covariates balance is not satisfactory for all seven out of the ten background covariates. The ASB values ranged from 0.01 to 1.58. We applied the proposed method using the ten variables. We also implemented the IPW method by estimating propensity scores from a linear logit specification of treatment-covariates relationship on all ten covariates. For each of the ten covariates, Figure 4 visualises the covariate balance obtained from the different weighting methods as measured by the conventional balance statistics – absolute standardized bias. A horizontal line at ASB = 0.25 was superimposed (as in the simulation study) to denote the balance threshold of the covariates.

Results from Figure 4 reveal that, though the proposed method maximized the improvement in balance, better than the IPW method, they both substantially improved the mean balance compared to the raw data. The proposed method has ASB values ranging from as small as zero to 0.062; while the IPW method had values ranging from 0.003 – 0.060.

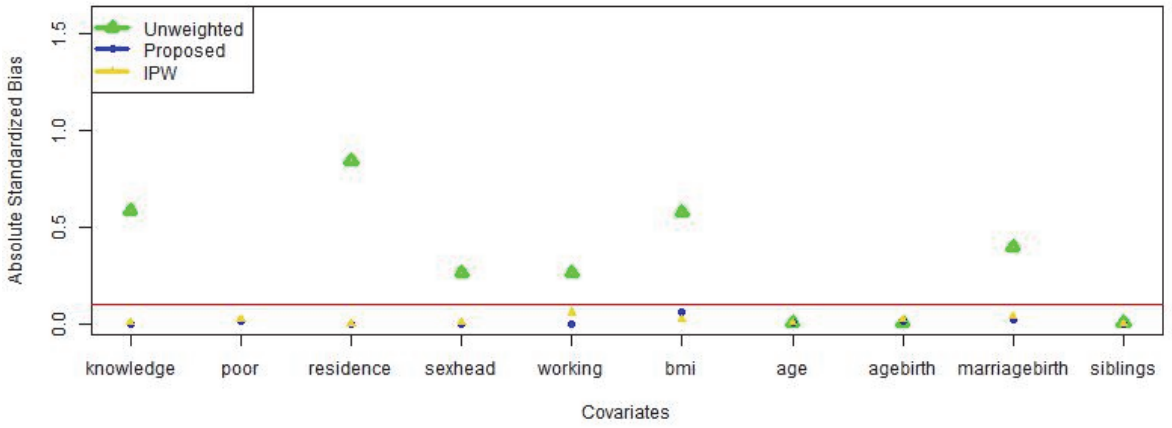

Figure 1. Plot of mean absolute standardized bias of covariates in the case study data, under each weighting method

Table 5 shows the point estimates, p-value, and associated 95% confidence interval from the weighting methods. The standard errors used to calculate confidence intervals for the proposed method and IPW are based on the robust sandwich variance estimator (Austin & Stuart, 2015; Joffe, Ten Have, Feldman, & Kimmel, 2004). Estimates from the weighted estimators were entirely different from the unweighted estimator. All the estimators were negative and statistically significant ($p<0.05$) at the 5% level, which suggests that educational exposure (having at least a secondary school education) decreases the expected number of desired children by married couples. The proposed weighted estimator produced a confidence interval with a slightly shorter length compared to the IPW estimator.

Table 5. Causal effect estimation of educational exposure on desired family size, using the various methods

Estimator	Point Estimate	95% Confidence Interval (CI)	CI Length	P-value
Unweighted	-0.444	(-0.4573, -0.4327)	0.0246	<0.0001
Proposed	-0.244	(-0.2644, -0.2242)	0.0402	<0.0001
IPW	-0.28	(-0.3095, -0.2505)	0.059	<0.0001

Note: Standard errors of the weighted estimators were based on the robust sandwich variance estimators

6. Discussion

Estimation of causal effects is central to health outcomes. In this study, we have proposed a new weighting method which is based on computations from a rank-based Mahalanobis distance. We showed through simulations and an empirical application, the effectiveness of the proposed method in terms of improvement in covariates balance, bias reduction and efficient estimation of treatment effects.

The proposed covariate balancing rank-based Mahalanobis distance (CBRMD) method is a novel approach to estimating causal effects, in the presence of confounding factors, as the case of observational studies. We have been able to demonstrate numerically, the excellent performance of the proposed method, to induce balance on background covariates, as well as, a notable reduction in the bias and increased efficiency of the estimated treatment effects. Notably also, is the fact that the CBRMD method performs at its best when the sample size is huge - this was evident from the results obtained from the simulations, and the case study from the large sample NDHS data, that was done in this study. Large sample sizes are typical of epidemiological studies and national surveys.

