ABSTRACT

ZHANG, YUNSHU. Doubly Robust Estimators of Causal Effects in Observational Studies: Theory and Practice. (Under the direction of Dr. Shu Yang).

Understanding causality has been a fundamental challenge throughout human history. In observational studies, the assignment of treatments is not random but rather determined by physicians or patients. Consequently, the treatment group often exhibits distinct characteristic distributions compared to the control group, posing a significant obstacle in estimating causal effects. In the absence of unmeasured confounders, statisticians have devised various estimators for the average treatment effect (ATE) in observational studies, such as inverse probability weighting, propensity score matching, and outcome regression estimators. However, these estimators are singly robust, meaning they can be substantially biased if the corresponding model is misspecified. To address this limitation, researchers have endeavored to develop doubly robust estimators that leverage both treatment and outcome information, thereby enhancing robustness. While the augmented inverse probability weighting estimator has been extensively studied, there is still limited development surrounding the doubly robust matching and regression estimators.

In this dissertation, our focus is on doubly robust estimators in observational studies, encompassing both theoretical developments and practical recommendations. In Chapter 2, we introduce a novel approach called double score matching (DSM), which combines the propensity score and prognostic score. To account for potential model misspecification, we consider multiple candidate models for each score. We demonstrate that the de-biased DSM estimator exhibits multiple robustness, as it remains consistent if any one of the score models is correctly specified. We characterize the asymptotic distribution of the DSM estimator, requiring only one correctly specified model. Additionally, we present a two-stage replication method for variance estimation and extend DSM for quantile estimation. Simulation results illustrate that DSM outperforms single score matching and prevailing multiply robust weighting estimators, particularly in scenarios involving extreme propensity scores.

In Chapter 3, we shift our focus to the practical aspects of the matching estimator. Through extensive simulations, we investigate the statistical and numerical properties of PSM, PGM, and DSM. Our findings support the superior performance of DSM, either matching or surpassing the two single score matching methods in terms of bias and variance. Notably, DSM achieves double robustness by only requiring either the correctly specified

propensity score model or prognostic score model for consistency. We recommend variable selection for the propensity score model and matching with replacement for DSM, supported by comprehensive simulation studies.

In Chapter 4, we transition from matching estimators to the doubly robust regression estimator. We present the framework of the double robust (DR) regression estimator for estimating the ATE. The DR regression estimator exhibits double robustness, as it remains consistent for the ATE even if either the propensity score or the outcome model is correctly specified, though not necessarily both. Notably, we establish, for the first time in the literature, the asymptotic distribution of the DR regression estimator by deriving its influence function (IF). The derived IF allows for misspecification in either the parametric component or the propensity score and shows that the estimator achieves the semiparametric efficiency bound when both models are correctly specified.

In Chapter 5, we aim to address the potential limitations of this assumption by considering the sensitivity analysis framework under informative sampling. We propose the semiparametric efficient estimator for the ATE by deriving the efficient influence function (EIF) based on semiparametric theory. A logistic selection specification is utilized to model the impact of unmeasured confounders. We obtain its asymptotic distribution and construct the corresponding variance estimator. Recognizing that the EIF-based estimator requires knowledge of the true sensitivity parameters, we propose an approach to calibrate these parameters. Finally, we apply our estimator to the Per- and polyfluoroalkyl substances (PFAS) dataset to investigate the health effects of PFAS.

Doubly Robust Estimators of Causal Effects in Observational Studies: Theory and Practice

by
Yunshu Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2023

APPROVED BY:

Dr. Wenbin Lu	Dr. Marie Davadian
Dr. Luo Xiao	Dr. Wensong Yu

Dr. Shu Yang
Chair of Advisory Committee

BIOGRAPHY

The author, born in the vibrant city of Beijing, China, began his academic journey at RDFZ - The High School Affiliated to Renmin University of China, where he completed his high school education. Showing exceptional promise in his studies, he went on to pursue higher education at the prestigious Tsinghua University, where he earned a bachelor's degree in mathematics and physics.

Driven by a passion for data analysis and quantitative research, he set his sights on furthering his knowledge in statistics. With a determined spirit, he embarked on an ambitious quest for knowledge and was admitted to North Carolina State University to pursue a Ph.D. degree in statistics.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to all those who have contributed to the completion of this dissertation. Their support, guidance, and encouragement have been invaluable throughout this academic journey.

First and foremost, I would like to extend my deepest gratitude to my supervisor, Dr. Shu Yang, for her unwavering support, insightful guidance, and constructive feedback. Her expertise and dedication have been instrumental in shaping the direction and quality of a researcher.

I am also grateful to the members of my dissertation committee, Dr. Wenbin Lu, Dr. Marie Davidian, and Dr. Luo Xiao, for their valuable insights, suggestions, and critical evaluation of my work. Their expertise and diverse perspectives have enriched the depth and breadth of this study.

I would like to acknowledge the assistance and cooperation of Dr. Brian Reich from NCSU, Wenyu Ye, Ilya Lipkovich, Douglas Faries, and Zbigniew Kadziola from Eli Lilly and Company, without whom this research would not have been possible. Their willingness to participate and provide valuable data and insights have been integral to the success of this study.

I am thankful to my friends for their support, especially my roommates Xianpeng Liu and Yutian Cai as well as my classmates Runbing Zheng, Nicolas Larsen, Joe Zhao, and Will Edwards. Their presence has made the academic environment and my life enjoyable and inspiring.

Lastly, I am deeply grateful to my family for their unwavering love, understanding, and encouragement. Their constant support and belief in my abilities have been the driving force behind my perseverance and dedication to finish the degree.

To all those who have contributed in various ways, whether big or small, I offer my heartfelt thanks. Your contributions have played an essential role in shaping this dissertation and my personal growth as a researcher.

Thank you all for your support and belief in me throughout this five-year journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Understanding causality has been a key problem throughout human history. People from different science disciplines are interested in discovering causal effects in the real world and utilizing them to make values. However, causal inference was not formalized until the introduction of the potential outcome framework by Rubin (1974). He defined the causal effect as the difference between the potential outcomes of different treatment levels. The potential outcome is the outcome of interest had the subject been given one specific treatment, which may not be same as the actual treatment the subject receives. Usually, we are interested in the causal effect of a treatment for a specific group of people. Thus, a common goal in causal inference is to estimate the average causal effect (ATE). The key challenge is that each subject can only receive one treatment, resulting in only one of the potential outcomes is observed and the other is not obtainable (Rubin 1978a). This is why causal inference can also be viewed as a missing data problem. The key to this problem is randomization. In randomized experiments, the treatment is independent with the potential outcomes. Under the Stable Unit Treatment Value Assumption (SUTVA) proposed by Rubin (1980), which assumes that the outcome observed for a subject will be the same as the corresponding potential outcome regardless of the other conditions and the treatment assignments of other individuals, the ATE can be estimated by the difference in means from

the dataset. A series of works have studied the inference of this estimator under different scenarios, for example, Fisher's exact test based on permutations of finite samples and Neyman's approach under repeated sampling (Fisher 1936; Neyman 1923).

However, the naive estimator only works in randomized experiments but not in observational studies, where treatments are not randomly assigned but determined by physicians or patients. The fundamental reason is that the treatment group has very different characteristic distributions from the control group, which is the core challenge in observational studies. One may argue to stick to randomized experiments to avoid this problem. However, it is much easier to obtain large-scale observational data in this big-data era than expensive experimental data. Thus, it is necessary to develop estimators for the treatment effects in observational studies under reasonable assumptions. A fundamental assumption is the no unmeasured confounders assumption, or equivalently, the strong ignorability assumption (Rubin 1978b; Rosenbaum and Rubin 983a; Rubin 1990). A confounder is defined as a pre-treatment characteristic that is associated with both the treatment assignments and potential outcomes. We assume all the confounders have been collected in the set of observed covariates. This leads to the conditional independence between the treatment assignments and potential outcomes. Another critical assumption is called the positivity assumption. It assumes that the treatment assignment probability, which is referred to as the propensity score is strictly between 0 and 1 almost surely. This implies that for a subject given any pre-treatment characteristics and treatment assignment, it is possible to find another individual with the same characteristics but the different treatment option. Under these two assumptions and SUTVA, statisticians have developed various estimators for the ATE in observational studies. Among them, weighting, matching, and regression estimators are widely used in practice.

Weighting estimators assign different weights to subjects to achieve unbiasedness. The idea of inverse probability weighting (IPW) estimator can be traced back to 1952 (Horvitz and Thompson 1952). It is practically appealing by its simple construction, where the weights are constructed by inverting the propensity scores. The propensity scores need to be estimated because they are always unknown in observational studies. Usually, people fit a logistic or probit model to obtain the estimate. However, the IPW estimator may be significantly biased if the propensity score model is misspecified. A common choice is augmented inverse probability weighting (AIPW) (Robins et al. 1994), where the outcome information is incorporated into the IPW estimator to increase efficiency. These conditional expectations need to be estimated via, for example, a generalized linear model on covariates and treatment or two independent models for the two treatment arms individually. Both

parametric and nonparametric methods can be used to model these relationships. Among all the existing estimators for the ATE, the AIPW estimator is the most popular because it is doubly robust, i.e., the estimator is consistent if either one of the propensity scores or outcome model is correctly specified (Scharfstein et al. 1999). This implies more reliability of the AIPW estimator. Moreover, the estimator achieves the so-called semiparametric efficiency bound (Tsiatis 2006), which is the minimum asymptotic variance that semiparametric estimators can achieve. Equivalently, the AIPW estimator is called a semiparametric efficient estimator because it reaches the semiparametric efficiency bound and thus its asymptotic variance must be no larger than that of any semiparametric estimator. However, weighting estimators may be unstable when propensity scores are close to zero or one due to their constructions (Kang and Schafer 2007; Guo and Fraser 2014). Also, even if the AIPW estimator is doubly robust, it may still be inconsistent if both models are misspecified. Thus, machine learning algorithms or ensemble methods, such as Super Learner (Van der Laan et al. 2007), may be used in model fitting to increase robustness further. Nevertheless, a significant drawback is that inference is primarily unavailable for machine learning methods. Van der Laan et al. (2007) proposed the targeted maximum likelihood learning estimator (TMLE) that uses the efficient influence function, which is the influence function of the semiparametric efficient estimator, to make the inference.

On the other hand, matching estimators as a hot dock imputation method is invulnerable to extreme propensity scores (Frölich 2004). The basic idea is to impute the missing potential outcome for each subject based on the most similar individual in the other treatment group. Thus, a distance metric is needed to define similarity between subjects, for example, the Mahalanobis distance based on observed pre-treatment covariates. The simplest setting is the one-to-one matching with replacement. The estimator can be extended by matching one unit with multiple units or matching without replacement, i.e., each subject can only be matched at most once. The theoretical and practical implications of these extensions have been discussed by researchers in recent years (Lin et al. 2021). Because the missing outcomes are all imputed via observed outcomes, the estimator will not be unstable even if the propensity scores are extreme. However, matching may suffer from the curse of dimensionality, and thus researchers choose to match on some reasonable summary statistics instead of matching on high-dimensional covariates directly. Rosenbaum and Rubin (1983b) proposed the popular propensity score matching (PSM) estimator by matching units with similar propensity scores, and Hansen (2008) proposed an alternative called the prognostic score matching (PGM) estimator by utilizing the outcome information. Unfortunately, both PSM and PGM are only singly robust, making them vulnerable

in practice. Similar to AIPW, researchers have tried to combine the two scores into the matching estimator to increase robustness (Leacy and Stuart 2014; Antonelli et al. 2018). However, only numerical properties and basic consistency results have been developed, and no asymptotic theories are establish for the doubly robust matching estimator. As a result, it is unclear how to make inference using this estimator,and whether the matching estimator can achieve the semiparametric efficiency bound.

Compared to the weighting and matching estimators, the idea of the outcome regression estimator is more straightforward. It applies regression models to estimate the missing potential outcomes. Outcome is regressed on the covariates and treatment assignment to estimate the conditional expectations, and the choice of the model may depend on the types of variables. Alternatively, two models can be trained independently from the treatment and control groups. They may not be in the same category if researchers believe the underlying data-generating processes for the two groups are different. Nevertheless, the conditional expectations are estimated from the regression models, and the outcome regression estimator for the ATE is the difference in means of the estimated potential outcomes. The outcome regression estimator is easy to implement and can be very efficient if the regression model is correct. However, the result may be significantly biased if the outcome model is misspecified. To improve the robustness of regression estimators, researchers have been working on incorporating the treatment information, i.e., the propensity score, into the outcome regression model. Doubly robust regression estimator, also named the penalized spline of propensity methods for treatment comparison (PENCOMP), has been proposed by Zhou et al. (2019) for missing data imputation and estimating treatment effects in observational studies. It assumes that the outcome follows a semiparametric partially linear regression model with a parametric component of individual covariates and a nonparametric component of the propensity score. The estimator is doubly robust to model misspecification of the parametric component or the propensity score and has good numerical performance; however, its theoretical properties are not fully established due to the challenges of the estimated propensity score as the regressor, so-called the “generated regressor” problem.

All of the aforementioned estimators, including weighting, matching, and regression estimators, rely on the assumption of no unmeasured confounders to ensure their consistency. However, verifying this assumption using observed data is impossible, and it is a fundamental concern when estimating causal effects in observational studies. This is precisely why randomized trials are considered the gold standard by the U.S. Food and Drug Administration (FDA) for testing medical effects (Ciociola et al. 2014). One approach

to address this issue is to expand the dimension of covariates by collecting more variables, which has become easier and more convenient in the era of big data. However, this high-dimensional scenario may introduce new potential problems, as demonstrated by the challenges encountered in the matching estimator with covariates. Therefore, dimension reduction techniques are necessary, and variable selection methods can enhance numerical efficiency even if the estimators are constructed based on propensity scores or outcomes. For instance, Brookhart et al. (2006) used simulations to suggest using covariates related to the outcome and avoiding the use of instrumental variables (IVs) when constructing the propensity score model for propensity score weighting. Similarly, Myers et al. (2011) reported similar findings in PSM, prioritizing the minimization of unmeasured confounding when selecting variables, even at the risk of conditioning on IVs. While most existing recommendations are focused on PSM, few studies have investigated these issues for propensity-based generalized methods and doubly robust matching estimators.

Another approach to address the strong ignorability assumption is through sensitivity analysis, which evaluates the impact of unmeasured confounders (Liu et al. 2013). One method involves bounding the influence of an unmeasured confounder by assuming that the odds ratio of treatment assignment can only change within a predetermined range. In other words, the minimum change required for the inference about the ATE to shift from statistically significant to insignificant is determined (Rosenbaum 2002). Another commonly used approach is to introduce sensitivity parameters that model the relationship between unmeasured confounders, treatment, and outcomes. For instance, a logistic selection specification can be employed to capture the conditional distributions of the outcome under different treatment groups (Franks et al. 2020). By assuming a parametric sensitivity framework, the semiparametrically efficient estimator can be derived using the corresponding efficient influence function (Scharfstein et al. 2021). It is worth noting that sensitivity analysis is closely tied to informative sampling and selection bias, where the sampling probability may depend on other variables even after accounting for observed covariates (Robins et al. 2000). However, to the best of our knowledge, there is currently no statistical framework that effectively handles both unmeasured confounders and informative sampling simultaneously.

In this dissertation, our focus is on doubly robust estimators in observational studies, encompassing both theoretical developments and practical recommendations. Our primary emphasis is on the development of the doubly robust matching estimator, extending beyond the well-studied doubly robust weighting estimator AIPW. In Chapter 2, we introduce a novel approach called double score matching (DSM), which combines the

propensity score and prognostic score. To account for potential model misspecification, we consider multiple candidate models for each score. We demonstrate that the de-biased DSM estimator exhibits multiple robustness, as it remains consistent if any one of the score models is correctly specified. We characterize the asymptotic distribution of the DSM estimator, requiring only one correctly specified model, by utilizing martingale representations of the matching estimators and theory for local Normal experiments. Additionally, we present a two-stage replication method for variance estimation and extend DSM for quantile estimation. Simulation results illustrate that DSM outperforms single score matching and prevailing multiply robust weighting estimators, particularly in scenarios involving extreme propensity scores.

In Chapter 3, we shift our focus to the practical aspects of the matching estimator. Through extensive simulations, we investigate the statistical and numerical properties of PSM, PGM, and DSM. Our findings support the superior performance of DSM, either matching or surpassing the two single score matching methods in terms of bias and variance. Notably, DSM achieves double robustness by only requiring either the correctly specified propensity score model or prognostic score model for consistency. We recommend variable selection for the propensity score model and matching with replacement for DSM, supported by comprehensive simulation studies.

In Chapter 4, we transition from matching estimators to the doubly robust regression estimator. We present the framework of the double robust (DR) regression estimator for estimating the ATE. The DR regression estimator exhibits double robustness, as it remains consistent for the ATE even if either the propensity score or the outcome model is correctly specified, though not necessarily both. Notably, we establish, for the first time in the literature, the asymptotic distribution of the DR regression estimator by deriving its influence function (IF). The derived IF allows for misspecification in either the parametric component or the propensity score and shows that the estimator achieves the semiparametric efficiency bound when both models are correctly specified. To construct valid variance estimators and confidence intervals, we propose an adaptive bootstrap method based on Newey (1997). Simulation studies reveal that the DR regression estimator may exhibit superior finite sample performance compared to other state-of-the-art estimators, including the augmented inverse probability weighting estimator and the target maximum likelihood estimator, particularly in scenarios with extreme propensity score distributions.

All of the aforementioned estimators rely on the strong ignorability assumption. However, in Chapter 5, we aim to address the potential limitations of this assumption by considering the sensitivity analysis framework under informative sampling. This motivation stems

from the dataset on Per- and polyfluoroalkyl substances (PFAS), a group of approximately 5,000 synthetic chemicals known to be resistant to heat, water, and oil, posing potential hazards to human health. Given the informative nature of the sampling process for measuring PFAS, where locations with higher concentrations are more likely to be selected, it becomes crucial to account for this sampling bias. Furthermore, unmeasured confounders between PFAS exposure and the health outcomes of interest may exist. Consequently, we adopt multiple sensitivity analysis models to accurately estimate the causal effect. Within this framework, we derive the semiparametric efficient estimator for the ATE, building upon the semiparametric theory. Inspired by the work of Franks et al. (2020); Scharfstein et al. (2021), we utilize a logistic selection specification to model the impact of unmeasured confounders. We establish the identification formula for the marginal mean and derive the efficient influence function (EIF) based on semiparametric theory. The semiparametric efficient estimator of the ATE is then constructed using the EIF, along with its asymptotic distribution and variance estimator. Recognizing that the EIF-based estimator requires knowledge of the true sensitivity parameters, we propose an approach to calibrate these parameters, building upon the work of Franks et al. (2020) with a modification to address an error. To assess the performance of the EIF-based estimator, we conduct a simulation study comparing it to other existing estimators. Finally, we apply our estimator to the PFAS dataset to investigate the health effects of PFAS. In Chapter 6, we summarize our findings and discuss about the potential future work on the doubly robust estimators in observational studies.

CHAPTER

2

MULTIPLY ROBUST MATCHING ESTIMATORS OF AVERAGE AND QUANTILE TREATMENT EFFECTS

2.1 Introduction

Causal inference plays an important role in science, education, medicine, policy, and economics. If all confounders of the treatment-outcome relationship are observed, one can use standard techniques, such as regression adjustment, inverse probability of treatment weighting (IPW), augmented IPW (AIPW), and matching to adjust for confounding (Imbens and Rubin 2015). Among them, the AIPW estimator is most popular because it achieves the so-called *double robustness* property by combining the use of models for the probability of treatment assignment, also known as the propensity score (Rosenbaum and Rubin 1983b), and the outcome mean function. More specifically, it consistently estimates the treatment effect if either one of these functions is modeled correctly (e.g., Lunceford and Davidian 2004; Bang and Robins 2005). However, inevitably, weighting estimators can have a high

variability by inverting the estimated propensity scores (e.g., Kang and Schafer 2007; Guo and Fraser 2014), especially if these probabilities are close to zero or one. Matching has multiple features that are desirable:

a) Matching does not involve weighting by the inverse of the propensity score and therefore avoids the possibly large variability due to weighting (Frölich 2004);

b) Matching is transparent and intuitively appealing with the goal of replicating a randomized experiment from an observational study (Rosenbaum 1989; Heckman et al. 1997a; Dehejia and Wahba 2002, 1999; Rubin 2006; Stuart 2010);

c) Matching can be viewed as a hot deck imputation method that can provide valid estimators of general parameters depending on the entire distribution, such as quantiles (Ford 1983).

Although matching has a substantial promise, it suffers from the issue of the curse of dimensionality. In the presence of many covariates, matching directly on high-dimensional covariates is incapable of removing all confounding biases. To overcome this challenge, researchers have proposed different dimension reduction techniques to facilitate matching. On the one hand, Rosenbaum and Rubin (1983b) demonstrated the central role of the propensity score as being a balancing score in the sense that the same propensity score distributions in different treatment groups lead to the same covariate distributions. Therefore, propensity score matching (PSM) can remove all confounding biases (e.g., Abadie and Imbens 2016). However, PSM has recently received major backlash for emulating an unconditional randomized experiment, contrary to matching on covariates that mimics a block randomized experiment (King and Nielsen 2019). On the other hand, Hansen (2008) proposed an alternative balancing score: the prognostic score, also called the disease risk score (i.e., a sufficient statistic for the potential outcomes given which the potential outcomes and covariates are independent). This score provides a balance of disease risks between the treatment groups, as distinct from the balance of treatment propensities provided by the propensity score. In economics, prognostic score matching (PGM) has been previously proposed in Imbens (2004) and Zhao (2004), where the prognostic score is a vector of linear predictors in treatment-specific outcome regressions. PGM is also similar to predictive mean matching (Yang and Kim 2018, 2019) in the missing data literature to compensate for nonresponse. As analogous to AIPW, it is advantageous to combine the use of the propensity and prognostic score in matching (Hansen 2008). Leacy and Stuart (2014) showed empirically that the joint use of two scores in matching (which we refer to as double

score matching, DSM) improves the treatment effect estimation. Antonelli et al. (2018) later established the double robustness of matching jointly on propensity and prognostic scores in the sense that the matching estimator is consistent for the ATE if either one of the score models is correctly specified. However, they only provided the rate of convergence but not the asymptotic distribution of the doubly robust matching estimator.

In this article, we propose new DSM estimators based on the propensity score and the prognostic score. Because each score creates a balance between the treated and control groups, the augmented score serves as a "double balancing score". To estimate the ATEs, existing DSM would require adjusting for the vector of the propensity score and possibly multiple treatment-specific prognostic scores (i.e., one for each treatment group). Instead of estimating the ATEs directly, we focus on estimating the average of the potential outcomes separately for each treatment level, which requires adjusting only for the propensity score and the prognostic score for that particular level of the treatment. This insight allows us to reduce the dimension of the double score further without giving up the double balancing property. This strategy also plays an important role in dimension reduction for constructing improved DSM estimators.

In practice, the double score is unknown and therefore requires modeling and estimation. The new DSM estimator is doubly robust, which includes one propensity score model and one prognostic score model. With an unknown data generating process, there is no guarantee that either of the two models is correctly specified. To gain additional protection against model misspecification, we posit multiple models for the propensity score and prognostic score. Doing so, however, may introduce bias due to matching discrepancy based on a moderately high-dimensional matching variable (Abadie and Imbens 2011), although our strategy of estimating the average of potential outcomes separately helps dimension reduction. In this case, we propose the de-biasing DSM estimator that corrects for the bias due to matching discrepancy. The current matching literature has focused primarily on estimating the ATEs; however, other aspects of the distribution such as quantiles may be more appropriate in certain applications. For example, a treatment strategy may not decrease average health cost but instead lowers the upper tail of the cost distribution, so focusing only on ATEs would not reveal the beneficial effect of the treatment strategy. In these cases, it is more informative to study quantile treatment effects (QTEs), which are defined as the differences in population quantiles of the potential outcome distributions. Taking the advantage of matching as a hot deck imputation method, we extend the multiply robust DSM framework to estimate QTEs.

We show that the DSM estimators have the *multiple robustness* property, which guar-

antees the estimation consistency if any one of the candidate models for the propensity score or prognostic score is correctly specified. This result is similar in essence to the multiply robust weighting (MRW) estimators in the missing data and survey literature (Han and Wang 2013; Han 2014; Han et al. 2019; Li et al. 2020; Chen and Haziza 2017b,a)). Naik et al. (2016) and Wang (2019) proposed MRW estimators for estimating the ATEs in causal inference. The DSM estimator has an intrinsic connection with the doubly-robust AIPW estimator in that they exhibit similar asymptotic expansions; see (2.6) and (2.7). However, AIPW is extremely unstable when some estimated propensity scores are close to zero or one due to the weight construction, and it is sensitive to slight model misspecification (Kang and Schafer 2007). Instead of inverting the estimated propensity score, the DSM estimator uses the matching weights and hence is more robust to extreme propensity score values. MRW also mitigates the instability of (A)IPW by estimating the weights directly under balance constraints. However, it requires exact covariate balance and thus can still have large variability in finite samples compared to DSM, especially when the overlap of the covariate distribution is limited.

The rest of this paper proceeds as follows. Section 2.2 introduces notation, assumptions, and lemmas for various balancing scores. Section 2.3 provides the new perspective of using the double score as a dimension reduction tool and proposes the new DSM estimator of the ATE with multiple candidate models for the double score. Section 2.4 establishes the multiple robustness of the DSM estimator and its limiting distribution, which allows quantifying the impact of the nuisance parameter estimation. Section 2.5 extends the DSM framework to the estimation of the QTE. Section 2.6 provides the extensions to the average treatment effect on the treated (ATT) and quantile treatment effect on the treated (QTT). Section 2.7 uses simulation to evaluate the finite-sample properties of the DSM estimators. The simulation results demonstrate that matching estimators outperform weighting estimators. Section 2.8 applies the DSM estimators to an observational study from the job training program. Section 2.9 concludes, the supplementary material contains the proofs and additional empirical results, and an R package *dsmatch* is available at https://github.com/Yunshu7/dsmatch.

2.2 Notation, assumptions, and balancing scores

Let X_i be a vector of pre-treatment covariates, A_i the binary treatment, and Y_i the outcome for unit $i = 1, \ldots, n$. We follow the potential outcomes framework. Let $Y_i(a)$ be the potential outcome had unit i been given treatment a ($a = 0, 1$). The observed outcome is $Y_i = Y_i(A_i) =$

$A_i Y_i(1)+(1-A_i)Y_i(0)$. We assume that $\{X_i, A_i, Y_i(0), Y_i(1)\}$, $i=1,\ldots,n$, are independent and identically distributed. Thus, (X_i, A_i, Y_i), $i=1,\ldots,n$, are also independent and identically distributed. Various causal estimands are useful to provide a comprehensive assessment of treatment effects. The ATE is $\tau = \mathbb{E}\{Y(1)-Y(0)\}$. For $\xi \in (0,1)$, the overall ξ-QTE is $\Delta_\xi = q_{1,\xi} - q_{0,\xi}$, where $q_{a,\xi} = \inf_q[\mathbb{P}\{Y(a)\le q\}\ge \xi]$, $a=0,1$. When the outcome data follow a skewed distribution, QTEs may be more informative measures of treatment effect. Similarly, the ATT is $\tau_{\text{ATT}} = \mathbb{E}\{Y(1)-Y(0)\mid A=1\}$, and the QTT is $\Delta_{\text{QTT},\xi} = q_{1,\xi|A=1} - q_{0,\xi|A=1}$, where $q_{a,\xi|A=1} = \inf_q[\mathbb{P}\{Y(a)\le q\}\ge \xi \mid A=1]$, $a=0,1$. We illustrate the methodology development for estimating τ and Δ_ξ and provide extensions to the ATT and QTT in Section 2.6. For simplicity of exposition, for a generic variable V, denote

$$\mu_a(V) = \mathbb{E}\{Y(a)\mid V\},\ \sigma_a^2(V) = \mathbb{V}\{Y(a)\mid V\},\ e(V) = \mathbb{P}(A=1\mid V),$$

where $\mu_a(V)$ is an outcome mean function, $\sigma_a^2(V)$ is a variance function, and $e(V)$ is the propensity score.

Hansen (2008) introduced the notion of the prognostic score $\Psi_a(X)$ as a sufficient statistic for $Y(a)$ in the sense that $Y(a) \perp\!\!\!\perp X \mid \Psi_a(X)$ for $a=0,1$. For example, if $Y(a)$ follows a location-shift family $f_a\{y-\mu_a(X)\}$, then $\Psi_a(X)=\mu_a(X)$ for $a=0,1$.

We focus on the setting where the standard positivity and treatment ignorability assumptions hold (Rosenbaum and Rubin 1983b).

Assumption 1 *(i) There exist constants c_1 and c_2 such that $0<c_1\le e(X)\le c_2<1$ almost surely; and (ii) $\{Y(0),Y(1)\}\perp\!\!\!\perp A\mid X$, where $\perp\!\!\!\perp$ means "independent of".*

Assumption 1 (i) implies a sufficient overlap of the covariate distribution between the treatment groups. If this assumption is violated, a common approach is to trim the sample; see Yang and Ding (2018). Generally, Assumption 1 (ii) can be made plausible by collecting detailed information on characteristics of the units that are related to treatment assignment and outcome. As a result, the dimension of X may be high. Balancing scores have been proposed for dimension reduction.

Lemma 1 (Balancing score; Antonelli et al, 2018) *Under Assumption 1, i) $\{Y(0),\ Y(1)\}\perp\!\!\!\perp A\mid\{e(X),h(X)\}$, and ii) $\{Y(0),\ Y(1)\}\perp\!\!\!\perp A\mid\{h(X),\Psi(X)\}$ for any $h(X)$.*

In particular, if $h(X)$ is a null set, Lemma 1 i) shows that the propensity score is a balancing score (Rosenbaum and Rubin 1983b) and ii) shows that the prognostic score is a

balancing score (Hansen 2008). Antonelli et al. (2018) proposed matching based on the estimated double score $\{\widehat{e}(X), \widehat{\Psi}(X)\}$. If either $\widehat{e}(X)$ or $\widehat{\Psi}(X)$ is consistent for the corresponding score, the matching estimator is consistent for the ATE by either i) or ii) in Lemma 1.

DSM is an attractive alternative to PSM; however, the dimension reduction property of Lemma 1 depends on the dimension of the prognostic score. The problem is that if the dimension of the matching variable is higher, the bias order of the matching estimator becomes larger, see Section 2.3.1, suggesting that the advantages of PSM do not carry over to DSM. To preserve the simplicity of matching (avoiding de-biasing), we show that further improvement of the dimension reduction property of the double score is possible without additional assumptions. Then, the advantage of PSM carries over to DSM, see Section 2.3.2. Moreover, because the double score is unknown in practice, one must posit models and estimate the double score from the observed data. To gain robustness to model misspecification, we posit multiple candidate models for the double score. We propose a multiply robust DSM procedure and show that the de-biasing matching estimator achieves multiple robustness, see Section 2.3.3.

2.3 DSM estimators of the ATE

2.3.1 General matching estimators

To fix ideas, we consider matching with replacement with the number of matches fixed at M. Matching estimators hinge on imputing the missing potential outcome for each unit. In practice, the most common choice of M is 1, then the matching procedure becomes nearest neighbor imputation (Little and Rubin 2002; Chen and Shao 2000, 2001). To be precise, for unit i, the potential outcome under A_i is the observed outcome Y_i; the (counterfactual) potential outcome under $1-A_i$ is not observed but can be imputed by the observed outcomes of the nearest M units with $1-A_i$.

To illustrate the properties of the matching estimator, we first consider a generic variable V as the matching variable. Table 2.1 summarizes the choices of V. To stabilize the numerical performance, it is desirable to standardize V such that each component has mean zero and variance one. Without loss of generality, we use the Euclidean distance to determine neighbors; the discussion applies to other distances (Abadie and Imbens 2006). We denote $\mathcal{J}_{V,i}$ as the index set for these matched units for unit i and $K_{V,i} = \sum_{l=1}^{n} \mathbf{1}(i \in \mathcal{J}_{V,l})$ as the number of times that unit i is used as a match, where the subscript "V" in $\mathcal{J}_{V,i}$ and $K_{V,i}$ indicates the name of the matching variable. Table 2.1 (I) illustrates the above

matching scheme to impute the missing potential outcomes. For unit i with $A_i = 1$, the imputed potential outcomes are $\hat{Y}_i(1) = Y_i$ and $\hat{Y}_i(0) = M^{-1}\sum_{j\in\mathscr{J}_{V,i}} Y_j$. For unit i' with $A_{i'} = 0$, $\hat{Y}_{i'}(1) = M^{-1}\sum_{j\in\mathscr{J}_{V,i'}} Y_j$ and $\hat{Y}_{i'}(0) = Y_{i'}$. Once we approximate both potential outcomes for all units, a simple matching estimator of τ is

$$\hat{\tau}_{\text{mat}} = n^{-1}\sum_{i=1}^{n}\{\hat{Y}_i(1) - \hat{Y}_i(0)\} = n^{-1}\sum_{i=1}^{n}(2A_i - 1)(1 + M^{-1}K_{V,i})Y_i.$$

To establish the asymptotic properties of $\hat{\tau}_{\text{mat}}$, Abadie and Imbens (2006) derived the following decomposition

$$n^{1/2}(\hat{\tau}_{\text{mat}} - \tau) = B_n + C_n,$$

where

$$\begin{aligned} B_n &= n^{-1/2}\sum_{i=1}^{n}(2A_i - 1)\left[M^{-1}\sum_{j\in\mathscr{J}_{V,i}}\left\{\mu_{1-A_i}(V_i) - \mu_{1-A_i}(V_j)\right\}\right], \qquad (2.1)\\ C_n &= n^{-1/2}\sum_{i=1}^{n}\left[\mu_1(V_i) - \mu_0(V_i) - \tau + (2A_i - 1)\left(1 + M^{-1}K_{V,i}\right)\left\{Y_i - \mu_{A_i}(V_i)\right\}\right]. \end{aligned}$$

By Assumption 1, for V to be X, propensity score, prognostic score or double score, we have $\mathbb{E}\{\mu_1(V) - \mu_0(V)\} = \tau$, and therefore $\mathbb{E}(C_n) = 0$. The difference $\mu_{A_i}(V_i) - \mu_{A_i}(V_j)$ in (2.1) accounts for the matching discrepancy, so B_n contributes to the asymptotic bias of the matching estimator. In general, if the matching variable is d_V-dimensional, we have $\mathbb{E}(B_n) = O(n^{1/2-2/d_V})$ (Abadie and Imbens 2006, Theorem 1). Table 2.2 demonstrates the relationship of the bias order and d_V. If $d_V \geq 4$, the bias is non-negligible. If $d_V = 3$, the bias shrinks to zero as n increases but the convergence rate $-1/6$ is slow. If $d_V = 2$ and 1, the bias shrinks to zero at much faster rates $-1/2$ and $-3/2$, respectively. Therefore, in finite samples, matching based on a 3-dimensional double score $\{e(X), \Psi(X)\}$ is likely to have a noticeable bias. Reducing d_V to 2 or 1 is worthwhile to make the bias achieve faster rates of converging to zero.

2.3.2 New simple DSM estimator

Lemma 2 is the key result (Antonelli et al. 2018).

Lemma 2 *Under Assumption 1,* $Y(a) \perp\!\!\!\perp A \mid \{h(X), \Psi_a(X)\}$, $Y(a) \perp\!\!\!\perp A \mid \{e(X), h(X)\}$ *for any* $h(X)$ *and* $a = 0, 1$.

Table 2.1: Two matching schemes for imputing potential outcomes. $\mathscr{J}_{V,i}$ denotes the index set for the matched units for unit i, where the subscript "V" represents the name of the matching variable. In (I), the matching variable V is the same for imputing the missing values of $Y(0)$ and $Y(1)$. In (II), the matching variables V_0 and V_1 are different for imputing the missing values of $Y(0)$ and $Y(1)$.

(I) Matching imputation					(II) New matching imputation				
Unit	A	Y	$\hat{Y}(0)$	$\hat{Y}(1)$	Unit	A	Y	$\hat{Y}(0)$	$\hat{Y}(1)$
i	0	Y_i	Y_i	$M^{-1}\sum_{l\in\mathscr{J}_{\mathbf{V},i}} Y_l$	i	0	Y_i	Y_i	$M^{-1}\sum_{l\in\mathscr{J}_{\mathbf{V_1},i}} Y_l$
i'	1	$Y_{i'}$	$M^{-1}\sum_{l\in\mathscr{J}_{\mathbf{V},i'}} Y_l$	$Y_{i'}$	i'	1	$Y_{i'}$	$M^{-1}\sum_{l\in\mathscr{J}_{\mathbf{V_0},i'}} Y_l$	$Y_{i'}$

	(I) Matching Variable		(II) Matching Variable		
	V	d_V	V_0	V_1	d_V
M.X	X	$\dim(X)$	X	X	$\dim(X)$
PSM	$e(X)$	1	$e(X)$	$e(X)$	1
PGM	$\{\Psi_0(X),\Psi_1(X)\}$	2	$\Psi_0(X)$	$\Psi_1(X)$	1
DSM	$S=\{e(X),\Psi_0(X),\Psi_1(X)\}$	3	$S_0=\{e(X),\Psi_0(X)\}$	$S_1=\{e(X),\Psi_1(X)\}$	2
DSM	$S=\{e^j(X),\Psi_0^k(X),\Psi_1^k(X):$ $j=1,\dots J;k=1,\dots,K\}$	$J+2K$	$S_0=\{e^j(X),\Psi_0^k(X):$ $j=1,\dots J;k=1,\dots,K\}$	$S_1=\{e^j(X),\Psi_1^k(X):$ $j=1,\dots J;k=1,\dots,K\}$	$J+K$

Table 2.2: The order of bias of the matching variable in terms of the dimension of the matching variable

d_V	1	2	3	4	>4
$O(n^{1/2-2/d_V})$	$O(n^{-3/2})$	$O(n^{-1/2})$	$O(n^{-1/6})$	$O(1)$	$O(n^{1/10})$

Lemma 2 implies that $\mathbb{E}\{Y(a)\} = \mathbb{E}[\mathbb{E}\{Y \mid A=a, e(X), \Psi_a(X)\}]$, $(a = 0,1)$. For its interpretation, it is useful to compare it to the result in Lemma 1. By Lemma 1, we create subpopulations where we can simultaneously compare the treated units and the control units. These subpopulations were defined by common values for $\{e(X),\Psi(X)\}$. By Lemma 2, we do not construct such populations. The key insight is that in order to estimate τ, it is not necessary to do so. Instead, we construct subpopulations where we can estimate the average value of the potential outcomes for $a=0$ and 1 separately. For a given a, these subpopulations are defined by the value of $\{e(X),\Psi_a(X)\}$. This difference allows us to reduce the dimension of the double score from three to two, a small reduction of the dimension of the matching variable, a big reduction of the bias order of the matching estimator.

We focus on estimating $\mu_a = \mathbb{E}\{Y(a)\}$ separately for $a=0,1$. Let the matching variable be the double score $S_a(X)=\{e(X),\Psi_a(X)\}$ or S_a for shorthand. Table 2.1 (II) illustrates the new

matching scheme to impute the missing potential outcomes. For unit i with $A_i = 1$, $\hat{Y}_i(1) = Y_i$ and $\hat{Y}_i(0) = M^{-1}\sum_{l \in \mathscr{J}_{S_0,i}} Y_l$. For unit i' with $A_{i'} = 0$, $\hat{Y}_{i'}(1) = M^{-1}\sum_{l \in \mathscr{J}_{S_1,i'}} Y_l$ and $\hat{Y}_{i'}(0) = Y_{i'}$. Importantly, the new matching scheme uses different matching variables, namely S_0 and S_1, to impute the missing values of $Y(0)$ and $Y(1)$. This is in contrast to matching scheme (I) that uses the same matching variable for imputing the missing values of $Y(0)$ and $Y(1)$. Once we approximate both potential outcomes for all units, a simple DSM estimator of τ is

$$\hat{\tau}^{(0)}_{\text{dsm}} = \hat{\mu}^{(0)}_{1,\text{dsm}} - \hat{\mu}^{(0)}_{0,\text{dsm}}, \tag{2.2}$$

where $\hat{\mu}^{(0)}_{a,\text{dsm}} = n^{-1}\sum_{i=1}^{n} \hat{Y}_i(a) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(A_i = a)\big(1 + M^{-1}K_{S_a,i}\big)\, Y_i$, for $a = 0, 1$. Because $\dim(S_a) = 2$, $\hat{\tau}^{(0)}_{\text{dsm}}$ is asymptotically unbiased.

2.3.3 Multiply robust DSM

In practice, S_0 and S_1 are unknown, requiring modeling and estimation from the observed data. Following the empirical literature, one can posit a logistic regression model for the propensity score and a generalized linear model for the prognostic score. To provide additional protection against model misspecification, we can posit multiple candidate models for both scores. The intuition is that if at least one of the candidate models is correctly specified, whether it is a propensity score model or a prognostic score model, balancing at least one score suffices to remove confounding biases. Therefore, the DSM estimator achieves the so-called multiple robustness.