There has been overwhelming usage of propensity score weighting methods among applied researchers in various disciplines who conduct causal inference in observational studies. We only acknowledge that propensity score methods have become more familiar, hence the reason for adoption.

The commonly used IPW method relies heavily on the correct propensity score model specification. Model

misspecification will substantially bias its estimation of treatment effects. However, as shown from the simulations, the proposed method still favourably competes with the IPW, under situations where the correct propensity score model is specified. An interesting bias-variance trade-off was observed from our simulations when the correct propensity score was known, with our proposed method consistently having the least absolute bias but slightly trailing the IPW method in terms of efficiency, as measured by root mean squared error.

One of the major strength of this study is the development of the simulation study based on notable existing real-life studies. This approach of designing simulations based on real-life studies is increasingly becoming the norm, as it allows the researcher to incorporate complex and realistic associations within the data structure. The fact that outcome variables from different distributions under the GLM framework were considered is also a strength of this study.

We acknowledge that the IPW method was only used as a benchmark for evaluating the performance of the proposed method. Though this study has briefly compared the IPW method under the situation where the correct propensity score model is known, future research is required for extensively comparing the two methods under varying scenarios before we can recommend one over the other. For now, evidence from this study can only advise researchers and applied practitioners to adopt the proposed method when the correct propensity score model is not known. Future researches may consider the combination of CBRMD with other pre-processing methods. We are currently exploring these.

7. Conclusions

When causal effects are of interest in the presence of confounding variables, as the case of observational studies, the proposed covariate balancing rank-based Mahalanobis Distance (CBRMD) method is a viable alternative method, that can improve covariates balance, bias reduction and efficient estimation of treatment effects.

Acknowledgements

We thank ORC macro, Measure DHS for giving us access to the data _le from which the dataset utilized for the empirical application was extracted. The corresponding author would also like to appreciate the University of Ilorin, Nigeria, for granting him a leave of absence to pursue his PhD studies in South Africa.

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal, 59*(5), 967-985.

Amusa, L. B. (2018). Reducing bias in Observational Studies: An Empirical Comparison of Propensity Score Matching Methods. *Turkiye Klinikleri Journal of Biostatistics, 10*(1).

Austin, P. C. (2010). A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics—Simulation and Computation®, 39*(3), 563-577.

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., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661-3679.

Austin, P. C., & Stuart, E. A. (2017). Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research, 26*(6), 2505-2525.

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

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932-945.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2*(3-4), 259-278.

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *Roy. Statist. Soc. A, 171*, 481-502.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician, 58*(4), 272-279.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science, 22*(4), 523-539.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604-620.

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

NDHS. (2013). Nigeria Demographic Health Survey. *Published by National Population Commission*.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.): New York: Springer-Verlag.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology, 66*(5), 688.