Following Han and Wang (2013), we postulate multiple candidate models

- $\mathscr{M}(\alpha) = \{e^j(X;\alpha^j) : j = 1,\ldots,J\}$ for $e(X)$ with unknown parameters $\alpha = (\alpha^{1,\mathrm{T}},\ldots,\alpha^{J,\mathrm{T}})^{\mathrm{T}}$;
- $\mathscr{M}_0(\beta_0) = \{\Psi_0^k(X;\beta_0^k) : k = 1,\ldots,K\}$ and $\mathscr{M}_1(\beta_1) = \{\Psi_1^k(X;\beta_1^k) : k = 1,\ldots,K\}$ for $\Psi_0(X)$ and $\Psi_1(X)$, respectively, with unknown parameters $\beta_0 = (\beta_0^{1,\mathrm{T}},\ldots,\beta_0^{K,\mathrm{T}})^{\mathrm{T}}$ and $\beta_1 = (\beta_1^{1,\mathrm{T}},\ldots,\beta_1^{K,\mathrm{T}})^{\mathrm{T}}$.

Let $\hat{\alpha}^j$, $\hat{\beta}_0^k$ and $\hat{\beta}_1^k$ be the maximum likelihood estimators or the method of moments estimators of α^j, β_0^k and β_1^k under the corresponding working model, respectively.

For each treatment level $a \in \{0,1\}$, let $S_a(\theta_a) = \{\mathscr{M}(\alpha), \mathscr{M}_a(\beta_a)\}$, where $\theta_a^{\mathrm{T}} = (\alpha^{\mathrm{T}}, \beta_a^{\mathrm{T}})$, be the set of candidate models for the propensity score and the prognostic score for treatment a, for $a = 0, 1$. Under matching scheme (II), we use $S_a(\hat{\theta}_a)$ to impute the missing values of $Y(a)$, separately for $a = 0, 1$. The corresponding dimension of the matching variable is $J + K$. Let $S(\theta) = \{\mathscr{M}(\alpha), \mathscr{M}_0(\beta_0), \mathscr{M}_1(\beta_1)\}$, where $\theta = (\alpha^{\mathrm{T}}, \beta_0^{\mathrm{T}}, \beta_1^{\mathrm{T}})^{\mathrm{T}}$, be the set of candidate

models for the propensity score and the prognostic score for both treatment groups. Under matching scheme (I), one would use $S(\hat{\theta})$ as the matching variable; the corresponding dimension is thus $J+2K$. If the number of candidate models for the prognostic score is large, the dimension reduction of the double score under new matching scheme (II) can be much larger than under matching scheme (I).

The initial DSM estimator of τ is given by $\hat{\tau}^{(0)}_{\text{dsm}}$ in (2.2) with S_a replaced by $S_a(\hat{\theta}_a)$ for $a=0,1$. We denote the initial estimator as $\hat{\tau}^{(0)}_{\text{dsm}}(\hat{\theta})$ to reflect its dependence on $\hat{\theta}$. As discussed in Section 2.3.2, if $J=K=1$, the dimension of $S_a(\hat{\theta})$ is two for $a=0,1$. In this case, the asymptotic bias of the matching estimator due to the matching discrepancy is negligible. We do not require further steps to correct the asymptotic bias of $\hat{\tau}^{(0)}_{\text{dsm}}$. This preserves the simplicity of matching in practice. However, if $J,K\geq 2$, the dimension of each matching variable is larger than or equal to four. Consequently, as shown in Table 2.2, the bias of the matching estimator due to matching discrepancy is not asymptotic negligible. In this case, we propose the de-biasing matching estimator that corrects the asymptotic bias due to matching discrepancy.

Let $\hat{\mu}_a(S_a)$ be a nonparametric estimator of $\mu_a(S_a)$, for $a=0,1$, e.g., using the method of sieves (Chen 2007). The de-biasing DSM estimator of τ is

$$\hat{\tau}_{\text{dsm}}(\hat{\theta})=\hat{\tau}^{(0)}_{\text{dsm}}(\hat{\theta})-n^{-1/2}\hat{B}_n, \tag{2.3}$$

where $\hat{B}_n$ is an estimator of B_n by replacing $\mu_a(S_a)$ with $\hat{\mu}_a(S_a)$ for $a=0,1$.

Before delving into the discussion of the theoretical properties of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$, we summarize the DSM algorithm that contains nuts and bolts as follows.

Step 1. Posit multiple candidate parametric models $\mathscr{M}(\alpha)$, $\mathscr{M}_0(\beta_0)$, and $\mathscr{M}_1(\beta_1)$ for $e(X)$, $\Psi_0(X)$, and $\Psi_1(X)$, respectively. Obtain the parameter estimators $\hat{\alpha}$, $\hat{\beta}_0$, and $\hat{\beta}_1$. For each unit i, calculate $S_{a,i}(\hat{\theta}_a)=\{\mathscr{M}(\hat{\alpha}),\mathscr{M}_a(\hat{\beta}_a)\}$ for $a=0,1$. The propensity scores are probability estimates, ranging from zero to one. To stabilize the numerical performance, it is desirable to use a monotone mapping to transform each propensity score estimate $e^j(X_i;\hat{\alpha}_j)\in(0,1)$ in $\mathscr{M}(\hat{\alpha})$, e.g., to $\text{logit}\{e^j(X_i;\hat{\alpha}_j)\}\in\mathscr{R}$. We also recommend standardize $S_{a,i}(\hat{\theta}_a)$ such that each component has mean zero and variance one for $a=0,1$.

Step 2. For each unit i with treatment $A_i=a$, find M nearest neighbors from the treatment group $1-a$ based on the matching variable $S_{1-a,i}=S_{1-a,i}(\hat{\theta})$. Obtain $\mathscr{J}_{S_{1-a}(\hat{\theta}),i}$ that contains the indexes of the matched units for unit i and calculate $K_{S_a(\hat{\theta}),i}$ that counts

the number of time that unit i is matched to other units. Obtain the initial matching estimator $\hat{\tau}^{(0)}_{\text{dsm}}(\hat{\theta})$ in (2.2) with S_a replaced by $S_a(\hat{\theta}_a)$.

If $J = K = 1$, let the DSM estimator be $\hat{\tau}_{\text{dsm}}(\hat{\theta}) = \hat{\tau}^{(0)}_{\text{dsm}}(\hat{\theta})$. If $J, K \geq 2$, we proceed to Steps 3 and 4 below. Even with $J = K = 1$, Steps 3 and 4 can help to reduce the matching discrepancy in finite samples.

Step 3. Obtain a nonparametric estimator of $\mu_a(S_a)$, denoted by $\hat{\mu}_a(S_a)$, e.g. by the method of sieves based on $[\{Y_i, S_{a,i}(\hat{\theta}_a)\} : A_i = a]$, for $a = 0, 1$.

Step 4. The DSM estimator of τ is given by $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ in (2.3) with $S_{a,i}$ replaced by $S_{a,i}(\hat{\theta}_a)$.

2.4 Main results

In this section, we establish the asymptotic properties of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$, which depends on the estimators of all nuisance parameters in the propensity score and prognostic score models. Without loss of generality, we consider the prognostic score $\Psi_a(X) = \mu_a(X; \beta_a)$ and multiple candidate models $\Psi^k_a(X; \beta^k_a) = \mu^k_a(X; \beta^k_a)$, for $k = 1, \ldots, K$ and $a = 0, 1$. Consider $\hat{\alpha}^j$, $\hat{\beta}^k_0$ and $\hat{\beta}^k_1$ that solve the estimating equation

$$n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} U^j_1(A_i, X_i; \alpha^j) \\ U^k_2(A_i, X_i, Y_i; \beta^k_0) \\ U^k_3(A_i, X_i, Y_i; \beta^k_1) \end{pmatrix} = 0, \tag{2.4}$$

where

$$\begin{aligned} U^j_1(A, X; \alpha^j) &= \frac{\partial e^j(X; \alpha^j)}{\partial \alpha^j} \frac{A - e^j(X; \alpha^j)}{e^j(X; \alpha^j)\{1 - e^j(X; \alpha^j)\}}, \\ U^k_2(A, X, Y; \beta^k_0) &= (1 - A) \frac{\partial \mu^k_0(X; \beta^k_0)}{\partial \beta^q_0} \{Y - \mu^k_0(X; \beta^k_0)\}, \\ U^k_3(A, X, Y; \beta^k_1) &= A \frac{\partial \mu^k_1(X; \beta^k_1)}{\partial \beta^k_1} \{Y - \mu^k_1(X; \beta^k_1)\}, \end{aligned}$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Then, $\hat{\theta}$ solves the joint estimating equation $\mathscr{U}_n(\theta) = n^{-1/2} \sum_{i=1}^{n} U(A_i, X_i, Y_i; \theta) = 0$, where $U(\theta)$ stacks $U^j_1(A_i, X_i, \alpha^j)$ for $j = 1, \ldots, J$, $U^k_2(A_i, X_i, Y_i; \beta^k_0)$ and $U^k_3(A_i, X_i, Y_i; \beta^k_1)$ for $k = 1, \ldots, K$.

Let θ^* be the probability limit of $\hat{\theta}$. We divide our theoretical investigation into two steps: first, we establish the asymptotic results for the DSM estimator when θ^* is known,

and second, building on the step-one results, we quantify the impact of the estimation of θ^* on the asymptotic distribution.

2.4.1 Asymptotic results with known θ^*

We allow possible model misspecification, so θ^* may not be the true parameter values. If $e^j(X;\alpha^j)$ is a correctly specified model, we have $e^j(X;\alpha^{j*}) = e(X)$; if $\mu_a^k(X;\beta_a^k)$ is a correctly specified model, we have $\mu_a^k(X;\beta_a^{k*}) = \mu_a(X)$, for $a = 0,1$. The key insight is that if any model of the propensity score or prognostic score is correctly specified, $S_a(\theta^*)$ remains as a balancing score in the sense that $Y(a) \perp\!\!\!\perp A \mid S_a(\theta^*)$ holds for $a = 0,1$ (Lemma 2). In the following theorem, we establish the multiple robustness and asymptotic distribution of $\hat{\tau}_{\text{dsm}}(\theta^*)$.

Theorem 1 *Under Assumption 1 and regularity conditions in Assumption 10, if any model of the propensity score or prognostic score is correctly specified, we have* $n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta^*) - \tau\} \to \mathcal{N}(0, V_\tau)$, *in distribution, as* $n \to \infty$, *where*

$$V_\tau = \mathbb{E}\big[\{\mu_1(S_1) - \mu_0(S_0) - \tau\}^2\big] + \mathbb{E}\left(\sigma_1^2(S_1)\left[\frac{1}{e(S_1)} + \frac{1}{2M}\left\{\frac{1}{e(S_1)} - e(S_1)\right\}\right]\right) + \mathbb{E}\left(\sigma_0^2(S_0)\left[\frac{1}{1-e(S_0)} + \frac{1}{2M}\left\{\frac{1}{1-e(S_0)} - 1 + e(S_0)\right\}\right]\right). \quad (2.5)$$

Remark 1 (Connection with AIPW) *The de-biased matching estimator has an intrinsic connection with the AIPW estimator. The AIPW estimator of* τ *is*

$$\hat{\tau}_{\text{aipw}} = n^{-1}\sum_{i=1}^{n}\left[\mu_1(X_i) - \mu_0(X_i) + \frac{2A_i - 1}{\widehat{\mathbb{P}}(A_i \mid X_i)}\big\{Y_i - \mu_A(X_i)\big\}\right] + o_{\text{P}}(n^{-1/2}). \quad (2.6)$$

It is well-known that $\widehat{\tau}_{\text{aipw}}$ *is doubly robust and semiparametrically efficient when the outcome model and the propensity score model are correctly specified. In the proof of Theorem 1, we show that the DSM estimator* $\hat{\tau}_{\text{dsm}}(\theta^*)$ *exhibits a similar asymptotic expansion as the AIPW estimator*

$$\hat{\tau}_{\text{dsm}}(\theta^*) = n^{-1}\sum_{i=1}^{n}\big[\mu_1(S_{1,i}) - \mu_0(S_{0,i}) + (2A_i - 1)\big(1 + M^{-1}K_{S_A,i}\big)\big\{Y_i - \mu_A(S_{A,i})\big\}\big] + o_{\text{P}}(n^{-1/2}). \quad (2.7)$$

When one of the prognostic score models is correctly specified, $\mu_a(S_{a,i}) = \mu_a(X_i)$ *for* $a = 0,1$. *From Theorem 1, the asymptotic variance of* $\hat{\tau}_{\text{dsm}}(\theta^*)$ *(2.5) does not achieve the semipara-*

metric efficiency bound (Hahn 1998) for a fixed M, but it becomes closer when M is larger. However, $\widehat{\tau}_{\text{aipw}}$ is extremely unstable when the estimated propensity scores are close to zero or one due to the weight construction, and it is sensitive to slight model misspecification. Instead of inverting the estimated propensity score, the DSM estimator uses the matching weights, $1+M^{-1}K_{S_A,i}$, for covariate balancing and hence is more robust in the scenarios with extreme propensity score values. This is confirmed in the simulation study.

Remark 2 *From Theorem 1, the consistency of the DSM estimator is guaranteed if any model for the propensity score or prognostic score is correctly specified. Both the number of the posited models and their functional forms can affect the efficiency of the DSM estimator in a very complex way. In addition, with a finite sample size, the matching performance can be unstable if there are a large number of working models. In particular, the discrepancy of the matched units may be large when some of the models are poorly constructed. To reduce the chance of running into these issues, we suggest positing a few well-constructed working models instead of a large number of poorly built ones.*

2.4.2 Asymptotic results with estimated θ^*

To acknowledge the fact that θ^* is estimated prior to matching, we will establish the approximate distribution of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ and examine the impact of nuisance parameter estimation on the properties of the DSM estimator. As in Abadie and Imbens (2016), the typical Taylor expansion technique can not be used because of the non-smooth nature of matching. Our derivation is based on the technique developed by Andreou and Werker (2012), which offers a general approach for deriving the limiting distribution of statistics that involve estimated nuisance parameters. This technique has been successfully used by Abadie and Imbens (2016) for the PSM estimators of the ATE and ATT based on a correctly specified propensity score model. We extend their results to the DSM estimator requiring only one of the double score models to be correctly specified.

Theorem 2 *Under Assumption 1 and regularity conditions in Assumptions 10–13, if any model of the propensity score or prognostic score is correctly specified, the approximate distribution of $n^{1/2}\left\{\hat{\tau}_{\text{dsm}}(\hat{\theta})-\tau\right\}$ is $\mathcal{N}(0, V_{\tau,\text{adj}})$, where*

$$V_{\tau,\text{adj}} = V_\tau - \gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1 + \gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2, \tag{2.8}$$

where V_τ is given in (2.5), $\Sigma_U = \mathbb{E}\{U(A,X,Y;\theta^*)\,U(A,X,Y;\theta^*)^{\mathrm{T}}\}$, $\Sigma_{\theta^*} = \Gamma_{\theta^*}^{-1}\Sigma_U(\Gamma_{\theta^*}^{-1})^{\mathrm{T}}$, $\Gamma_{\theta^*} = \mathbb{E}\{\partial U(A,X,Y;\theta^*)/\partial\theta^{\mathrm{T}}\}$, γ_1 and γ_2 are given in (A.14) and (A.10), respectively.

We discuss the impact of estimating the nuisance parameters on the matching estimators. Abadie and Imbens (2016) showed that for τ, matching on the estimated propensity score always improves the estimation efficiency compared to matching on the true propensity score. This improvement is due to the correlation of the matching estimator and the score function for the parameters in the propensity score. In our context, comparing the asymptotic variances in Theorems 1 and 2, the difference between $V_{\tau,\text{adj}}$ and V_τ, $-\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1+\gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2$, can be either positive, negative, or zero; i.e., matching on the estimated double score can either increase, decrease, or maintain the estimation efficiency compared to matching on the true double score. To explain the difference, we note that the variance reduction term $-\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1$ is still due to the correlation of the matching estimator and the score function for the parameters in the double score, while the variance inflation term $\gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2$ is because if either the prognostic score model or the propensity score model is misspecified, τ may depend on the nuisance parameters through $\tau=\mathbb{E}\big[\mu_1\{S_1(\theta^*)\}-\mu_0\{S_0(\theta^*)\}\big]$, which contributes to the variance inflation term. On the other hand, Abadie and Imbens (2016) focused on the setting when the propensity score model is the only nuisance model and is correctly specified. In this case, τ does not depend on α^*, γ_2 is zero, and therefore the variance inflation term is zero.

2.4.3 Variance estimation and inference

Theorem 2 provides guidance for variance estimation of the DSM estimators that can take all sources of variability into account. However, such variance estimators are complicated to construct. We consider variance estimation based on replication methods (Efron 1979; Wolter 2007). Lack of smoothness makes the standard replication methods invalid for the matching estimator. When the number of matches remains fixed, Abadie and Imbens (2008) demonstrated the failure of the bootstrap for matching estimators. This is because the non-parametric bootstrap cannot preserve the distribution of the number of times that each unit is used as a match. In this case, Otsu and Rai (2017) proposed a wild bootstrap procedure for the matching estimator when matching is directly based on the covariates. Yang and Kim (2019) proposed a replication-based procedure for predictive mean matching in survey data.

Given the two-stage estimation procedure for the DSM estimator, the variability of the matching estimator results from two sources: first, the estimation of the double score function, and second, matching. Following Yang and Kim (2019), we propose a two-stage replication variance estimation procedure, in parallel to the two-stage point estimation

procedure. First, we construct replicates of the nuisance parameter estimators in the double score. Second, based on the asymptotic linear representations of the DSM estimator, we construct replicates of the DSM estimator directly based on the linear terms with the replicated nuisance parameters. In this way, the distribution of the number of times that each unit is used as a match can be retained.

Specifically, the replication variance estimation algorithm proceeds as follows.

VE-Step 1. Obtain a bootstrap sample, or equivalently the bootstrap replication weights $\omega_i^* = n^{-1}m_i^*$ with $(m_1^*, \ldots, m_n^*)$ is a multinomial random vector with n draws on n equal probability cells. Obtain a bootstrap replicate of $\hat{\theta}$, $\hat{\theta}^*$, by solving the estimating equation $n^{-1/2}\sum_{i=1}^n \{\omega_i^* U(A_i, X_i, Y_i; \theta)\} = 0$. For each unit i, calculate $S_{a,i}(\hat{\theta}^*)$ for $a = 0, 1$.

VE-Step 2. Obtain a bootstrap replicate of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ as

$$\begin{aligned}\hat{\tau}_{\text{dsm}}^*(\hat{\theta}^*) &= n^{-1}\sum_{i=1}^n \omega_i^* \big[\hat{\mu}_1\{S_{1,i}(\hat{\theta}^*)\} - \hat{\mu}_0\{S_{0,i}(\hat{\theta}^*)\}\big] \\ &\quad + n^{-1}\sum_{i=1}^n \omega_i^*(2A_i - 1)\Big\{1 + M^{-1}K_{S_{A_i}(\hat{\theta}),i}\Big\}\big[Y_i - \hat{\mu}_{A_i}\{S_{A_i,i}(\hat{\theta}^*)\}\big].\end{aligned}$$

VE-Step 3. Repeat VE-Steps 1 and 2 a large number of times. Calculate the bootstrap variance estimator of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ as the empirical variance of $\hat{\tau}_{\text{dsm}}^*(\hat{\theta}^*)$ over a large number of bootstrap replicates.

2.5 Multiply robust matching estimator of the QTE

Matching is attractive for general causal estimation because it can be viewed as a hot deck imputation method (Ford 1983), where for each unit the donors for the missing potential outcome are actually observed values from the opposite treatment group. An advantage of hot deck imputation is that it preserves the distribution of the potential outcomes so that valid estimators for parameters depending on the entire distribution of the potential outcomes such as the mean and quantiles can be obtained based on the imputed data set. In this section, we extend the DSM framework to estimate the QTE.

We focus on estimating $q_{a,\xi}$ separately for $a = 0, 1$. By Lemma 1, we have

$$q_{a,\xi} = \inf_q \big(\mathbb{E}[\mathbb{P}\{Y \le q \mid A = a, e(X), \Psi_a(X)\}] \ge \xi\big).$$

Based on the above equation, we propose the DSM estimator of $q_{a,\xi}$ as

$$\hat{q}_{a,\xi,\text{dsm}} = \inf_q \{\hat{F}_{a,\text{dsm}}(q) \geq \xi\}, \tag{2.9}$$

where

$$\begin{aligned}
\hat{F}_{a,\text{dsm}}(q) &= \hat{F}^{(0)}_{a,\text{dsm}}(q) - n^{-1/2}\hat{B}_{a,n}(q), &(2.10)\\
\hat{F}^{(0)}_{a,\text{dsm}}(q) &= n^{-1}\sum_{i=1}^{n} \mathbf{1}(A_i = a)\big(1 + M^{-1}K_{S_a,i}\big)\mathbf{1}(Y_i \leq q),\\
\hat{B}_{a,n}(q) &= -n^{-1/2}\sum_{i=1}^{n} \mathbf{1}(A_i = 1-a)M^{-1}\sum_{j\in\mathscr{J}_{S_a,i}} \big\{\hat{F}_a(q;S_{a,i}) - \hat{F}_a(q;S_{a,j})\big\}, &(2.11)
\end{aligned}$$

and $\hat{F}_a(q;S_a)$ is a semi/nonparametric estimator of $F_a(q;S_a) = \mathbb{P}\{Y(a) \leq q \mid S_a\}$, for $a = 0,1$. Note that $\hat{F}^{(0)}_{a,\text{dsm}}(q)$ is an initial matching estimator of $F_a(q) = \mathbb{P}\{Y(a) \leq q\}$, $\hat{B}_{a,n}(q)$ is the bias correction term. Then the DSM estimator of Δ_ξ is $\hat{\Delta}_{\xi,\text{dsm}} = \hat{q}_{1,\xi,\text{dsm}} - \hat{q}_{0,\xi,\text{dsm}}$.