Appendix B: Selected R codes

Some selected R codes for Chapters 3, 4, 5, and 6

```
# Chapter 3
# R codes for Scenario 1
rm(list=ls())
require(cem)
require(MatchIt)
require(Matching)
library(ggplot2)
library(ggpubr)
library(MASS)
ATT=1
sigma<- diag(10)
sigma[1,5]=sigma[5,1]=sigma[8,3]=sigma[3,8]=0.2
sigma[2,6]=sigma[6,2]=sigma[4,9]=sigma[9,4]=0.9

MCSim <- 1000
fun=function(rate,n)
{
set.seed(123456)
tmp <- matrix(NA, MCSim, 8)
sizes <- matrix(NA, MCSim, 8)
colnames(sizes) <- c("RAW(n)", "CEM(n)", "MAH1(n)", "MAH2(n)", "PSC1(n)", "PSC2(n)",
    "FULL1(n)", "FULL2(n)")
```

```
colnames(tmp) <- c("RAW", "CEM", "MAH1", "MAH2", "PSC1", "PSC2", "FULL1", "FULL2")
for(MC in 1:MCSim){
x <- mvrnorm(n, mu = rep(0, 10), Sigma = sigma,empirical=T)
x1<- x[,1]; p1=pnorm(x1);  xx1=rbinom(n,1,p1)
x2<- x[,2]
x3<- x[,3]; p3=pnorm(x3);  xx3=rbinom(n,1,p3)
x4<- x[,4]
x5<- x[,5]; p5=pnorm(x5);  xx5=rbinom(n,1,p5)
x6<- x[,6]; p6=pnorm(x6);  xx6=rbinom(n,1,p6)
x7<- x[,7]
x8<- x[,8]; p8=pnorm(x8);  xx8=rbinom(n,1,p8)
x9<- x[,9]; p9=pnorm(x9);  xx9=rbinom(n,1,p9)
x10<- x[,10]
X=cbind(xx1,x2,xx3,x4,xx5,xx6,x7)
b0=rate
b1<- 0.8; b2 <- -0.25; b3<- 0.6; b4<- -0.4; b5<- -0.8; b6<- -0.5; b7<- 0.7
a1<- log(2); a2<- log(2); a3<- log(2); a4<- log(1.75); a5<- log(1.75)
a6<- log(1.75); a7<- log(1.5)
M <- cbind(rep(1, n),xx1,x2,xx3,x4,xx5,xx6,x7)
mod.coeffs <- as.matrix(c(b0,b1,b2,b3,b4,b5,b6,b7))
mu = M %*% mod.coeffs
Tr.pred <- exp(mu)/(1+exp(mu))
t <- matrix(nrow=n, ncol=1)
for(i in 1:n)
t[i] = sample(0:1, 1, prob=c(1-Tr.pred[i],Tr.pred[i]))
dat=data.frame(treat=t,X)
y <- I(ATT*t) + a1*xx1+ a2*x2+ a3*xx3+ a4*x4+ a5*xx5+ a6*xx6+ a7*x7+ rnorm(n,0,1)
tsubjects <- which(t==1)
csubjects <- which(t==0)
nt <- length(tsubjects)
nc <- length(csubjects)
dati1 <- data.frame(dat,y)
```

```
cem.mat <- cem("treat", dat)
cem.tr <- which(cem.mat$groups=="1" & cem.mat$matched==TRUE)
cem.ct <- which(cem.mat$groups=="0" & cem.mat$matched==TRUE)
cem.idx <- unique(c(cem.tr, cem.ct))

mah.mat1 <- Match(Tr=t, X=X,Weight=2,M=1,replace=F)
mah.mat2 <- Match(Tr=t, X=X,Weight=2,caliper=0.25,M=1,replace=F)
mah.tr1 <- mah.mat1$index.treated; mah.tr2 <- mah.mat2$index.treated
mah.ct1 <- mah.mat1$index.control;  mah.ct2 <- mah.mat2$index.control
mah.idx1 <- unique(c(mah.tr1, mah.ct1))
mah.idx2 <- unique(c(mah.tr2, mah.ct2))

pscore <- glm(treat~., family=binomial, data=dat)
psc.mat1 <- Match(Tr=t, X=pscore$fitted,M=1,replace=F)
psc.mat2 <- Match(Tr=t, X=pscore$fitted,M=1,replace=F,caliper = 0.25)
psc.tr1 <- psc.mat1$index.treated
psc.tr2 <- psc.mat2$index.treated
psc.ct1 <- psc.mat1$index.control
psc.ct2 <- psc.mat2$index.control
psc.idx1 <- unique(c(psc.tr1, psc.ct1))
psc.idx2 <- unique(c(psc.tr2, psc.ct2))