For estimating Δ_ξ, Steps 1 and 2 of DSM in Section 2.3.3 remain the same; Steps 3′ and 4′ proceed as follows:

Step 3′. Obtain a semiparametric estimator of $F_a(q;S_a)$ based on $[\{Y_i, S_{a,i}(\hat{\theta})\} : A_i = a]$, for $a = 0,1$.

Step 4′. The DSM estimator of $q_{a,\xi}$ is given by (2.9) with S_a replaced by $S_a(\hat{\theta})$. We denote the final estimator of $q_{a,\xi}$ as $\hat{q}_{a,\xi\text{dsm}}(\hat{\theta})$ to reflect its dependence on $\hat{\theta}$, for $a = 0,1$. Then, the DSM estimator of Δ_ξ is $\hat{\Delta}_{\xi,\text{dsm}}(\hat{\theta}) = \hat{q}_{1,\xi,\text{dsm}}(\hat{\theta}) - \hat{q}_{0,\xi,\text{dsm}}(\hat{\theta})$.

Remark 3 *In Step 3', many choices can be considered for estimating $F_a(q;S_a)$: e.g., the method of sieves for the normal linear model after a Box-Cox transformation of Zhang et al. (2012), the single-index conditional distribution model of Chiang and Huang (2012), or the distribution regression models of Foresi and Peracchi (1995) and Chernozhukov et al. (2013). The data at hand and subject matter knowledge can be used to guide the choice of the models. For example, if the transformed outcome is believed to follow a normal distribution, Zhang et al. (2012)'s method is preferable; otherwise, other semiparametric approaches are desirable.*

Under Assumption 1, regularity conditions in Assumptions 10 (i) and 14, if any model of the propensity score or prognostic score is correctly specified, similar to the proof in

Section A.1, we have $\mathrm{d}\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi})/\mathrm{d}q = f_a(q_{a,\xi}) + o_{\mathrm{P}}(n^{-1/2})$, and then we express $\hat{q}_{a,\xi,\mathrm{dsm}}$ as

$$\hat{q}_{a,\xi,\mathrm{dsm}} - q_{a,\xi} = -\frac{\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}) - F_a(q_{a,\xi})}{f_a(q_{a,\xi})} + o_P(n^{-1/2}), \tag{2.12}$$

$q_{a,\xi}$ lies in a closed interval $\mathscr{I}$. Expression (2.12) is called the Bahadur-type representation for $\hat{q}_{a,\xi,\mathrm{dsm}}$ (Francisco and Fuller 1991). With the representation (2.12), we can then extend the multiple robustness and asymptotic distributions of the ATE estimation to $\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)$ and $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$.

Theorem 3 *Under Assumption 1, regularity conditions in Assumptions 10 (i) and 14, if any model of the propensity score or prognostic score is correctly specified, for all* $\xi \in \widetilde{\mathscr{I}} = \{\xi : F_a(x) = \xi, x \in \mathscr{I}\}$, $n^{1/2}\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*) - \Delta_\xi\} \to \mathcal{N}(0, V_\xi)$, *in distribution, as* $n \to \infty$, *where* V_ξ *is given in (A.17).*

Theorem 4 *Under Assumption 1, regularity conditions in Assumptions 10 (i)–14, if any model of the propensity score or prognostic score is correctly specified, the approximate distribution of* $n^{1/2}\left\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*) - \Delta_\xi\right\}$ *is* $\mathcal{N}(0, V_{\xi,\mathrm{adj}})$, *where*

$$V_{\xi,\mathrm{adj}} = V_\xi - \gamma_3^{\mathrm{T}}\Sigma_U^{-1}\gamma_3 + \gamma_4^{\mathrm{T}}\Sigma_{\theta^*}\gamma_4, \tag{2.13}$$

V_ξ *is given in (A.17),* Σ_U *and* Σ_{θ^*} *are given in Theorem 3,* γ_3 *and* γ_4 *are given in (A.18) and (A.19), respectively.*

For variance estimation of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$, VE-Step 1 in Section 2.4.3 remains the same; VE-Steps 2 and 3 proceed as follows:

VE-Step 2′. For $a = 0, 1$, obtain a bootstrap replicate of $\hat{q}_{a,\xi\mathrm{dsm}}(\hat{\theta})$, $\hat{q}^*_{a,\xi\mathrm{dsm}}(\hat{\theta}^*)$, by solving

$$\begin{aligned}\hat{F}^*_{a,\mathrm{dsm}}(q) = n^{-1}\sum_{i=1}^n \omega_i^* \hat{F}_a\{q; S_{a,i}(\hat{\theta}^*)\} \\ + n^{-1}\sum_{i=1}^n \omega_i^* \mathbf{1}(A_i = a)\left\{1 + M^{-1}K_{S_a(\hat{\theta}),i}\right\}\left[\mathbf{1}(Y_i \le q) - \hat{F}_a\{q; S_{a,i}(\hat{\theta}^*)\}\right] = \xi,\end{aligned}$$

for q. Then a bootstrap replicate of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$ is $\hat{\Delta}^*_{\xi,\mathrm{dsm}}(\hat{\theta}^*) = \hat{q}^*_{1,\xi,\mathrm{dsm}}(\hat{\theta}^*) - \hat{q}^*_{0,\xi,\mathrm{dsm}}(\hat{\theta}^*)$.

VE-Step 3′. Repeat VE-Steps 1 and 2' for a large number of times. Calculate the bootstrap variance estimator of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$ as the empirical variance of $\hat{\Delta}^*_{\xi,\mathrm{dsm}}(\hat{\theta}^*)$ over a large number of bootstrap replicates.

2.6 Extensions to the causal effects on the treated

In this extension, we estimate the average causal effect on the treated τ_{ATT} and the quantile treatment effect on the treated $\Delta_{\text{ATT},\xi} = q_{1,\xi|A=1} - q_{0,\xi|A=1}$, where $q_{a,\xi|A=1} = \inf_q[P\{Y(a) \leq q\} \geq \xi \mid A = 1]$, $a = 0,1$. Here, because $f\{Y(1) \mid A = 1\} = f(Y \mid A = 1)$, the outcome distribution for the treated is identifiable. Therefore, $\mathbb{E}\{Y(1) \mid A = 1\} = \mathbb{E}(Y \mid A = 1)$ and $q_{1,\xi|A=1} = \inf_q \{\mathbb{P}(Y \leq q \mid A = 1) \geq \xi\}$.

To identify the outcome distribution for the control, Assumption 1 can be relaxed (Heckman et al. 1997b).

Assumption 2 *i)* $Y(0) \perp\!\!\!\perp A \mid X$*; and ii) there exists a constant* c *such that* $e(X) \leq c < 1$ *almost surely.*

For the causal effects on the treated, the prognostic score $\Psi_0(X)$ is a sufficient statistic for $Y(0)$ in the sense that $Y(0) \perp\!\!\!\perp X \mid \Psi_0(X)$ according to Hansen (2008). Then, under Assumption 2,

$$\begin{aligned} \tau_{\text{ATT}} &= \mathbb{E}[\mathbb{E}(Y \mid A = 1) - \mathbb{E}\{Y \mid A = 0, e(X)\} \mid A = 1] \\ &= \mathbb{E}[\mathbb{E}(Y \mid A = 1) - \mathbb{E}\{Y \mid A = 0, \Psi_0(X)\} \mid A = 1], \end{aligned}$$

and

$$\begin{aligned} q_{0,\xi|A=1} &= \inf_q \big(\mathbb{E}[\mathbb{P}\{Y \leq q \mid A = 0, e(X)\} \mid A = 1] \geq \xi\big) \\ &= \inf_q \big(\mathbb{E}[\mathbb{P}\{Y \leq q \mid A = 0, \Psi_0(X)\} \mid A = 1] \geq \xi\big), \end{aligned}$$

encoding the double balancing properties of $S = \{e(X), \Psi_0(X)\}$.

The DSM estimators for τ_{ATT} and $\Delta_{\text{ATT},\xi}$ follow similar steps as for τ and Δ_ξ. We describe the differences below.

In the matching step, for each unit i with treatment $A_i = 1$, find M nearest neighbors from the control group $A_i = 0$ based on the matching variable $S_i = S_i(\hat{\theta})$. Let these matched units for unit i be indexed by $\mathscr{J}_{S(\hat{\theta}),i}$.

The initial and de-biasing DSM estimators of τ_{ATT} are

$$\begin{aligned}
\hat{\tau}^{(0)}_{\mathrm{ATT,dsm}} &= n_1^{-1}\sum_{i=1}^{n} A_i\{Y_i - \hat{Y}_i(0)\},\ \hat{Y}_i(0) = M^{-1}\sum_{j\in\mathscr{I}_{S,i}} Y_j,\\
\hat{\tau}_{\mathrm{ATT,dsm}} &= \hat{\tau}^{(0)}_{\mathrm{ATT,dsm}} - n_1^{-1}\sum_{i=1}^{n} A_i\left\{\hat{\mu}_0(S_i) - M^{-1}\sum_{j\in\mathscr{I}_{S,i}}\hat{\mu}_0(S_j)\right\}.
\end{aligned}$$

Let the estimator of $F_1(q \mid A=1) = \mathbb{P}\{Y(1) < q \mid A=1\}$ be

$$\hat{F}_1(q \mid A=1) = n_1^{-1}\sum_{i=1}^{n} A_i \mathbf{1}(Y_i \le q).$$

Then, we estimate $q_{1,\xi|A=1}$ by

$$\hat{q}_{1,\xi|A=1} = \inf_q\{\hat{F}_1(q \mid A=1) \ge \xi\}.$$

The initial and de-biasing DSM estimators of $F_0(q \mid A=1) = \mathbb{P}\{Y(0) < q \mid A=1\}$ are

$$\begin{aligned}
\hat{F}^{(0)}_{0,\mathrm{dsm}}(q \mid A=1) &= n_1^{-1}\sum_{i=1}^{n} A_i M^{-1}\sum_{j\in\mathscr{I}_{S,i}} \mathbf{1}(Y_j \le q) = n_1^{-1}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{S,i}\mathbf{1}(Y_i \le q),\\
\hat{F}_{0,\mathrm{dsm}}(q \mid A=1) &= \hat{F}^{(0)}_{0,\mathrm{dsm}}(q \mid A=1) - n_1^{-1/2}\hat{B}_{0,n}(q),\\
\hat{B}_{0,n}(q) &= -n_1^{-1/2}\sum_{i=1}^{n} A_i M^{-1}\sum_{j\in\mathscr{I}_{S,i}}\left\{\hat{F}_0(q;S_i) - \hat{F}_0(q;S_j)\right\}.
\end{aligned}$$

Then, we estimate $q_{0,\xi|A=1}$ by

$$\hat{q}_{0,\xi|A=1,\mathrm{dsm}} = \inf_q\{\hat{F}_{0,\mathrm{dsm}}(q \mid A=1) \ge \xi\}.$$

Lastly, the DSM estimator of $\Delta_{\mathrm{ATT},\xi}$ is $\hat{\Delta}_{\mathrm{ATT},\xi,\mathrm{dsm}} = \hat{q}_{1,\xi|A=1} - \hat{q}_{0,\xi|A=1,\mathrm{dsm}}$.

For variance estimation, we replace the VE-Step 2 and VE-Step 2' for τ and Δ_ξ by the following steps:

ATT-VE-Step 2. Obtain a bootstrap replicate of $\hat{\tau}_{\mathrm{ATT,dsm}}(\hat{\theta})$,

$$\hat{\tau}^*_{\mathrm{ATT,dsm}}(\hat{\theta}^*) = n_1^{-1}\sum_{i=1}^{n}\omega_i^* A_i\left[\hat{\mu}_1\{S_i(\hat{\theta}^*)\} - \hat{\mu}_0\{S_i(\hat{\theta}^*)\}\right] + n_1^{-1}\sum_{i=1}^{n}\omega_i^*\left\{A_i - (1-A_i)M^{-1}K_{S(\hat{\theta}),i}\right\}\left[Y_i - \hat{\mu}_{A_i}\{S_i(\hat{\theta}^*)\}\right].$$

QTT-VE-Step 2′. For $a=1$, obtain a bootstrap replicate of $\hat{q}_{1,\xi|A=1}(\hat{\theta})$, $\hat{q}^*_{1,\xi|A=1}(\hat{\theta}^*)$, by solving

$$\hat{F}^*_1(q \mid A=1) = n_1^{-1}\sum_{i=1}^{n}\omega_i^* A_i \mathbf{1}(Y_i \le q) = \xi.$$

For $a=0$, obtain a bootstrap replicate of $\hat{q}_{0,\xi|A=1,\mathrm{dsm}}(\hat{\theta})$, $\hat{q}^*_{0,\xi|A=1,\mathrm{dsm}}(\hat{\theta}^*)$, by solving

$$\hat{F}^*_{0,\mathrm{dsm}}(q \mid A=1) = n_1^{-1}\sum_{i=1}^{n}\omega_i^* A_i \hat{F}_0\{q; S_i(\hat{\theta}^*)\} + n_1^{-1}\sum_{i=1}^{n}\omega_i^* \mathbf{1}(A_i=0)M^{-1}K_{S(\hat{\theta}),i}\left[\mathbf{1}(Y_i \le q) - \hat{F}_0\{q; S_i(\hat{\theta}^*)\}\right] = \xi,$$

for q. Then a bootstrap replicate of $\hat{\Delta}_{\mathrm{ATT},\xi,\mathrm{dsm}}(\hat{\theta})$ is

$$\hat{\Delta}^*_{\mathrm{ATT},\xi,\mathrm{dsm}}(\hat{\theta}^*) = \hat{q}^*_{1,\xi|A=1}(\hat{\theta}^*) - \hat{q}^*_{0,\xi|A=1,\mathrm{dsm}}(\hat{\theta}^*)$$

2.7 Simulation study

We conduct a simulation study to investigate the finite-sample performance of the proposed DSM estimators relative to existing weighting and matching estimators. In the causal inference and missing data literature, previous simulations (e.g., Kang and Schafer 2007) have found that weighting estimators can have high variability, especially if the probabilities are close to zero or one. Frölich (2004) found that the weighting estimator was inferior to matching estimators in terms of root mean squared error. It has been found that matching on high-dimensional covariates is not practical for commonly found sample sizes (e.g., Abadie and Imbens 2006). In the comparative effectiveness research, PGM has been shown to be more advantageous than PSM when the propensity score distributions are strongly separated (Wyss et al. 2015; Kumamaru et al. 2016). Imbens (2004) noted that if the re-

gression models are misspecified, PGM may be inconsistent. These results motivate us to compare the weighting and matching estimators in a setting with complex data generative models, and where the propensity scores may be extreme (i.e., close to zero or one) or non-extreme.

Let the sample size be $n = 1000$. Confounder $X \in \mathbb{R}^{10}$ is generated by $X_j \overset{\text{iid}}{\sim}$ Uniform $[1-\sqrt{3}, 1+\sqrt{3}]$ for $j = 1, \ldots, 10$. To introduce nonlinear relationships between X and dependent variables, let $Z \in \mathbb{R}^{10}$ be a nonlinear transformation of X, where $Z_1 = \exp(X_1/2)$, $Z_2 = \exp(X_2/3)$, $Z_3 = \log\{(X_3+1)^2\}$, $Z_4 = \log\{(X_4+1)^2\}$, $Z_5 = \mathbf{1}(X_5 > 0.5)$, $Z_6 = \mathbf{1}(X_6 > 0.75)$, $Z_7 = \sin(X_7 - X_8)$, $Z_8 = \cos(X_7 + X_8)$, $Z_9 = \sin(X_9)$, and $Z_{10} = \cos(X_{10})$, which are further scaled and centered such that $\mathbb{E}(Z_j) = 1$ and $\mathbb{V}(Z_j) = 1$ for all j. The potential outcomes are $Y(0) = \beta_0^{\mathrm{T}} Z + \epsilon(0)$ and $Y(1) = Y(0) - \epsilon(0) + \epsilon(1)$, where $\beta_0^{\mathrm{T}} = (0.1, 1, 1, 1, 1, -1, -1, -1, -1, -1)$, $\epsilon(0) \sim \mathcal{N}(0, 2^2)$, and $\epsilon(1) \sim \mathcal{N}(0,1)$. Under the data generative model, the ATE τ is 0 and the 75th QTE is -0.45. An additional simulation with heterogeneous treatment effects and log-normal errors is presented in the supplementary material. The treatment indicator A follows Bernoulli$\{e(X)\}$, where logit$\{e(X)\} = \alpha_0^{\mathrm{T}} Z$. We consider two scenarios for the propensity score distribution: in the first case, $\alpha_0^{\mathrm{T}} = (5, -5, 1, 1, 2, -2, -2, 1, -1, -1)$, resulting in some extreme values of $e(X)$ that are close to zero or one; and in the second case, the propensity score distribution is not extreme, where $\alpha_0^{\mathrm{T}} = (5, -5, 1, 1, 2, -2, -2, 1, -1, -1)/4$. Visualization of the propensity score distributions in the two cases is presented in Figure A.2.

To assess the multiple robustness property of the DSM estimators, we consider two model specifications for the propensity score: 1) a correctly specified logistic regression model logit$\{e^1(X;\alpha^1)\} = \alpha^{1,\mathrm{T}} Z$ and 2) a misspecified logistic regression model logit$\{e^2(X;\alpha^2)\} = \alpha^{2,\mathrm{T}} X$; we also consider two model specifications for the prognostic score: 3) a correctly specified regression model $\mu_a^1(X;\beta_a^1) = \beta_a^{1,\mathrm{T}} Z$ for $a = 0, 1$ and 4) a misspecified regression model $\mu_a^2(X;\beta_a^2) = \beta_a^{2,\mathrm{T}} X$ for $a = 0, 1$.

We compare the following estimators:

a) naive, which is the simple difference of standard estimators from two treatment groups;

b) the weighting estimators including the IPW, AIPW, and MRW estimators ("ipw", "aipw", and "mrw");

c) the matching estimators based on X ("m.x"; bias-corrected Abadie and Imbens 2011), or propensity score ("psm") or prognostic score ("pgm") or double score ("dsm") with

$M = 5$.

Each weighting and matching estimator is assigned a name in the form of "method-0000", where each digit of the four-digit number, from left to right, indicates if $e^1(X;\alpha^1)$, $e^2(X;\alpha^2)$, $\{\mu_a^1(X;\beta_a^1)\}_{a=0}^1$, or $\{\mu_a^2(X;\beta_a^2)\}_{a=0}^1$ is used in the construction of the method, with "1" meaning yes and "0" meaning no, respectively. For example, "ipw1000" is the IPW estimator with the propensity score model $e^1(X;\alpha^1)$ and "dsm1110" is the DSM estimator with two propensity score models $e^1(X;\alpha^1)$, $e^2(X;\alpha^2)$ and one prognostic score model $\{\mu_a^1(X;\beta_a^1)\}_{a=0}^1$. We implement standard IPW and AIPW estimators for the ATE estimation and the corresponding estimators of Zhang et al. (2012) for the QTE estimation. MRW is implemented by the R package "MultiRobust". For all matching estimators, the conditional outcome mean functions are approximated using power series, and the conditional distribution functions are approximated based on the power series for the Normal linear model (Zhang et al. 2012).