m.out1 <- matchit(treat ~ xx1+x2+xx3+x4+xx5+xx6+x7, data =dat, method ="full")
s1=summary(m.out1,standardize = T)
m1=match.data(m.out1)

m.out2 <- matchit(treat ~ xx1+x2+xx3+x4+xx5+xx6+x7, data =dat,
    method ="full",distance="mahalanobis")
s2=summary(m.out2,standardize = T)
m2=match.data(m.out2)

nt.cem <- length(unique(cem.tr)); nc.cem <- length(unique(cem.ct)); n.cem=nt.cem+nc.cem
nt.mah1 <- length(unique(mah.tr1)); nc.mah1 <- length(unique(mah.ct1))
```

```
n.mah1=nt.mah1+nc.mah1
nt.mah2 <- length(unique(mah.tr2)); nc.mah2 <- length(unique(mah.ct2))
n.mah2=nt.mah2+nc.mah2

nt.psc1 <- length(unique(psc.tr1)); nc.psc1 <- length(unique(psc.ct1))
n.psc1=nt.psc1+nc.psc1

nt.psc2 <- length(unique(psc.tr2)); nc.psc2 <- length(unique(psc.ct2))
n.psc2=nt.psc2+nc.psc2

nt.full1<- s1$nn[2,1]; nc.full1 <- s1$nn[2,2]; n.full1=nt.full1+nc.full1
nt.full2<- s2$nn[2,1]; nc.full2 <- s2$nn[2,2]; n.full2=nt.full2+nc.full2

RAW <- mean(y[tsubjects]) - mean(y[csubjects])
CEM <- weighted.mean(y[cem.tr],cem.mat$w[cem.tr]),weighted.mean(y[cem.ct],
    cem.mat$w[cem.ct])
MAH1 <- mean(y[mah.tr1]) - mean(y[mah.ct1])
MAH2 <- mean(y[mah.tr2]) - mean(y[mah.ct2])
PSC1 <- mean(y[psc.tr1]) - mean(y[psc.ct1])
PSC2 <- mean(y[psc.tr2]) - mean(y[psc.ct2])
FULL1<- lm(y~treat,weights = weights,data=m1)$coefficients[2]
FULL2<- lm(y~treat,weights = weights,data=m2)$coefficients[2]

tmp[MC,] <-  c(RAW, CEM, MAH1, MAH2, PSC1, PSC2, FULL1, FULL2)
sizes[MC,] <-  c(n, n.cem, n.mah1,n.mah2, n.psc1,n.psc2,n.full1,n.full2)
  }
 return(list("est"=tmp,"size"=sizes))
}
res1a=fun(-1.43,600); res1b=fun(-1.43,800); res1c=fun(-1.43,1000)
res2a=fun(-0.94,600);res2b=fun(-0.94,800); res2c=fun(-0.94,1000)

cat("\n\nAverage absolute bias:\n")
```

```
bias1=abs(colMeans(res1a$est)-ATT); bias2=abs(colMeans(res1b$est)-ATT)
bias3=abs(colMeans(res1c$est)-ATT)
biasA=c(bias1,bias2,bias3)

bias4=abs(colMeans(res2a$est)-ATT); bias5=abs(colMeans(res2b$est)-ATT)
bias6=abs(colMeans(res2c$est)-ATT)
biasB=c(bias4,bias5,bias6)

cat("RMSE:\n")
mse1=sqrt(colMeans((res1a$est-ATT)^2)); mse2=sqrt(colMeans((res1b$est-ATT)^2))
mse3=sqrt(colMeans((res1c$est-ATT)^2))
rmseA=c(mse1,mse2,mse3)

mse4=sqrt(colMeans((res2a$est-ATT)^2)); mse5=sqrt(colMeans((res2b$est-ATT)^2))
mse6=sqrt(colMeans((res2c$est-ATT)^2))
rmseB=c(mse4,mse5,mse6)

cat("Average Matched units:\n")
tt1 <- colMeans(res1c$size,na.rm=TRUE)
tt2 <- colMeans(res2c$size,na.rm=TRUE)

tab1 <- matrix(round(tt1), 8,1, byrow=TRUE)
tab1; write.csv(tab1,"tab1.csv")

tab2 <- matrix(round(tt2), 8,1, byrow=TRUE)
tab2; write.csv(tab2,"tab2.csv")

df1a=matrix(c(biasA),8)
df1b=matrix(c(biasB),8)
df2a=matrix(c(rmseA),8)
df2b=matrix(c(rmseB),8)

colnames(df1a)=colnames(df1b)=colnames(df2a)=colnames(df2b)=c("n=500","n=800","n=1000")
```

```
par(mar = c(3,4,4,4.8)) #left bottom top right
par(mfrow=c(2,2))
barplot(df1a, beside = TRUE,col=1:8,main="A",ylab="bias")
barplot(df1b, beside = TRUE,col=1:8,main="B",ylab="bias")
barplot(df2a, beside = TRUE,col=1:8,main="C",ylab="rmse")
barplot(df2b, beside = TRUE,col=1:8, main="D",ylab="rmse",legend.text = c("RAW",
   "CEM","MD1", "MD2", "PS1", "PS2", "FUL1", "FUL2"),args.legend = list(x =
   "topright",bty='n', inset=c(-0.4,-0.5), xpd = TRUE))


# Chapter 4
## Simulation from Kang and Shafer (2007).
rm(list=ls())
k=4; MCSim=1000; ATT=0
fun=function(n)
{
set.seed(123456)
tmp <- matrix(NA, MCSim, 2)
colnames(tmp) <- c("Unweighted", "Proposed")
library(MASS)
for (MC in 1:MCSim){
smahal=
function(t,X){
X<-as.matrix(X)
for (j in 1:k) X[,j]<-rank(X[,j])
n<-dim(X)[1]
rownames(X)<-1:n
k<-dim(X)[2]
m<-sum(t)
cv<-cov(X)
vuntied<-var(1:n)
rat<-sqrt(vuntied/diag(cv))
cv<-diag(rat)%*%cv%*%diag(rat)
```

```
out<-matrix(NA,m,n-m)
Xc<-X[t==0,]
Xt<-X[t==1,]
rownames(out)<-rownames(X)[t==1]
colnames(out)<-rownames(X)[t==0]
library(MASS)
icov<-ginv(cv)
for (i in 1:m) {
out[i,]<-mahalanobis(Xc,Xt[i,],icov,inverted=T)}
out
    }
 X <- mvrnorm(n, mu = rep(0, 4), Sigma = diag(4))
z=-X[,1] +X[,2] -X[,3] -X[,4]
prop <- 1 / (1 + exp(-z))
t <- rbinom(n, 1, prop)

dat=data.frame(t,X)
dat2=dat[order(-t),]
dist=smahal(dat2$t, dat2[,2:length(dat2)])
m=sum(t)
min2 = function(x){
nn=c(min(x),which.min(x))
return(nn)
  }
gg = apply(dist,1,min2)
#t(gg)
p=as.vector(t(gg)[,2])
p2=p+m

d=rep(NA,n)
for (i in 1:n)
d[i]=sum(p2==i)
d=d[-(1:m)]
```

```
w1=c(rep(1,m),d)
size=sum(d==0)
w1[which(w1==0)]=1/size
w0=rep(1,n)

ps=glm(t~.,data=dat2,family=binomial(link="logit"))
pred=predict(ps,type="response")
w2=ifelse(dat2$t==1,1,pred/(1-pred))

outcome <- 210 + 27.4*X[,1] + 13.7*X[,2] + 13.7*X[,3] + 13.7*X[,4] + rnorm(n)
dati1 <- data.frame(dat,outcome)
dati2=dati1[order(-t),]; dati2=data.frame(w1,dati2)

RAW <- summary(lm(outcome~t,data=dati1))$coef[2]
MAH <- summary(lm(outcome~t,weights=w1,data=dati2))$coef[2]
tmp[MC,] <- c(RAW,MAH)
}
return(tmp)
}
res1a=fun(200); res1b=fun(500); res1c=fun(1000); res1d=fun(2000)

cat("\n\nAverage absolute bias:\n")

bias1=abs(colMeans(res1a)-ATT);bias2=abs(colMeans(res1b)-ATT)
bias3=abs(colMeans(res1c)-ATT; bias4=abs(colMeans(res1d)-ATT)
bias=c(bias1,bias2,bias3,bias4)

cat("RMSE:\n")
mse1=sqrt(colMeans((res1a-ATT)^2));mse2=sqrt(colMeans((res1b-ATT)^2))
mse3=sqrt(colMeans((res1c-ATT)^2)); mse4=sqrt(colMeans((res1d-ATT)^2))
rmse=c(mse1,mse2,mse3,mse4)
```

```
# Chapter 5
## Obtaining the optimal lamda value for a dataset

rm(list=ls())
k=4; MCSim=100; n=200
library(survey)