Figure 2.1 shows the distributions of the estimation error (i.e.; the estimator minus the true parameter value) based on 1000 repeated sampling. The naive estimator is biased for the 75th QTE and ATE. Matching directly based on 10-dimensional X (indicated by "m.x") is biased for the QTE and ATE even with bias correction. This suggests that matching on high-dimensional covariates is not practical and calls for dimension reduction. IPW is unstable and sensitive to the extreme values of the propensity score. Even when the propensity score model is correctly specified (indicated by "ipw1000"), IPW is biased of the QTE and ATE. By theory, AIPW is supposed to be doubly robust: it should have small biases if either the propensity score model or the prognostic model is correctly specified. However, as indicated by "aipw1001", AIPW is biased of the QTE and ATE when the propensity score is extreme even if its model is correctly specified. We examine the empirical distribution of the estimated propensity score weights and find that there are extremely large weights that dominate other weights. Therefore, weighting estimators by inverting the estimated propensity scores are sensitive to outliers of the propensity score estimates. To mitigate this issue, one can stabilize the weighting estimators by normalizing the weights (Hernán et al. 2001). However, this strategy is not effective in our setting. Although AIPW is constructed to be semiparametrically efficient, its performance can be poor when it involves large weights. By construction, matching does not invert the estimated propensity scores and therefore is more robust to outliers of the propensity score estimates. We now compare the performances of the score-based matching estimators. The single score matching estimators (indicated by "psm1000", "psm0100", "pgm0010", "pgm0001") are singly robust

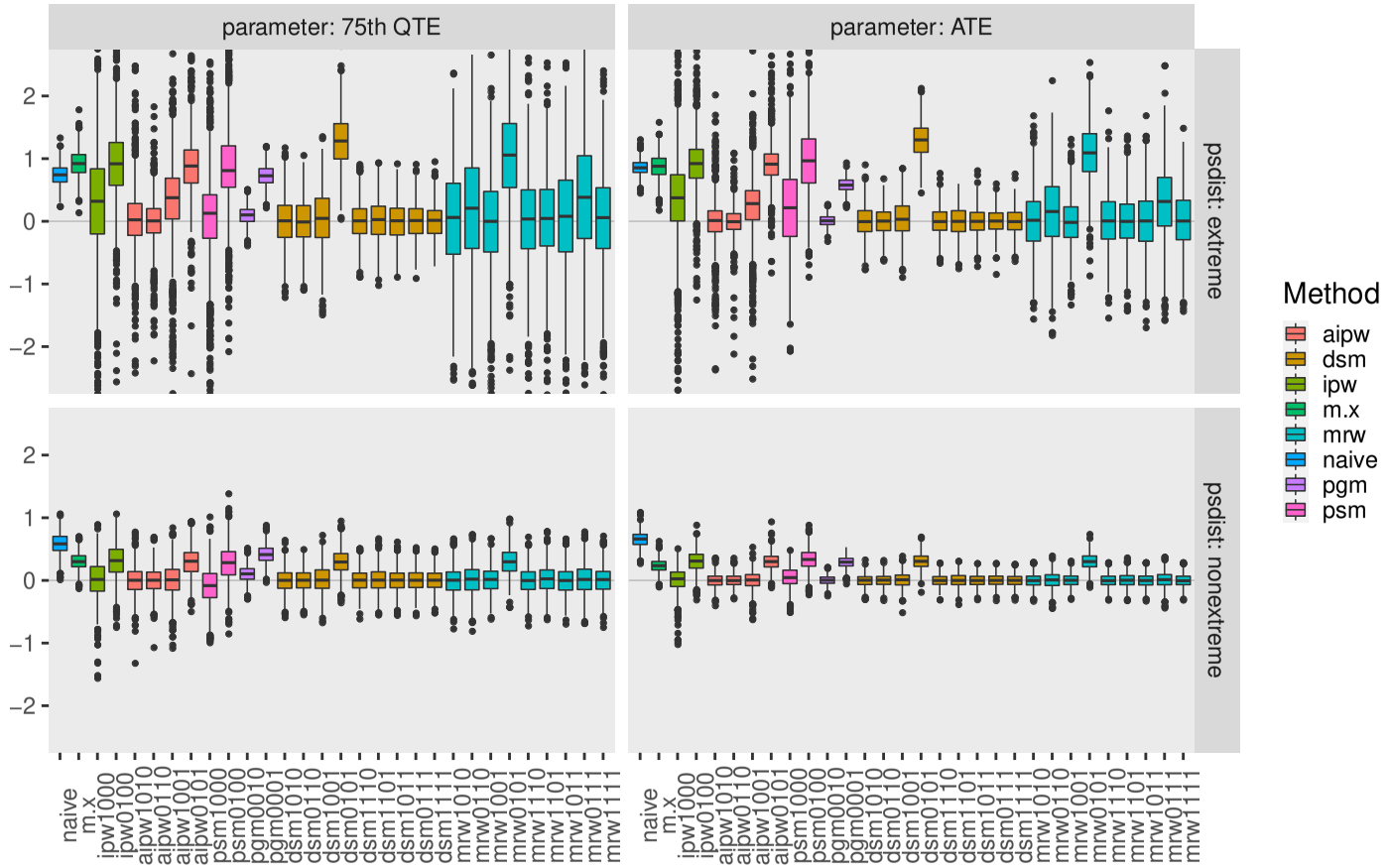


Figure 2.1: Simulation results of various weighting and matching estimators. There are four panels of results: the left for the 75th QTE, the right for the ATE, the top for the extreme propensity score, and the bottom for the non-extreme propensity score. Each box plot shows the distribution of the estimator subtracting the true parameter value based on 1000 Monte Carlo simulated datasets.

Table 2.3: Simulation results based on 1000 Monte Carlo simulated datasets for the coverage properties of the DSM estimators using the replication-based method: empirical coverage rate and (empirical coverage rate ± 1.96×Monte Carlo standard error)

	Non-extreme PS				Extreme PS			
	75th QTE		ATE		75th QTE		ATE	
"dsm1010"	94.8	(93.4,96.2)	95.2	(93.9,96.5)	96.0	(94.8, 97.2)	95.8	(94.6, 97.0)
"dsm0110"	94.4	(93.0,95.8)	96.0	(94.8,97.2)	95.7	(94.4, 97.0)	95.2	(93.9, 96.5)
"dsm1001"	95.4	(94.1,96.7)	95.9	(94.7,97.1)	95.9	(94.7, 97.1)	95.7	(94.4, 97.0)
"dsm0101"	72.9	(70.1,75.6)	29.2	(26.4,32.0)	19.4	(16.9, 21.9)	0.2	(-0.1, 0.5)
"dsm1111"	95.0	(93.6,96.4)	95.5	(94.2,96.8)	95.4	(94.1, 96.7)	95.8	(94.6, 97.0)
"dsm1110"	95.4	(94.1,96.7)	95.6	(95.5,97.7)	95.7	(94.4, 97.0)	96.2	(95.0, 97.4)
"dsm1101"	94.8	(93.4,96.2)	95.4	(94.1,96.7)	96.3	(95.1, 97.5)	95.4	(94.1, 96.7)
"dsm1011"	94.0	(92.5,95.5)	95.8	(94.6,97.0)	95.7	(94.4, 97.0)	95.8	(94.6, 97.0)
"dsm0111"	95.3	(94.0,96.6)	95.2	(93.9,96.5)	96.0	(94.8, 97.2)	95.4	(94.1, 96.6)

and rely on a correct specification of the underlying score model. DSM and MRW are multiply robust in that they have small biases for the QTE and the ATE if any model of the propensity score or prognostic score is correctly specified. Compared to MRW, DSM is more robust to extreme values of the propensity score estimates. Therefore, DSM is advantageous in practice compared to weighting.

Table 2.3 reports the coverage rates of the DSM estimators of the 75th QTE and the ATE using the proposed replication-based method. Under the multiple robustness condition (i.e., if any model of the propensity score or prognostic score is correctly specified), the coverage rates are all close to the nominal coverage except for "dsm0101".

2.8 Real-data application

In this section, we apply the proposed DSM method as well as other matching methods in Section 2.7 to the well-known National Supported Work (NSW) data (LaLonde 1986; Firpo 2007). This dataset documented the effect of a job training program for the unemployed on future earnings. Following Dehejia and Wahba (1999), we include the comparison group from Westat's Matched Current Population Survey-Social Security (CPS) Administration File. In our analysis, we include 185 treated units and 689 control units from the NSW, as well as 429 comparison units from the CPS-3, a subset of the CPS data (LaLonde 1986; Firpo 2007). Seven baseline confounding covariates are used for this application: age, education, race, Hispanic, married, having no college degree, and real earnings in 1975. The outcome

Table 2.4: Covariate balance check before and after DSM

		age	educ	black	hisp	married	nodegr	re75
Before Matching	Treatment group mean	24.63	10.38	0.80	0.09	0.17	0.73	3066
	Control group mean	26.25	10.21	0.50	0.13	0.34	0.70	2745
	Stand diff. in means	-0.19	0.08	0.61	-0.10	-0.37	0.06	0.07
After Matching	Treatment group mean	24.63	10.38	0.80	0.09	0.17	0.73	3066
	Control group mean	25.01	10.31	0.69	0.12	0.21	0.73	2876
	Stand diff. in means	-0.04	-0.03	0.22	-0.08	-0.10	0.00	0.04

Table 2.5: Estimated ATT and QTTs at the 0.1, 0.25, 0.3, 0.5, 0.75 and 0.9 quantiles, and 95% Wald confidence intervals

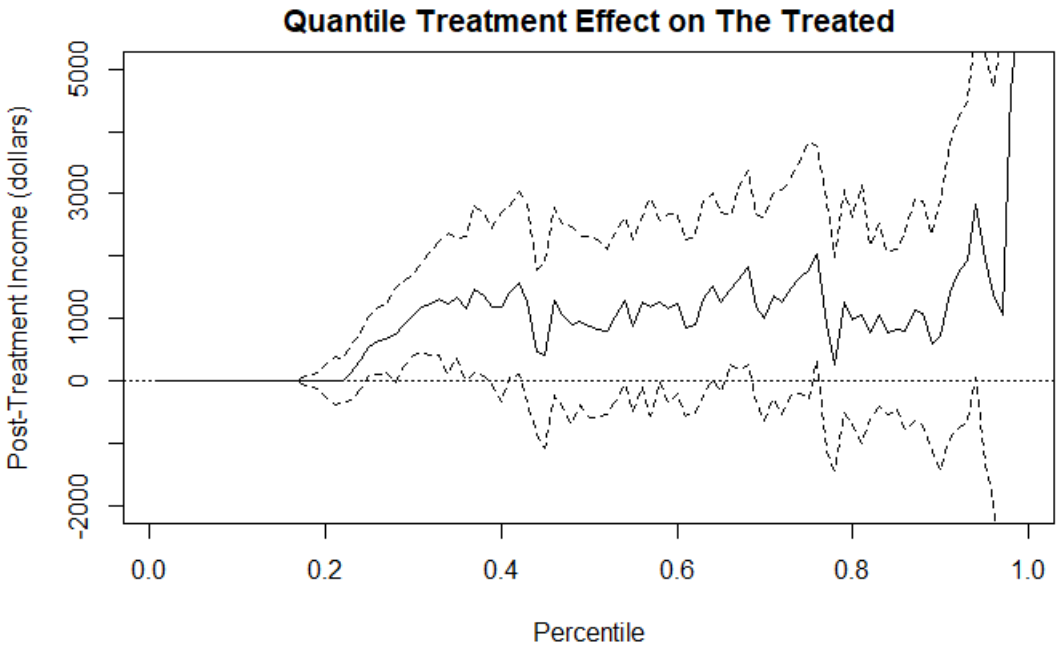
Estimand	m.x	psm	pgm	dsm	Naive(ATE&QTE)
ATT	372 (-746,1489)	918 (-222,2058)	-150 (-1215,914)	641 (-429,1711)	-65 (-957,827)
0.1-QTT	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
0.25-QTT	549 (-90,1189)	549 (-55,1154)	549 (-104,1202)	549 (-62,1160)	549 (-55,1153)
0.3-QTT	935 (57,1813)	1064 (268,1860)	1039 (96,1982)	1036 (211,1861)	604 (-641,1170)
0.5-QTT	524 (-1606,2654)	889 (-741,2519)	1296 (-252,2844)	680 (-766,2126)	69 (-1093,1230)
0.75-QTT	-391 (-2752,1970)	617 (-1451,2686)	737 (-2070,3544)	558 (-1297,2414)	-195 (-1441,1578)
0.9-QTT	-963 (-3482,1556)	897 (-1750,3543)	-936 (-4770,1795)	-718 (-3711,2275)	-2326 (-2021,2158)

of interest is the real earnings in 1978.

Because the outcome distributions are highly skewed (see Figure A.3), the average treatment effect may not provide a comprehensive evaluation of the job training program. Therefore, we estimate the ATT and QTTs. To gain the robustness and reliability of the results, we posit two propensity score models and two prognostic score models. Following Dehejia and Wahba (1999), one propensity score model is a logistic regression model with all first-order terms of the covariates and second-order terms of numerical variables, and one prognostic score model is a linear regression of the earnings with the same predictors as in the propensity score models for the control group. Given the popularity of probit models, we consider the second propensity score model to be a probit regression model with the same predictors in the first propensity score model. Given the skewness of the outcome distribution, we consider the second prognostic score to be a linear regression model for the logarithm of the real earnings in 1978.

Matching admits a transparent assessment of covariate balance before and after matching. Table 2.4 presents the means of all covariates by treatment group and the standardized difference in means before and after DSM. The standardized difference is calculated as the difference of the group means divided by the overall standard error in the original sample.

Figure 2.2: Quantile effect plot on the treated from the DSM algorithm



DSM makes standardized differences fall between −0.05 to 0.05 for all covariates, reducing the differences of the observed covariates in the treated and the control.

Table 2.5 shows the estimated ATTs and QTTs at the 0.1, 0.25, 0.3, 0.5, 0.75 and 0.9 quantiles, and 95% Wald confidence intervals from the four matching methods, as well as ATE and QTE estimated by the naive method. All four matching estimators show that the job training program does not have a significant effect on the average earning for the treated. Figure 2.2 shows the QTT plot estimated by DSM algorithm. A closer inspection of the QTT plot reveals that the effect is, in fact, significant around the percentile of 0.3, which suggests that the program is beneficial for the lower middle class.

2.9 Discussion

We have developed multiply robust matching estimators for general causal estimands. This framework offers a new "metric" to summarize the differential roles of different covariates and also serves as a powerful dimensional reduction tool in high-dimensional confounding. The improved robustness comes from multiple model specifications for the propensity score and prognostic score. The proposed DSM thus provides multiple protections to model

misspecification and therefore is an attractive alternative to existing weighting estimators.

Several issues are worth discussing. As with PSM, although the matching variables are well balanced, individual covariates may not for a given application. In this case, if the researchers know important confounders based on substantive knowledge, they can augment the double score by adding those confounders to ensure balance for these confounders; however, adding too many variables will results in potential bias as demonstrated in our simulation. Alternatively, one can use regression adjustment for the matched sample Abadie and Spiess (2016), which can remove remaining confounding biases. We focus on a binary treatment. Yang et al. (2016) has developed the generalized propensity score matching for estimating the treatment effects for more than two treatments. It is of interest to extend our DSM algorithm to more than two treatment comparisons. The current DSM framework focuses on continuous and binary outcomes, and it would be an important task to extend DSM to handle survival outcomes (Tang et al. 2019) and clustered data (Yang 2018) and estimate the heterogeneous treatment effects Huang and Yang (2021). It is crucial to highlight that as for all existing matching methods, the DSM method cannot account for unmeasured confounding. Following Rosenbaum and Rubin (1983a) and Yang and Lok (2017), we will develop sensitivity analyses to no unmeasured confounding in the matching framework.

CHAPTER

3

PRACTICAL RECOMMENDATIONS ON DOUBLE SCORE MATCHING FOR ESTIMATING CAUSAL EFFECTS

3.1 Introduction

Randomized clinical trials (RCTs) are the touchstone for treatment effect evaluation. By trial design, treatment randomization guarantees that treatment groups are comparable and thus bias can be minimized to the extent possible. However, in practice, it may be infeasible to conduct an RCT due to financial, logistic or ethical reasons. In these settings, comparative analyses using observational data may be of particular value. Unlike RCTs, confounding control is critical for estimating the causal effects from observational studies due to the lack of treatment randomization. Under the unconfoundedness assumption, i.e., all pre-treatment variables that are predictors of treatment and outcome are observed, matching methods Cochran and Rubin (1973); Stuart (2010) can be used to emulate an RCT that is hidden in the observational study. Researchers from various disciplines have expanded this

field in both theory and practice for decades; see Stuart (2010) for a comprehensive review. To ensure the key assumptions hold, the effort is often made to collect a large number of possible confounders, rendering dimension reduction imperative in matching.

In their seminal paper, Rosenbaum and RubinRosenbaum and Rubin (1983b) demonstrated the vital role of the propensity score, which is defined as the conditional probability of receiving treatment given the confounders, as a balancing score. The key implication is that matching on the scalar propensity score can dramatically reduce the confounding bias. Since then, propensity score matching (PSM) has been the most common method in the industry. However, recently, PSM has been criticized to be ineffective in that it attempts to emulate a completely randomized trialKing and Nielsen (2019), which is rarely if at all implemented in practice.

The prognostic score is an important alternative score summarizing covariate correlations with the outcomes. It is also known as the disease risk score in the epidemiology literature with a rich historyWyss et al. (2016). Its fundamental theory as a balancing score was established by HansenHansen (2008) in 2008, and he suggested various ways to utilize prognostic score in matching, called prognostic score matching (PGM). PGM balances the disease risk between the treatment groups and thus attempts to emulate a special blocked randomized trial, where the blocks are formed by the risk levels. The reason we call it a special trial is that not all the covariates are balanced within blocks as in standard blocked randomized design. Only prognostic factors that are highly related to the outcomes are adjusted for. More importantly, the potential outcomes are balanced within blocks, which is our ultimate goal in a trial. By not adjusting for unimportant variables that are not related to the outcomes, the special design may obtain higher efficiency. Wyss *et al.*Wyss et al. (2015) used simulations and an empirical example to illustrate the superior efficiency of PGM compared to PSM when the propensity score distributions are separated. Another notable advantage of PGM is that it is less sensitive than PSM to the practical violation of the overlap assumption. This is because treatment selection often shifts the propensity score distribution more dramatically than the prognostic score distribution between the treatment groups (see, e.g., Figure 3.3). As a result, the matching rate in PGM can be larger than that in PSM. However, this may not be true when the prognostic score is highly correlated with the propensity score. In this case, the prognostic score cannot provide any additional information compared to the propensity score. Thus, it is often reasonable to check the correlation between the two scores*et al.*Aikens et al. (2020). Moreover, Stuart *et al.*Stuart et al. (2013) showed empirically the positive correlation between the prognostic score and the bias of the treatment effect estimator and thus the prognostic score is useful

for balance check after matching. Nguyen and Debray extended the use of the prognostic score to the case of general treatment regimesNguyen and Debray (2019).

Combining the propensity score and prognostic score is indeed a sensible alternative to form a balancing score. Leacy and StuartLeacy and Stuart (2014) were the first in the statistical and epidemiological literature to assess the performance of jointly matching or stratifying on both the propensity score and prognostic score via simulation, although this idea was first raised by Hansen in a technical reportHansen (2006). In addition to the full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores and the full matching on the estimated prognostic propensity score (i.e. the propensity score predicted by the estimated prognostic score) within propensity score calipers, their paper also included subclassification on the two scores. As a result, although PGM achieves the best performance when the prognostic score model is correctly specified, full matching using the Mahalanobis distance combining the estimated propensity and prognostic scores is more robust to model misspecifications, while its performance is similar to PGM. Antonelli *et al.*Antonelli et al. (2018) later established the double robustness and convergence rate of the double score estimator of the average treatment effect (ATE). Yang and ZhangYang and Zhang (2020) coined the term "double score matching" (DSM) and derived the asymptotic distribution of the double score estimator for the average and quantile treatment effects. Both studies underscore the advantage of DSM being doubly robust against model misspecification of either the propensity score model or the prognostic score model. Besides matching, the propensity score and the outcome information are used in the augmented inverse probability weighting estimator that achieves semiparametric efficiency bound and double robustness. Hu *et al.* utilized the double score in the context of nonparametric estimationHu et al. (2012).

In short, propensity score, prognostic score, and double score are useful for dimension reduction and reducing confounding bias, resulting in three matching schemes. However, a comprehensive comparison among the three matching schemes is lacking and research towards the best practices including variable selection and choices of whether using caliper and replacement is still needed. First, an important topic is variable selection in the propensity score and prognostic score. In fact, model fitting has been an important and challenging issue in PSM. Brookhart*et al.*Brookhart et al. (2006) used simulations to suggest using covariates related to outcome and not using instrumental variables (IVs) to construct the propensity score model for propensity score weighting. Yang et alYang et al. (2020) recommended to select all variables that are predictive of either treatment or outcome for robustness consideration. De Luna *et al.* proposed an algorithm to select the minimal sets

of covariatesDe Luna et al. (2011). Myers *et al.*Myers et al. (2011) reported similar findings in PSM but prioritized minimizing unmeasured confounding when selecting variables, even at the risk of conditioning on IVs. At the same time, PearlPearl (2011) showed that the rate of bias amplification from IVs may be faster than the rate of bias reduction. Some researchers suggested that IVs may be less detrimental if the prognostic score is included in matchingWyss et al. (2016). Most existing recommendations are confined to PSM, and very few studies have investigated these issues for PGM and DSM.

Secondly, matching constraints are also important issues in matching methods, including whether to include caliper and whether to match with or without replacement. In PSM, a caliper of 0.25 standard deviation of linear propensity score was generally suggested by Rosenbaum and RubinRosenbaum and Rubin (1985). Other researchers used Monte Carlo methods to find the optimal caliper width equal to 0.2 of the standard deviation in some special circumstancesAustin (2011); Wang et al. (2013). Matching without replacement is claimed to reduce variance since each control sample is used only onceImbens and Rubin (2015). But when the control group is not large enough, bias can often be decreased by matching with replacementDehejia and Wahba (1999). However, these recommendations are restricted to PSM and may not extend to PGM and DSM given that the prognostic score distribution is less affected by treatment assignment than the propensity score distribution.

In this article, we explore the statistical and numerical properties of the three score-based matching methods via extensive simulations. Our study supports conclusion that DSM performs comparably with, if not better than, the other two single score matching. Table 3.1 summarizes the key features for comparison. Importantly, by linking matching scheme and trial design, we show that DSM emulates a hybrid design of complete randomization and blocked randomization that incorporates the blocking benefit of PGM, while retaining the balancing guarantee of PSM. We also provide bolts and nuts for DSM and illustrate the recommendations with comprehensive simulation studies. In essence, we propose the following steps for DSM to achieve its best performance in terms of bias and variance.

Step 1. At the variable selection stage, fit a prognostic score model with penalization (e.g. LASSO) to select all prognostic variables.

Step 2. Fit the propensity score model restricted to the selected prognostic variables and fit the prognostic score model with all the covariates.

Step 3. Calculate the double score and assess the overlap of the propensity score and

Table 3.1: Comparison among the three matching estimators. Results with * are based on the correct model specifications.

Matching scheme	The emulating trial design	Model specification requirement	Overlap requirement	Optimal configuration*			Performance under optimal configuration*
				Variable selection	Replacement	Caliper	
PSM	Completely randomized experiment	The propensity score model	Sufficient overlap of the propensity score distribution	Yes	Yes	Yes	Lower matching rate and large variance
PGM	Special blocked randomized experiment	The prognostic score model	Sufficient overlap of the prognostic score distribution	No	Yes	Insen-sitive	High matching rate and small variance
DSM	Hybrid randomized experiment	Either the propensity score or prognostic score model (double robustness)	Sufficient overlap of either the propensity score or prognostic score distribution	Yes for the propensity score model	Yes	Insen-sitive	High matching rate and small variance

the prognostic score, or alternatively, assess the matching rate after implementing matching based on the double score in the following step. We require sufficient overlap for at least one score or a large matching rate to apply DSM.

Step 4. Carry out matching with replacement based on the estimated double score and calculate the average treatment effect estimator.

Thanks to variable selection strategy, DSM can achieve the best efficiency across different settings in the simulation studies, in contrast to Leacy and Stuart's result that PGM performed best when models were correctly specified. Also, because all the variables are kept in the prognostic score model, the estimator will not suffer from confounding bias. This is a special advantage of DSM over PSM and PGM in consideration of variable selection. The double robustness property of DSM is also verified by extensive simulations across various scenarios. However, it is worth noting that all the above conclusions are based on limited simulation studies where covariates and outcomes are normally distributed continuous variables. Readers should be cautious not to apply these results to other scenarios before a general theoretical proof is derived.

The remaining sections of this paper proceed as follows. Section 3.2 introduces the three matching methods: PSM, PGM, and DSM. Section 3.3 focuses on model selection and can be divided into two parts: the first part uses a hypothetical simulating example to illustrate the performance of different variable selection strategies; the second part turns to a more realistic simulation setting using the REFLECTIONS datasetFaries et al. (2020). Because of the complex correlation in the covariate set, LASSO is applied to select variables into the outcome models. Simulation setup and detailed results are both presented in this part.

Section 3.4 presents the simulation results regarding the choice of caliper and replacement. Section 3.5 compares the three matching estimators based on their best configurations of variable selection strategy and choice of caliper and replacement. Section 3.6 concludes the paper and discusses possible directions of future work. Derivations of the asymptotic results for the three matching estimators and the comparison of their theoretical efficiency are included in Appendix B.1.

3.2 Methodology: matching estimators

We follow the potential outcome framework to fomulate causal effects of treatmentImbens and Rubin (2015). Let X_i be the set of covariates, A_i be the treatment indicator, and Y_i be the observed outcome for unit $i = 1, ..., n$. Let $Y_i(a)$ be the potential outcome had unit i been given treatment a, where $a = 1$ is the treatment of interest and $a = 0$ is the control group. It is assumed that $\{X_i, A_i, Y_i(0), Y_i(1)\}$, $i = 1, ..., n$ are independent and identically distributed. We intend to estimate the average treatment effect on the treated $\tau_{\text{ATT}} = E\{Y(1) - Y(0) \mid A = 1\}$.

We use matching to impute missing potential outcomes in causal inference. The main intuition is to find the closest subject in the control group for each individual in the treatment group. Note that here we restrict our analysis within the estimation of the average treatment effect on the treated (ATT) and one-to-one matching. Thus, the key step in matching is to define a proper distance measure to determine "closeness". PSM, PGM, and DSM are three variants with different measures of distance, and we will introduce them in detail in this section. Besides, calipers can be incorporated into the definition of distance so that subjects who cannot find a good match will be excluded from the analysis. Moreover, matching with or without replacement determines whether a control individual can be matched multiple times or not.

3.2.1 Propensity score matching

The propensity score is defined as the probability of receiving treatment given certain values of covariatesRosenbaum and Rubin (1983b):

$$e(X) = E(A \mid X) = P(A = 1 \mid X).$$

Here we restrict our analysis to binary treatment scenarios. As illustrated in Figure 3.1a, the path between covariates and treatment is blocked by the propensity score. Under positivity

and no unmeasured confounder assumptions (stated formally as Assumptions 1 and 2 later in Section B.1), the propensity score is a balancing score conditional on which the potential outcomes and treatment assignment are independent:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid e(X).$$

This implies that the estimand ATT can be written as:

$$\begin{aligned}\tau_{\text{ATT}} &= E\{Y(1) - Y(0) \mid A = 1\} \\ &= E[E(Y \mid A = 1) - E\{Y \mid A = 0, e(X)\} \mid A = 1].\end{aligned}$$

This ensures that we can use PSM to estimate the ATT. We define the distance D_{ij} between individuals i and j for matching based on the linear propensity score:

$$D_{ij} = \left|\text{logit}\{e(X_i)\} - \text{logit}\left\{e\left(X_j\right)\right\}\right|. \tag{3.1}$$

The linear version is used because the bias may be effectively reducedRosenbaum and Rubin (1985); Rubin and Thomas (1996); Rubin (2001). Later on, we will use the "propensity score" to refer to either the probability version or the linear version. Suppose there are n units and individuals indexed by $1, ..., n_1$ are from the treatment group. For each treated unit i, we denote J_i as the index of the closest control unit in the sense of distance defined by the propensity score:

$$J_i = \underset{j=n_1+1,...,n}{\arg\min}\ D_{ij}. \tag{3.2}$$

As a result, the ATT can be estimated by

$$\hat{\tau}_{\text{ATT}} = n_1^{-1} \sum_{i=1}^{n_1} \left(Y_i - Y_{J_i}\right). \tag{3.3}$$

To improve matching quality, we can incorporate a caliper into the definition of matching index J_i:

$$J_i = \begin{cases} \underset{j=n_1+1,...,n}{\arg\min}\ D_{ij} & \text{if } \min\limits_{j=n_1+1,...,n} D_{ij} \le c \\ 0 & \text{if } \min\limits_{j=n_1+1,...,n} D_{ij} > c, \end{cases} \tag{3.4}$$

where c is the preset caliper. This excludes treated units that are dissimilar from anyone in

the control group. Let $\mathscr{J}$ be the set of treated indices that will be included in matching:

$$\mathscr{J} = \{i : J_i \neq 0, i = 1, ..., n_1\}.$$

Then the ATT can be estimated by

$$\hat{\tau}_{\mathrm{ATT}} = |\mathscr{J}|^{-1} \sum_{i \in \mathscr{J}} \big(Y_i - Y_{J_i}\big). \tag{3.5}$$

Note that the above estimators are based on matching with replacement. When matching without a caliper and without replacement, J_i is chosen from the remaining control indices:

$$J_i = \underset{j \in \{n_1+1,...,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)}{\arg\min} D_{ij}. \tag{3.6}$$

When matching with a caliper and without replacement, the definition of J_i can be adapted in a similar way:

$$J_i = \begin{cases} \underset{j \in \{n_1+1,...,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)}{\arg\min} D_{ij} & \text{if } \min_{j \in \{n_1+1,...,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)} D_{ij} \leq c \\ 0 & \text{if } \min_{j \in \{n_1+1,...,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)} D_{ij} > c. \end{cases} \tag{3.7}$$

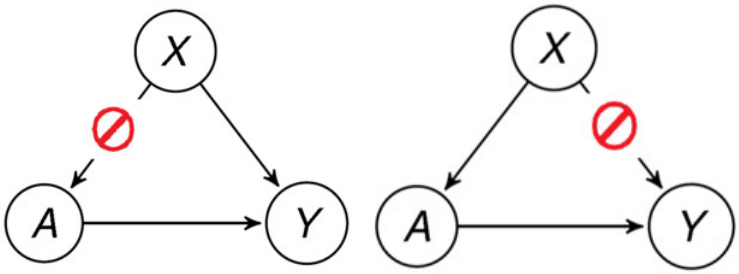
Note that in practice we don't have access to the true propensity score. Thus, the estimated propensity score $\hat{e}(X)$ will replace $e(X)$ in (3.1). In the simulation part of this paper, a logistic model is always used to fit the propensity score, but the set of variables may vary depending on our model selection strategy.

3.2.2 Prognostic score matching

The prognostic score $\Psi(X)$ is formally defined by HansenHansen (2008) as a balancing score in the sense that $Y(0) \perp\!\!\!\perp X \mid \Psi(X)$. We illustrate the prognostic score using the following examples.

Example 1 *If $Y(0)$ follows a generalized linear model with mean $\mu_0(X) = X^T\beta_0$ and constant variance, then $\Psi(X) = E(Y(0) \mid X) = X^T\beta_0$.*

Example 2 *If $Y(0)$ follows a location-shift family $f_0\{y - \mu_0(X)\}$, then $\Psi(X) = \mu_0(X)$.*



(a) DAG of PSM (b) DAG of PGM

Figure 3.1: DAGs to illustrate the different intuitions behind PSM and PGM. PSM blocks the path from covariates to treatment, while PGM blocks the path from covariates to outcome.

As illustrated in Figure 3.1b, the path between covariates and outcome is blocked by the prognostic score. Hansen showed that if there is no hidden bias, treatment ignorability holds by conditioning on the prognostic score:

$$Y(0) \perp\!\!\!\perp A \mid \Psi(X).$$

This implies that the ATT can be estimated by matching via the prognostic score:

$$\begin{aligned}\tau_{\text{ATT}} &= E\{Y(1) - Y(0) \mid A = 1\} \\ &= E[E(Y \mid A = 1) - E\{Y \mid A = 0, \Psi(X)\} \mid A = 1].\end{aligned}$$

Similarly, we define the distance D_{ij} between individuals i and j for matching based on the prognostic score:

$$D_{ij} = \left\| \Psi(X_i) - \Psi\left(X_j\right) \right\|. \tag{3.8}$$

Then we can estimate the ATT via formulas (3.2)-(3.7) depending on our choice of caliper and replacement. Similarly, an estimated prognostic score $\hat{\Psi}(X)$ will replace $\Psi(X)$ in (3.8), and a generalized linear model is used throughout this paper as in Example 1. Note that only the control group is used in the model fitting step but prognostic scores of units from both groups are required to be estimated.

3.2.3 Double score matching

Antonelli *et al.*Antonelli et al. (2018) showed that the double score as the combination of the propensity score and prognostic score is also a balancing score, and that conditioning on the double score deconfounds the potential outcomes from the treatment assignment:

$$Y(0) \perp\!\!\!\perp A \mid e(X), \Psi(X).$$

Note that this result holds even if only one score is correctly specified. This is the basis of the double robustness property of the DSM estimator. As a result, we can estimate the ATT by matching on the double score:

$$\begin{aligned}\tau_{\mathrm{ATT}} &= E\{Y(1) - Y(0) \mid A = 1\} \\ &= E[E(Y \mid A = 1) - E\{Y \mid A = 0, e(X), \Psi(X)\} \mid A = 1].\end{aligned}$$

To be specific, the distance metric is defined as the Mahalanobis distance combining propensity score and prognostic score:

$$D_{ij} = \begin{pmatrix} \mathrm{logit}\{e(X_i)\} - \mathrm{logit}\{e(X_j)\} \\ \Psi(X_i) - \Psi(X_j) \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \mathrm{logit}\{e(X_i)\} - \mathrm{logit}\{e(X_j)\} \\ \Psi(X_i) - \Psi(X_j) \end{pmatrix}, \tag{3.9}$$

where Σ is the covariance matrix of $\big(\mathrm{logit}\{e(X)\}, \Psi(X)\big)^T$. Mahalanobis distance removes the imbalance between the scales of two scores.

Formulas (3.2)-(3.7) can still be used to estimate the ATT, but Yang and ZhangYang and Zhang (2020) recommended to include a bias correction term into the estimator:

$$\hat{\tau}_{\mathrm{ATT}} = n_1^{-1} \sum_{i=1}^{n_1} \Big(\big(Y_i - Y_{J_i}\big) - \big[\hat{\mu}_0\{e(X_i), \Psi(X_i)\} - \hat{\mu}_0\big\{e\big(X_{J_i}\big), \Psi\big(X_{J_i}\big)\big\}\big] \Big), \tag{3.10}$$

where $\mu_0\{e(X_i), \Psi(X_i)\} = E\{Y(0) \mid e(X_i), \Psi(X_i)\}$ and we can obtain its rough estimate by the method of sieves. Note that (3.10) only works when caliper is not included. When matching with a caliper, we can simply calculate the bias correction term using individuals who remained in the matching set as (3.5). Although Yang and Zhang showed that this bias is asymptotically negligible, correcting for bias may increase finite sample performance in practice.

There is another special issue about DSM when a caliper is included. HansenHansen (2008) suggested using an ordinary propensity score caliper and matching on the prognostic

propensity score. Leacy and StuartLeacy and Stuart (2014) found that this is not better than matching based on the Mahalanobis distance of the two scores. Alternatively, we would like to incroporate a caliper in both matching by the propensity score and the prognostic score, in a procedure similar to PSM and PGM. However, we don't need to require both scores to be close enough. The intuition is that individuals within a matching pair are comparable if at least one of the scores satisfy the caliper constraint. To be specific, the matching index J_i in (3.4) changes its definition to:

$$J_i = \begin{cases} \underset{j=n_1+1,\ldots,n}{\arg\min}\, D_{ij} & \text{if } \min\limits_{j=n_1+1,\ldots,n} \min\left\{\left|\text{logit}\{e(X_i)\} - \text{logit}\{e(X_i)\}\right|, \left|\Psi(X_i) - \Psi\left(X_j\right)\right|\right\} \le c \\ 0 & \text{if } \min\limits_{j=n_1+1,\ldots,n} \min\left\{\left|\text{logit}\{e(X_i)\} - \text{logit}\{e(X_i)\}\right|, \left|\Psi(X_i) - \Psi\left(X_j\right)\right|\right\} > c. \end{cases}$$

When matching with replacement, the matching index J_i in (3.7) can be adapted similarly. As a result, the matching rate should be higher than both PSM and PGM. More importantly, overlapping assumptions are also relaxed. For units outside the overlapping region of the propensity score, they can still find matching pairs if they are in the overlapping region of the prognostic score. This is illustrated by the increased matching rate in the simulation results, and we hope to complete the theoretical proof in the future.

3.2.4 Summary of the three matching methods

We provide some insights by connecting the three matching methods with randomized trials; see also Table 3.1. King and NielsenKing and Nielsen (2019) linked PSM with a completely randomized experiment. On the other hand, PGM mimics a more efficient fully blocked randomized experiment with special design where only potential outcomes and prognostic factors are adjusted for. This explains why PGM may have a smaller variance than PSM when models are correctly specified. Meanwhile, DSM approximates a hybrid randomized experiment, which can be more efficient with the correctly specified prognostic score. When the outcome model is misspecified but the propensity score model is correct, DSM is still valid because it inherits the unbiasedness property from the complete randomized experiment.

So far, we have introduced the three types of estimators from PSM, PGM, and DSM, as well as their variants depending on different choices of caliper and replacement. It is natural to derive the theoretical results for these estimators. In the Appendix B.1, we provide some asymptotic results for estimating ATT. These are based on some simplifications of the problem, following Abadie and ImbensAbadie and Imbens (2016). First, we only

consider matching with replacement and without a caliper. Second, the asymptotic results are established when the coefficients in the propensity score and the prognostic score models are known. Under these assumptions, the theory justifies the high efficiency of PGM compared to DSM, which invokes the importance of involving variable selection algorithms into DSM. However, these two assumptions we made may oversimplify the problem as the errors from the estimation of coefficients are not addressed, which may have a significant impact especially in high dimensional settings. Thus, the theory can only offer suggestions instead of guarantees. Meanwhile, when considering additional factors such as variable selection, it is difficult to make comparisons theoretically. In the following three sections, we will use simulations to study the effect of variable selection strategy, matching constraints, and distance metric on the matching estimator and make some recommendations based on the results.

3.3 Variable selection in matching

In this section, we would like to address the questions: how variable selection will change the performance of each matching estimator and what kind of variable selection strategy we should use. To better answer these questions, we will first introduce an illustrating example that helps us understand the categories of variables and gives us guidance about the appropriate variable selection strategies. Then we will turn to a more comprehensive simulation study based on a more realistic dataset.

3.3.1 Importance and strategy of variable selection: an illustrating example

Simulation setup

We generated covariates for 3000 subjects from a 16-dimensional multivariate normal distribution, where the mean vector was 0 and the covariance matrix was a 16×16 identity matrix. The covariate distribution was highly hypothetical and unrealistic, but our goal here is to have a clear categorization of variables. Specifically, we generated the propensity score using $X_1, ..., X_7$ and generate the continuous potential outcomes using $X_1, ..., X_4$ and $X_8, ..., X_{10}$. Figure 3.2 illustrates this generating process. As a result, four covariates $(X_1, ..., X_4)$ were confounders, three $(X_5, ..., X_7)$ were IVs, three $(X_8, ..., X_{10})$ were predictors of outcome only, and six $(X_{11}, ..., X_{16})$ were noise variables. The generating mechanism follows the work

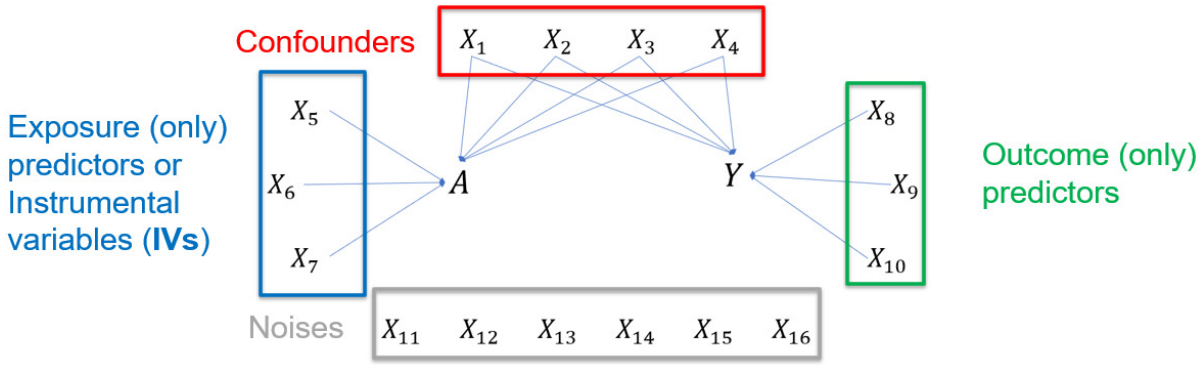


Figure 3.2: Data generation structure of the illustrating example.

of Leacy and StuartLeacy and Stuart (2014), but with noise variables included.

Figure 3.3 shows the distributions of the propensity score (on the original scale and logit scale) and prognostic score for both the treatment group and control group. Interestingly, the prognostic score had much better overlap than the propensity score. This was not caused by the special design of the simulation but by the definition and nature of the two scores. This difference in overlap will become an important explanation when we later compare the performance of the three matching estimators in Section 3.5. It is also worth noting that the difference in overlap relies on the correlation between the propensity score and the prognostic score. In this simple simulation, the correlation is very weak (around -0.14). When the propensity score is highly correlated with the prognostic score, the difference between PSM and PGM can be negligibleAikens et al. (2020). We provide an illustrating example in the supplementary material.