library(tableone)
library(causalsens)
library(twang)
library(MASS)

fun=function(lamda)
{
tmp <- rep(NA,MCSim)
set.seed(123456)
for (MC in 1:MCSim)
{
smahal=
function(t,X){
X<-as.matrix(X)
for (j in 1:k) X[,j]<-rank(X[,j])
n<-dim(X)[1]
rownames(X)<-1:n
k<-dim(X)[2]
m<-sum(t)
cv<-cov(X)
vuntied<-var(1:n)
rat<-sqrt(vuntied/diag(cv))
cv<-diag(rat)%*%cv%*%diag(rat)
out<-matrix(NA,m,n-m)
Xc<-X[t==0,]
Xt<-X[t==1,]
rownames(out)<-rownames(X)[t==1]
```

```
colnames(out)<-rownames(X)[t==0]
library(MASS)
icov<-ginv(cv)
for (i in 1:m) {
out[i,]<-mahalanobis(Xc,Xt[i,],icov,inverted=T)}
out
}
X <- mvrnorm(n, mu = rep(0, 4), Sigma = diag(4))
x1=X[,1]; x2=X[,2]; x3=X[,3]; x4=X[,4]

z=-x1 +x2 -x3 -x4
prop <- 1 / (1 + exp(-z))
t <- rbinom(n, 1, prop)

dat=data.frame(t,X)
dat2=dat[order(-t),]
dist=smahal(dat2$t, dat2[,2:length(dat2)])
m=sum(t)

gg2=apply(dist,1,rank)
ggg=t(gg2)
ggg=lamda^(ggg-1)
est=as.vector(apply(ggg,2,sum))

w=c(rep(1,m),est)
 tmp[MC]<- sd(w)/mean(w)
}
return(tmp)
}
lamda=seq(0,1,0.01)
res=sapply(lamda,fun)

cv=colMeans(res)
```

```
df=data.frame(lambda=lamda,cv=cv)
minlamda=subset(df,cv==min(cv))$lambda

plot(lamda,cv,type="b",ylab="Coefficient of variation")
abline(v=minlamda,lty=2)
abline(h=min(cv),lty=3)



# Chapter 6
# Covariate balance diagnostics
rm(list=ls())
library(cobalt)
library(WeightIt)
library(survey)
library(tableone)
library(ggplot2)
library(MASS)
library(ggpubr)

ATT=-0.4
MCSim=1000
k=7
fun=function(n,rate,rate.y)
{
set.seed(123456)
tmp <- matrix(NA, MCSim, 2)
se <- matrix(NA, MCSim, 2)
coverage <- matrix(NA, MCSim, 2)
SB.ebal=matrix(NA,k,MCSim)
SB.ipw=matrix(NA,k,MCSim)

colnames(tmp) <- c("EBAL","IPW")
colnames(se) <- c("EBAL","IPW")
```

```
colnames(coverage) <- c("EBAL","IPW")

 for(MC in 1:MCSim){

sigma<- diag(10)
sigma[1,5]=sigma[5,1]=sigma[8,3]=sigma[3,8]=0.2
sigma[2,6]=sigma[6,2]=sigma[4,9]=sigma[9,4]=0.9

x <- mvrnorm(n, mu = rep(0, 10), Sigma = sigma,empirical=T)
x1<- x[,1]; p1=pnorm(x1);  xx1=rbinom(n,1,p1)
x2<- x[,2]
x3<- x[,3]; p3=pnorm(x3);  xx3=rbinom(n,1,p3)
x4<- x[,4]
x5<- x[,5]; p5=pnorm(x5);  xx5=rbinom(n,1,p5)
x6<- x[,6]; p6=pnorm(x6);  xx6=rbinom(n,1,p6)
x7<- x[,7]
x8<- x[,8]; p8=pnorm(x8);  xx8=rbinom(n,1,p8)
x9<- x[,9]; p9=pnorm(x9);  xx9=rbinom(n,1,p9)
x10<- x[,10]
b0=rate; a0=rate.y
    #=-1.047 for 35%, =0 for 50%, =0.7 for 67%
b1<- 0.8; b2 <- -0.25; b3<- 0.6; b4<- -0.4; b5<- -0.8; b6<- -0.5; b7<- 0.7
a1<- 0.3; a2<- -0.36; a3<- -0.73; a4<- -0.2; a5<- 0.71; a6<- -0.19; a7<- 0.26
mu=b0 + b1*xx1 + b2*x2 + b3*xx3 + b4*x4 + b5*xx5 + b6*xx6 + b7*x7
prop <- exp(mu) / (1 + exp(mu))
t <- rbinom(n, 1, prop)
dat=data.frame(xx1,x2,xx3,x4,xx5,xx6,x7,t)

w.out1=weightit(t~.,estimand="ATT",data=dat,method="ebal")