In this simple setting, the individual treatment effect was a constant 3 for all the 3000 subjects in the population. Thus, the ATT was also 3, and we would use PSM, PGM, and DSM to estimate the ATT. For simplicity, we applied one-to-one matching without a caliper and with replacement in all the three methods. In Section 3.3.2, we will see that the conclusions could be extended to other configurations of caliper and replacement.

Based on different categories of variables, we used the following five sets of covariates to fit each score.

- All variables: $X_1, ..., X_{16}$
- All except noise variables: $X_1, ..., X_{10}$
- Confounders + IVs: $X_1, ..., X_7$

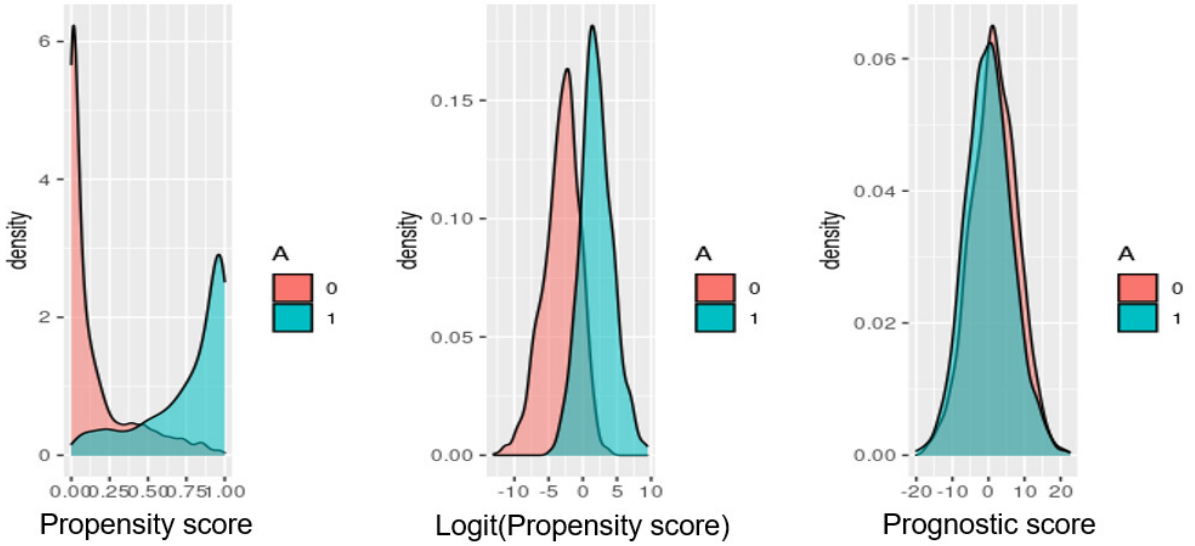


Figure 3.3: Overlapping of propensity score and prognostic score.

- Confounders + outcome predictors: $X_1, ..., X_4$ and $X_8, ..., X_{10}$
- Only confounders: $X_1, ..., X_4$

Note that we always included confounders when fitting scores, otherwise there would be significant confounding bias. We replicated the simulation 100 times and recorded the estimated ATT from all the three matching methods with all the five variable selection strategies. We seek to find the estimator with the smallest bias and variance. The results will be presented in the following section.

Results

Figure 3.4 shows the performance of the PSM estimator under the five variable selection configurations. It can be easily seen that removing instrumental variables is beneficial for PSM: the variance decreased significantly compared to the regular PSM estimator while the bias didn't increase much. This is consistent with existing results in the literatureBrookhart et al. (2006); Myers et al. (2011). Interestingly, even though the propensity score model was misspecified in these two configurations, the estimator was still consistent and even more accurate. This may be because a closer match on IVs is not useful in getting a better match on the potential outcome while overlap of the propensity score becomes poor when including these strong IVs. On the other hand, removing noise variables or outcome predictors

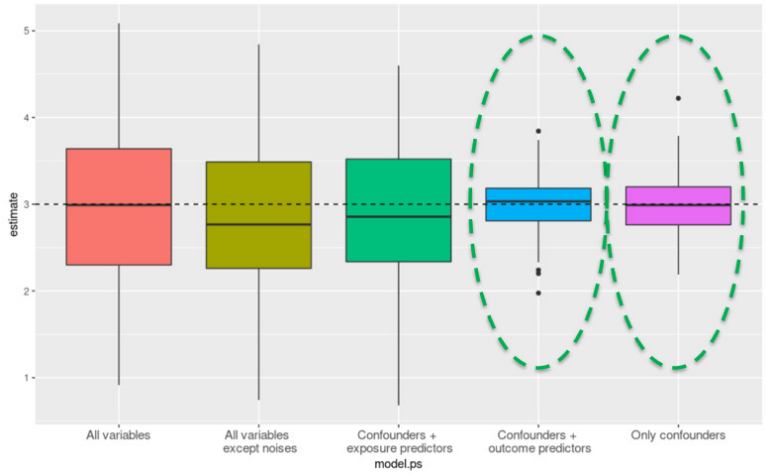


Figure 3.4: Performance of PSM estimator under different variable selection strategies in the illustrating example. Best configurations are marked by green circles.

does not help reduce the variance but may even increase the bias. This suggests that any model selection method not incorporating outcome may not be effective in improving the PSM estimator. It is also worth mentioning the sample bias observed in the propensity model that includes IVs. Since the theory of PSM guarantees the unbiasedness of the estimatorRosenbaum and Rubin (1983b), the observed bias should be a finite sample bias. This was verified by additional simulations reported in the supplementary material. When IVs are included, the finite sample bias may become larger as the variance becomes larger. To sum up, we recommend excluding IVs and noise variables before fitting the propensity score model in PSM.

Figure 3.5 illustrates the performance of the PGM estimator under the five variable selection strategies. Different from PSM, removing IVs was not very helpful in reducing the variance of the PGM estimator. However, falsely removing outcome predictors could be very harmful: the variance increased a lot, which may increase the finite sample bias of the estimator. This also shows the importance of outcome information. Note that keeping all the variables in the prognostic score model is one of the best configurations. Thus, when fitting the prognostic score model with all covariates is feasible, we recommend skipping the variable selection process in PGM.

In DSM, we have both the propensity score and prognostic score to fit, each with five possible sets of variables. Thus, there are $5 \times 5 = 25$ configurations in DSM. As illustrated in Figure 3.6, the best strategy is a combination of our previous strategies in PSM and PGM:

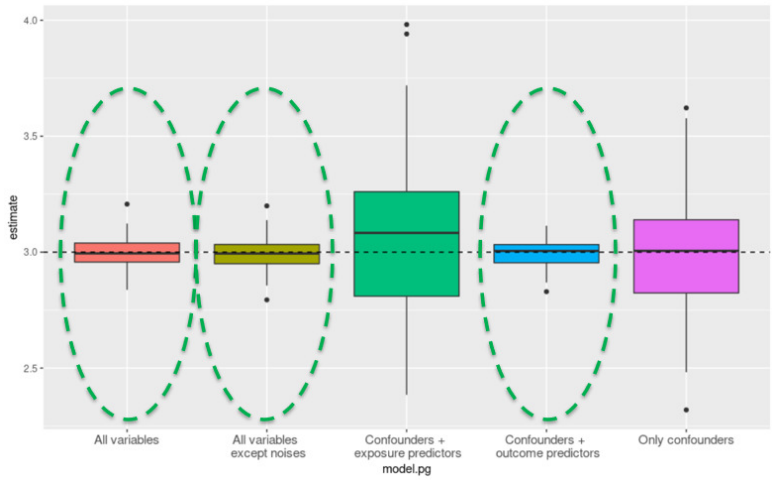


Figure 3.5: Performance of PGM estimator under different variable selection strategies in the illustrating example. Best configurations are marked by green ellipses.

we should remove IVs from the propensity score model and not remove outcome predictors from the prognostic score model. As a result, we recommend selecting confounders and outcome predictors to fit the propensity score model and using all the variables to fit the prognostic score model.

3.3.2 Variable selection in real world: REFLECTIONS dataset

We have figured out the importance and basic strategy of variable selection in matching methods. However, in real datasets, we can never know which variables are outcome predictors and which variables are IVs. As a result, we need some algorithms to select prognostic variables. What's more, various correlation patterns in real-world covariates can increase the complexity of variable categorization. To illustrate these problems, we will introduce the REFLECTIONS dataset and create simulations based on this.

Simulated REFLECTIONS dataset

REFLECTIONS stands for the Real World Examination of Fibromyalgia: Longitudinal Evaluation of Cost and Treatments datasetRobinson et al. (2012). It was a prospective observational study conducted at 58 clinical sites in the United States and Puerto Rico between 2008 and 2011. The goal was to examine the outcome for patients receiving new treatments for fibromyalgia. The observational data was collected from multiple sources including

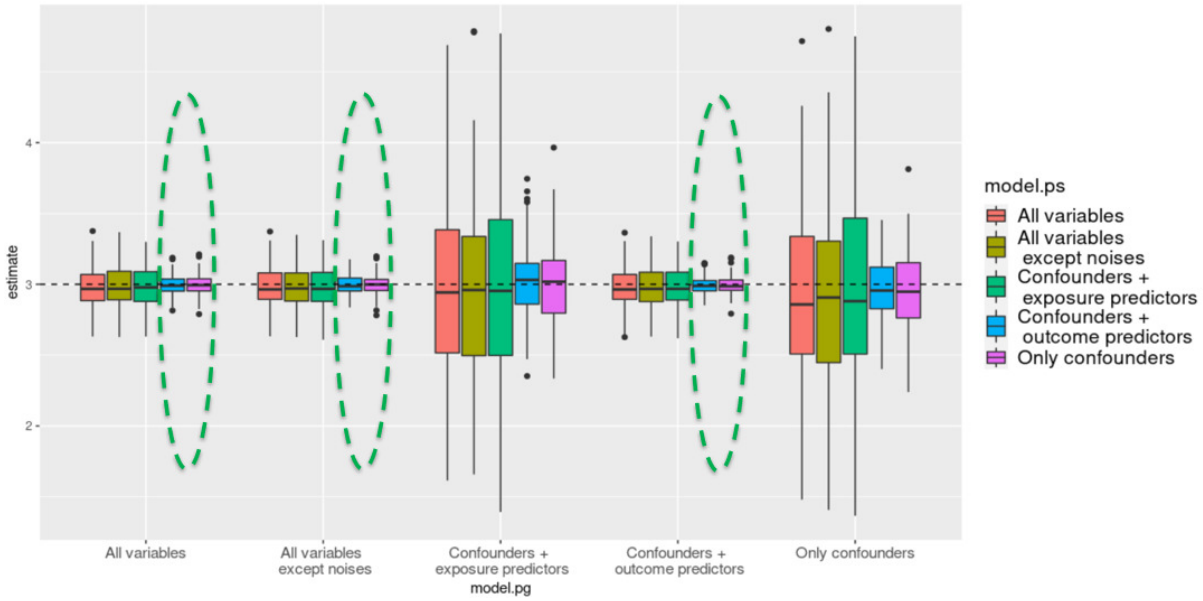


Figure 3.6: Performance of DSM estimator under different variable selection strategies in the illustrating example. Different colors imply different propensity score models, and x-axis differentiate various prognostic score models. Best configurations are marked by green ellipses.

physician surveys and telephone patient interviews. There were three treatment groups based on patients' treatment at initiation: opioid treatments, non-narcotic opioid like treatment, and all other treatmentsPeng et al. (2015). In our simulations, we only compare the first two groups so that the treatment is binary. 16 continuous variables and 11 categorical variables were recorded for each patient, and we only keep the continuous covariates in our simulations for simplicity. Detailed information for each covariate is exhibited in the supplementary material.

Faries *et al.*Faries et al. (2020) implemented the Iman-Conover method to generate simulated REFLECTIONS datasets. Thus, our simulated data did not include any actual data from the REFLECTIONS study. However, the distributions of variables were almost identical in the simulated and real datasets. More importantly, the realistic correlations within the covariates set were retained.

Methods for variable selection: LASSO

From Section 3.3.1, we know that it is important to remove IVs and noise variables when fitting the propensity score model, which is equivalent to omitting variables that are irrelevant to the potential outcome. Thus, it is reasonable to apply variable selection algorithms on covariates with respect to the outcome data. For simplicity, here we used LASSOTibshirani (1996) on the control group data, where we consider all the covariates as predictors and the observed outcomes as response. 10-fold cross-validation was used to select the best tuning parameter λ. We kept two choices of λ: λ_{min} and λ_{1se}, where the first one gave the minimum mean cross-validation error and the latter one was the largest value of λ such that the cross-validation deviance is within 1 standard error of the minimumFriedman et al. (2010). After LASSO regression, we would use covariates with non-zero coefficients to fit scores in the matching methods. To sum up, we had three sets of variables to fit the propensity score or the prognostic score:

- All variables
- Variables selected by λ_{min}
- Variables selected by λ_{1se}

According to its definition, λ_{1se} is more aggressive in removing variables than λ_{min}. As a result, the number of variables used in the selected model decreases when we change from the first strategy to the last strategy. Although more efficient variable selection algorithms could be employed, here we only used LASSO to illustrate the importance of variable selection in the matching methods.

Simulation setup informed by the REFLECTIONS dataset

We generated 100 different REFLECTIONS datasets based on nonparametric sampling from the original dataset, and there were 3000 individuals in each dataset. Although these datasets were not identical, the correlations among covariates were well retained. Each covariate was standardized before fitting the propensity score and prognostic score. A similar generating structure was used to produce the treatment assignment and potential outcomes, as illustrated in Figure 3.7. However, because of the complex correlations between covariates, BMI_B, DxDur, PHQ8_B may not be IVs since they may be correlated with other prognostic variables. In fact, every covariate in this REFLECTIONS dataset is a confounder,

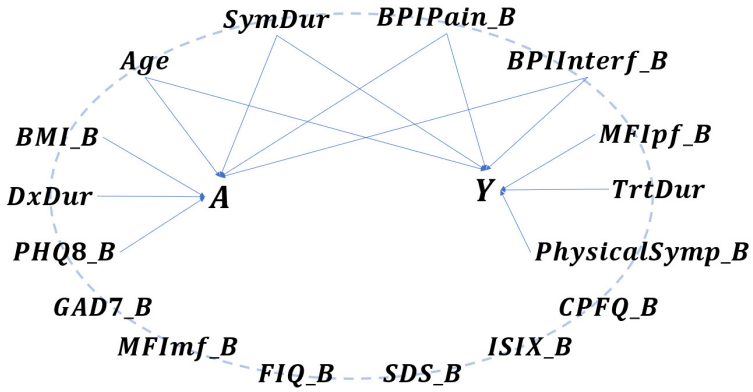


Figure 3.7: Data generation structure of the REFLECTIONS dataset. The dotted ellipse implies complex correlation within the covariates set.

but some of them can be categorized as IVs or noise variables if they are weakly associated with outcome. Moreover, we opted for two different generating processes for both propensity score and prognostic score: a linear model and a nonlinear model, where the latter case is more difficult with possible overfitting issues in the modeling process. We also considered both constant and heterogeneous treatment effect models in our simulation, where the true effect in the first case was 3. Detailed specifications for the heterogeneous effect as well as models for the propensity and prognostic score can be checked in the supplementary materials. To sum up, there were $2 \times 2 = 4$ scenarios to be considered:

- Linear model + constant effect
- Nonlinear model + constant effect
- Linear model + heterogeneous effect
- Nonlinear model + heterogeneous effect

In each scenario, we used PSM, PGM, and DSM to estimate the ATT. All the possible caliper and replacement choices were considered, and as a result, there were $2 \times 2 = 4$ estimators for each matching method:

- Match with a caliper and with replacement
- Match with a caliper and without replacement

- Match without a caliper and with replacement
- Match without a caliper and without replacement

The caliper was always set as 0.25 times the standard deviation of each score. In DSM, at least one score that satisfies the caliper constraint is sufficient for a matching pair, as explained in Section 3.2.3. Same as in the illustrating example, we recorded the estimated ATT from each estimator in each scenario, and we sought to find the best configuration with the smallest bias and variance across different situations. Because we were considering the heterogeneous treatment effect, we calculated the bias of all estimates and made the corresponding box-plots, as shown in the following section.

Results: variable selection in propensity score matching

Figure 3.8 presents the results of all the PSM estimators in different scenarios. Clearly, blue boxes were narrower than red and green boxes in most of the scenarios, indicating that variable selection based on the LASSO algorithm with tuning parameter λ_{1se} may significantly increase the accuracy of the estimator. On the other hand, bias may be induced when we remove the weak confounders too aggressively (depeding on scenarios). In linear model cases, the bias didn't change or even decrease when we used variable selection. This might be because the matching quality was improved when we matched without caliper or the matching rate was increasing when we matched with caliper, see Figure 3.9. However, when the generating process became a complicated nonlinear model and more higher-order or interaction terms were included, we might falsely remove too many weak confounders and the bias was induced, as illustrated in the nonlinear model with constant treatment effect scenario when we match with replacement. Thus, there was a bias and variance trade-off. We think that variable selection based on λ_{1se} is acceptable because 0 was contained in the blue box and thus the bias was not significantly large. It is also worth noting that these were the only two cases when bias increased. In all the other cases the bias decreased or did not change. Thus, we recommend always using LASSO with tuning parameter λ_{1se} to select variables before running PSM.

Results: variable selection in prognostic score matching

As illustrated in Figure 3.10, the performance of PGM was not sensitive to the variable selection strategies. This is consistent with our conclusion in Section 3.3.1 that there is no significant change in the variance unless we falsely remove outcome predictors. To avoid

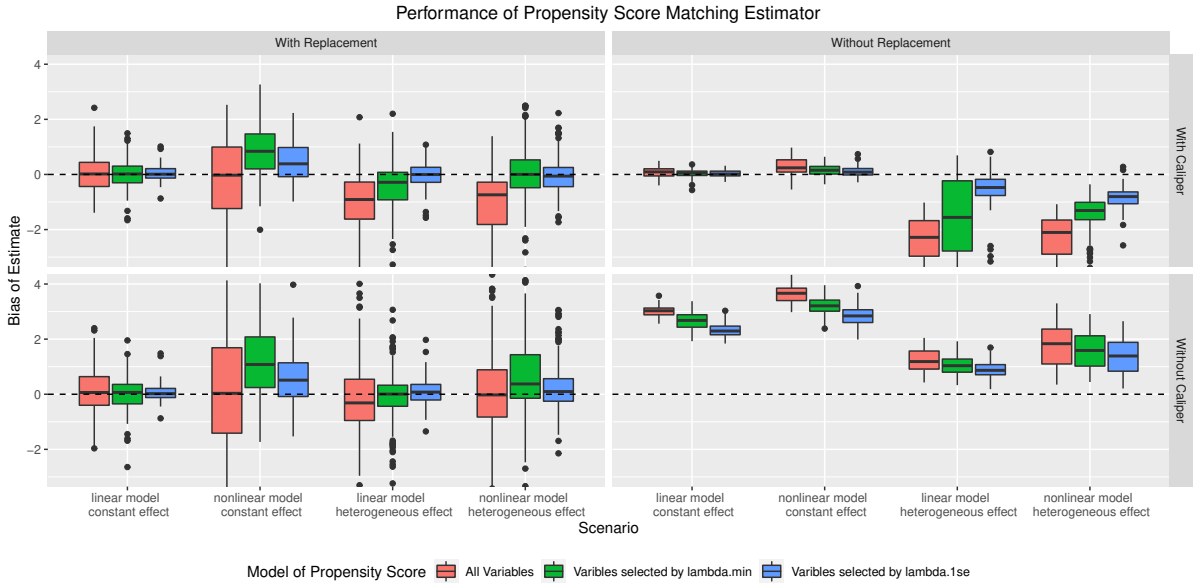


Figure 3.8: Performance of PSM estimator under different variable selection strategies in REFLECTIONS dataset.