w.ebal=get.w(w.out1)
w.out2=weightit(t~.,estimand="ATT",data=dat,method="ps")
w.ipw=get.w(w.out2)
```

```
b1=bal.tab(w.out1); b2=bal.tab(w.out2)
SB.ebal[,MC]=b1$Balance[16][,1];   SB.ipw[,MC]=b2$Balance[16][-1,1]

mu2=a0+ I(ATT*t)+ a1*xx1+ a2*x2+ a3*xx3+ a4*x4+ a5*xx8+ a6*xx9+ a7*x10
prop2 <- exp(mu2) / (1 + exp(mu2))
y <- rbinom(n, 1, prop2)
w.ebal_c=w.ebal[t==0]; w.ipw_c=w.ipw[t==0]
y_t=y[t==1]; y_c=y[t==0]
dat2=data.frame(y,w.ebal,w.ipw,dat)
weighteddata1=svydesign(ids=~1,data=dat2,weights=w.ebal)
weighteddata2=svydesign(ids=~1,data=dat2,weights=w.ipw)
ebal=summary(svyglm(y~t,design=weighteddata1,family=quasibinomial))$coef[2]
ipw=summary(svyglm(y~t,design=weighteddata2,family=quasibinomial))$coef[2]
se.ebal=summary(svyglm(y~t,design=weighteddata1,family=quasibinomial))$coef[4]
se.ipw=summary(svyglm(y~t,design=weighteddata2,family=quasibinomial))$coef[4]

c1a=confint(svyglm(y~t,family=quasibinomial,design=weighteddata1))[2]
c2a=confint(svyglm(y~t,family=quasibinomial,design=weighteddata1))[4]
c1b=confint(svyglm(y~t,family=quasibinomial,design=weighteddata2))[2]
c2b=confint(svyglm(y~t,family=quasibinomial,design=weighteddata2))[4]

indx.ebal <- (c1a <= ATT) & (c2a >= ATT); cov.ebal = sum(indx.ebal)
indx.ipw <- (c1b <= ATT) & (c2b >= ATT); cov.ipw = sum(indx.ipw)

tmp[MC,] <- c(ebal,ipw)
se[MC,] <- c(se.ebal,se.ipw)
 coverage[MC,] <- c(cov.ebal,cov.ipw)
  }
return(list("ASMD1"=rowMeans(SB.ebal),"ASMD2"=rowMeans(SB.ipw),"est"=tmp,"se"=se,
     "coverage"=coverage))
}
res1a.ya=fun(300,-1.047,-2.485)
res1a.yb=fun(300,-1.047,0.785)
```

```
res1b.ya=fun(1000,-1.047,-2.485)
res1b.yb=fun(1000,-1.047,0.785)

res2a.ya=fun(300,0,-2.485)
res2a.yb=fun(300,0,0.785)
res2b.ya=fun(1000,0,-2.485)
res2b.yb=fun(1000,0,0.785)

res3a.ya=fun(300,0.7,-2.485)
res3a.yb=fun(300,0.7,0.785)
res3b.ya=fun(1000,0.7,-2.485)
res3b.yb=fun(1000,0.7,0.785)


ASMD1a=cbind(res1a.ya$ASMD1,res1a.ya$ASMD2); ASMD1b=cbind(res1b.ya$ASMD1,res1b.ya$ASMD2)
ASMD2a=cbind(res2a.ya$ASMD1,res2a.ya$ASMD2); ASMD2b=cbind(res2b.ya$ASMD1,res2b.ya$ASMD2)
ASMD3a=cbind(res3a.ya$ASMD1,res3a.ya$ASMD2); ASMD3b=cbind(res3b.ya$ASMD1,res3b.ya$ASMD2)

colnames(ASMD1a)=colnames(ASMD1b)=colnames(ASMD2a)=colnames(ASMD2b)=
        colnames(ASMD3a)=colnames(ASMD3b)=c("EBAL","IPW")
par(mfrow=c(2,3))
##boxplots of n=300
boxplot(ASMD1a,ylab="ASMD",main="(A) N=300, Prevalence=33%")
boxplot(ASMD2a,ylab="ASMD",main="(B) N=300, Prevalence=50%")
boxplot(ASMD3a,ylab="ASMD",main="(C) N=300, Prevalence=67%")
##boxplots of n=1000
boxplot(ASMD1b,ylab="ASMD",main="(D) N=1000, Prevalence=33%")
boxplot(ASMD2b,ylab="ASMD",main="(E) N=1000, Prevalence=50%")
boxplot(ASMD3b,ylab="ASMD",main="(F) N=1000, Prevalence=67%")

windows()
```