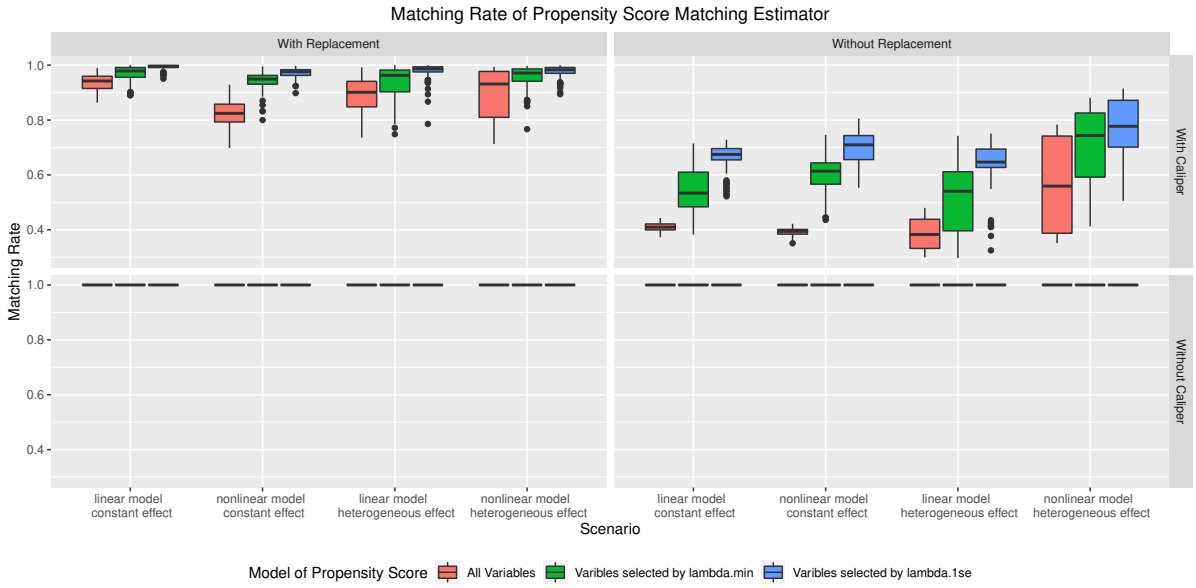


Figure 3.9: Matching rate of PSM estimator under different variable selection strategies in REFLECTIONS dataset.

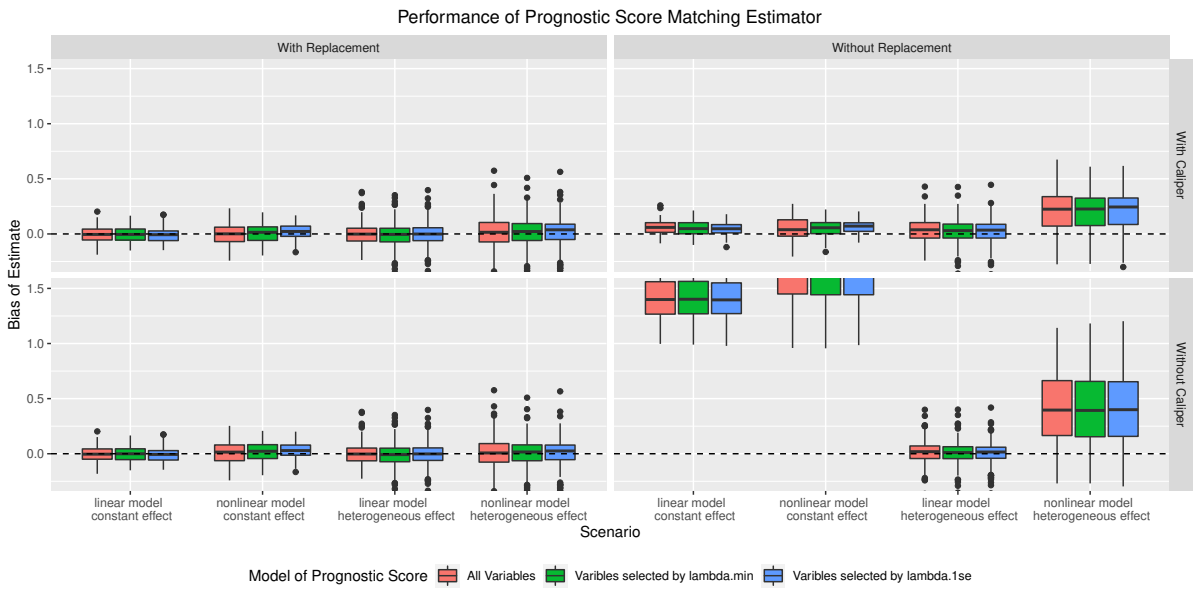


Figure 3.10: Performance of PGM estimator under different variable selection strategies in REFLECTIONS dataset.

this risk, we recommend keeping all the variables when we fit the prognostic score in PGM since there is a potential drawback instead of significant improvement from the variable selection.

Results: variable selection in double score matching

Similar to Section 3.3.1, each score could be fit by the three possible sets of variables. In total, there were $3 \times 3 = 9$ variable selection strategies in double score matching. For simplicity, we only showed the results for matching with a caliper and with replacement in the scenario of a linear generating model and heterogeneous treatment effect, see Figure 3.11. Detailed results are posted in the supplementary materials, and the conclusions are consistent across different scenarios.

Same as the result from Section 3.3.1, variable selection in the propensity score model help reduce the variance, while variable selection in the prognostic score model was not important. From the perspective of bias, it is worth noting that the bias increased as we removed variables from the prognostic score model when we had already selected variables in the propensity score model. The explanation may be that keeping all the variables in the prognostic score model helps avoid the confounding bias induced by removing weak

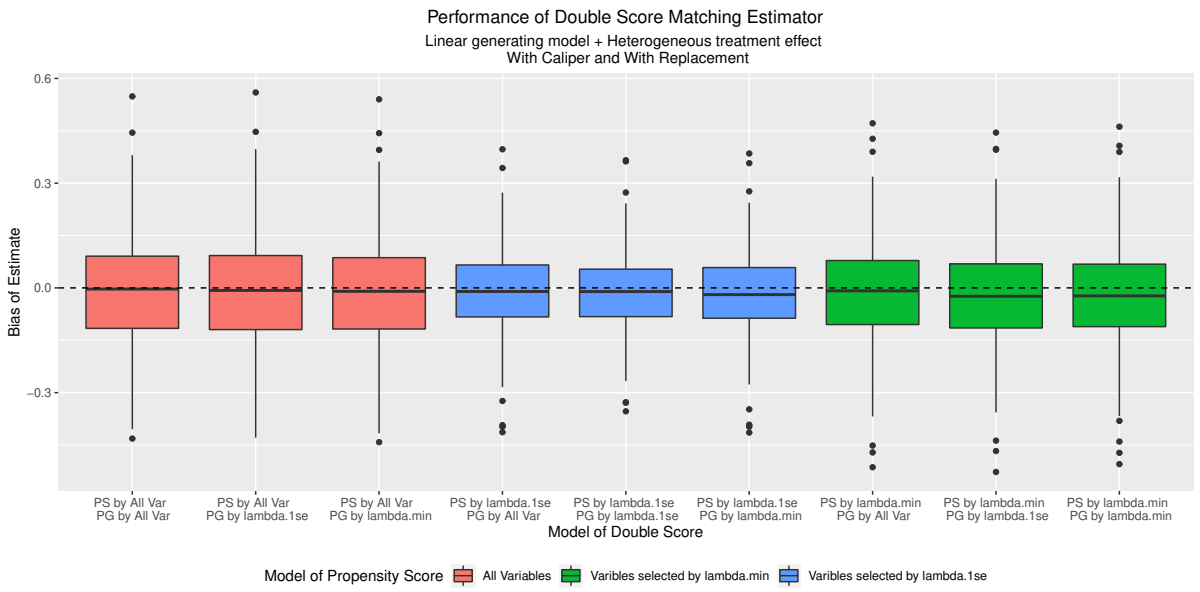


Figure 3.11: Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was linear and the treatment effect was heterogeneous. Only results for matching with a caliper and with replacement are shown here for simplicity.

confounders in the propensity score model, which is a special and important property of DSM. Variable selection in PSM is not recommended by some researchers due to the risk of omitting confounders, but DSM can enjoy the benefits from variable selection without taking the risk of confounding bias. As a result, we strongly recommend applying LASSO with tuning parameter λ_{1se} to select variables for the propensity score model but using all the covariates to fit prognostic score in DSM.

3.4 Choice of caliper and replacement in matching

In Section 4, we explored the best variable selection strategy in the matching methods. In this section, we will find the best choice of caliper and replacement based when combined with best model selection strategy from Section 4. That is, we will only compare PSM estimators where the variables are selected by LASSO based on λ_{1se}. The criterion is still that the bias and variance should be small across various situations.

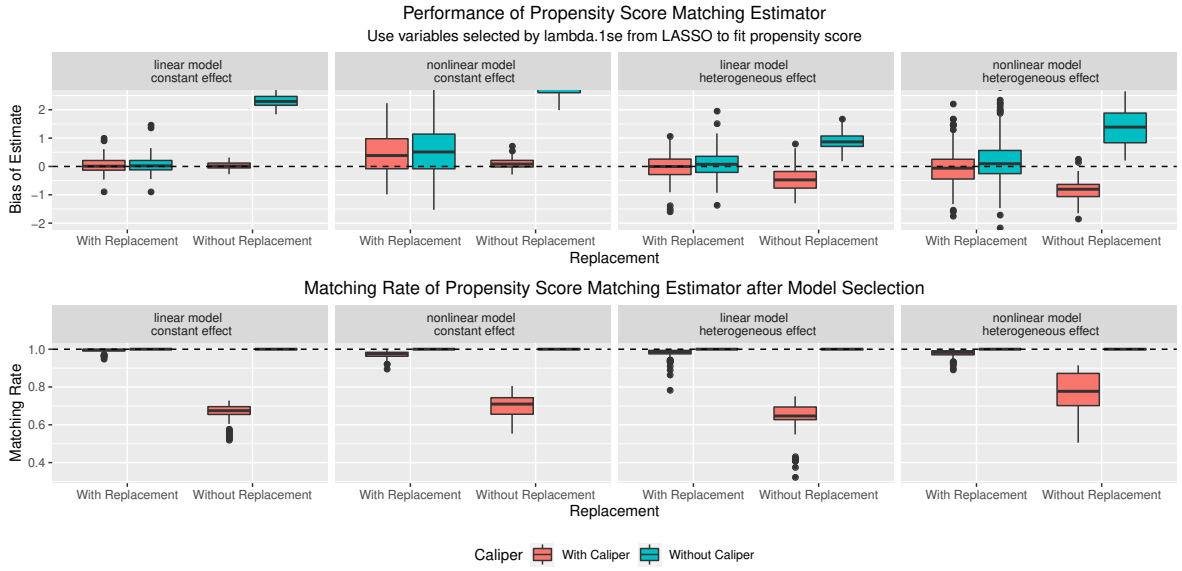


Figure 3.12: Performance and matching rate of PSM estimator under different choices of caliper and replacement in REFLECTIONS dataset, using variables selected by λ_{1se} from LASSO.

3.4.1 Matching constraints in propensity score matching

Figure 3.12 illustrated the performance of PSM estimators under different caliper and replacement configurations. Matching without replacement reduced the variance of the PSM estimator by removing duplicate samples in the after-match dataset, which was consistent with the findings from the literatureBrookhart et al. (2006); Yang et al. (2020); Myers et al. (2011). However, if we matched without a caliper, there was a large bias due to the poor matching quality. If we matched with a caliper, there was still a significant bias when the treatment effect was heterogeneous. This perhaps was because the matched sample from the low matching rate cannot fully represent the target population, and the bias was induced from the difference of estimands. Thus, matching with a caliper and without replacement is only recommended when the constant treatment effect assumption is guaranteed. On the other hand, matching with replacement had a more stable performance because the distance within matching pairs was smaller and the matching rate was higher. Relatively speaking, matching with a caliper could avoid bad matching pairs thus the bias and variance were both smaller. To sum up, we recommend matching with a caliper and with replacement in PSM unless the treatment effect is constant.

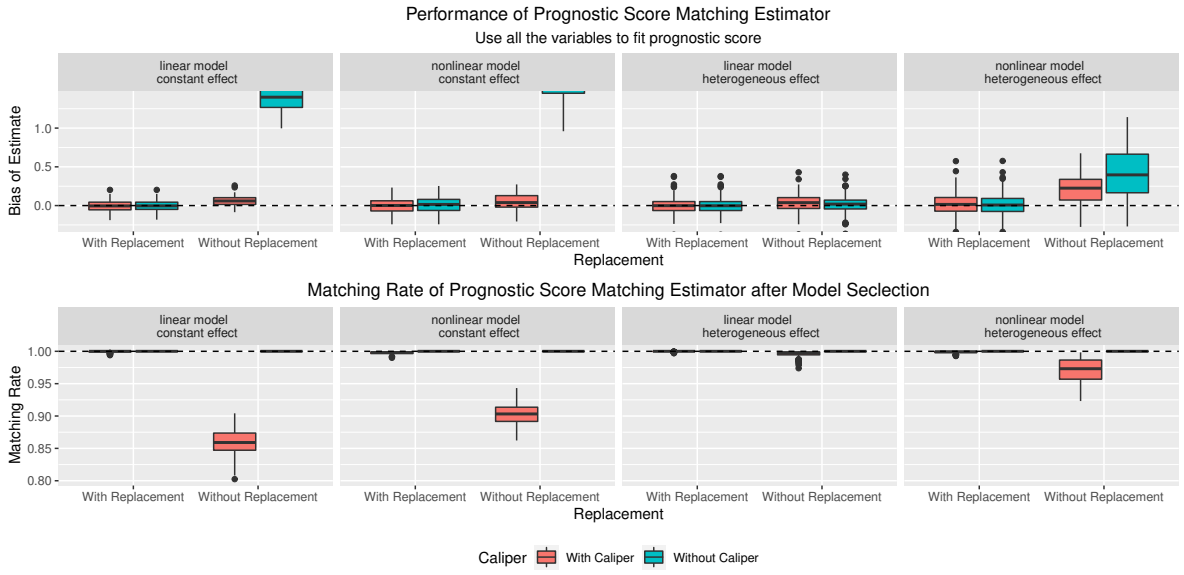


Figure 3.13: Performance and matching rate of PGM estimator under different choices of caliper and replacement in REFLECTIONS dataset, all the variables were used.

3.4.2 Matching constraints in prognostic score matching

As shown in Figure 3.13, matching without replacement was still accompanied by possible bias in PGM. Even in scenarios with constant treatment effect, the bias was still higher compared to matching with replacement. However, in PGM, there is no significant difference between matching with and without a caliper if we match with replacement. Thus, we only need matching with replacement to be guaranteed in PGM.

3.4.3 Matching constraints in double score matching

Compared to PSM and PGM, DSM was not sensitive to the choice of caliper and replacement, as illustrated in Figure 3.14. Interestingly, matching without caliper and replacement in DSM had very stable performance across all the scenarios, while PSM and PGM performed extremely badly in this setting. When the generating model was nonlinear and the treatment effect was heterogeneous, adding caliper even increased the bias if we matched without replacement. This is also different from our results in Section 3.4.1 and 3.4.2. Matching without replacement may slightly reduce the variance, while matching with replacement may slightly reduce the bias. However, matching without caliper and replacement may

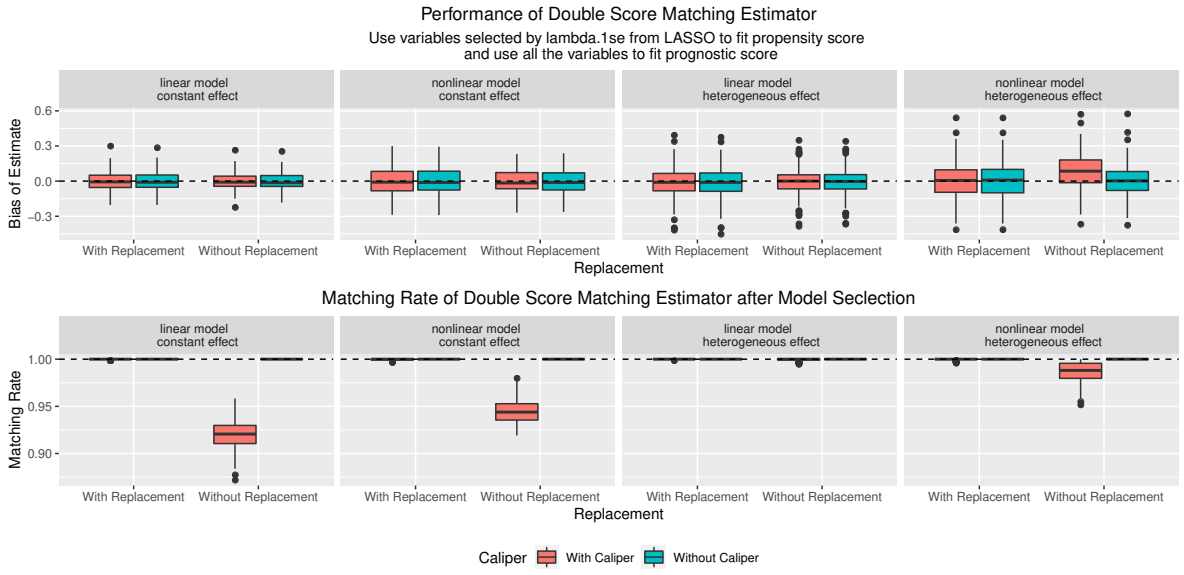


Figure 3.14: Performance and matching rate of DSM estimator under different choices of caliper and replacement in REFLECTIONS dataset, using variables selected by λ_{1se} from LASSO to fit propensity score and all the variables to fit prognostic core.

weaken the double robustness of DSM, as shown in the supplementary materials. As a result, we recommend matching with replacement in DSM. Similar to PGM, caliper didn't make a significant difference in DSM.

3.5 Comparison among matching estimators

At this point, we have found the best choice of caliper, replacement, and variable selection strategy for PSM, PGM, and DSM. Now we are interested in comparing these three matching methods based on each estimator's best configuration. We will use the same four simulation scenarios as we showed in Section 3.3.2. However, here we also consider model specification problems. We start with comparing the performance of different matching estimators under correct model specification, and then results for different model misspecification scenarios will be presented in the following section.

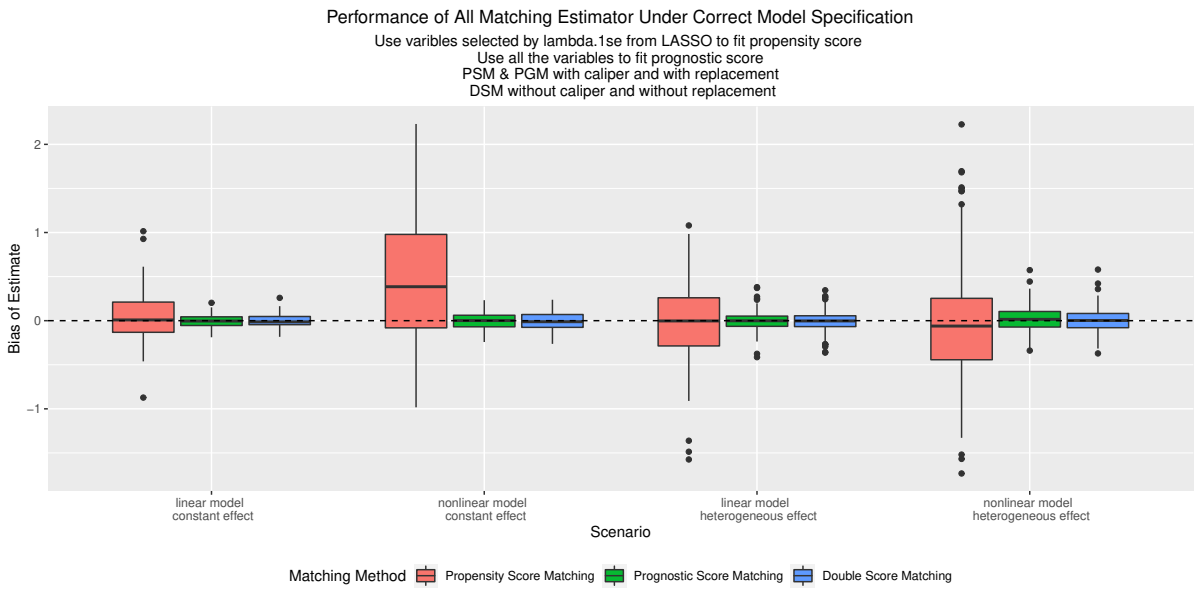


Figure 3.15: Comparison of PSM, PGM, and DSM when both models were correctly specified.

3.5.1 Comparison under correct model specification

Results under correct model specification are illustrated in Figure 3.15. In all the four scenarios, PGM and DSM achieved much smaller variance than PSM with negligible bias, suggesting that in practice PSM should be replaced by PGM or DSM if the outcome information is available. Meanwhile, PGM and DSM had very similar performance across different situations. This is different from Leacy and Stuart's resultLeacy and Stuart (2014) that PGM had a much smaller variance than DSM. The reason for superior performance of DSM in our simulations is that we used the variable selection strategy to improve the accuracy of DSM. The following section will show the advantage of DSM over PGM when models are misspecified.

3.5.2 Comparison under different model specifications

In this section we consider four different model specifications:

- Both models are correctly specified
- Both models are misspecified

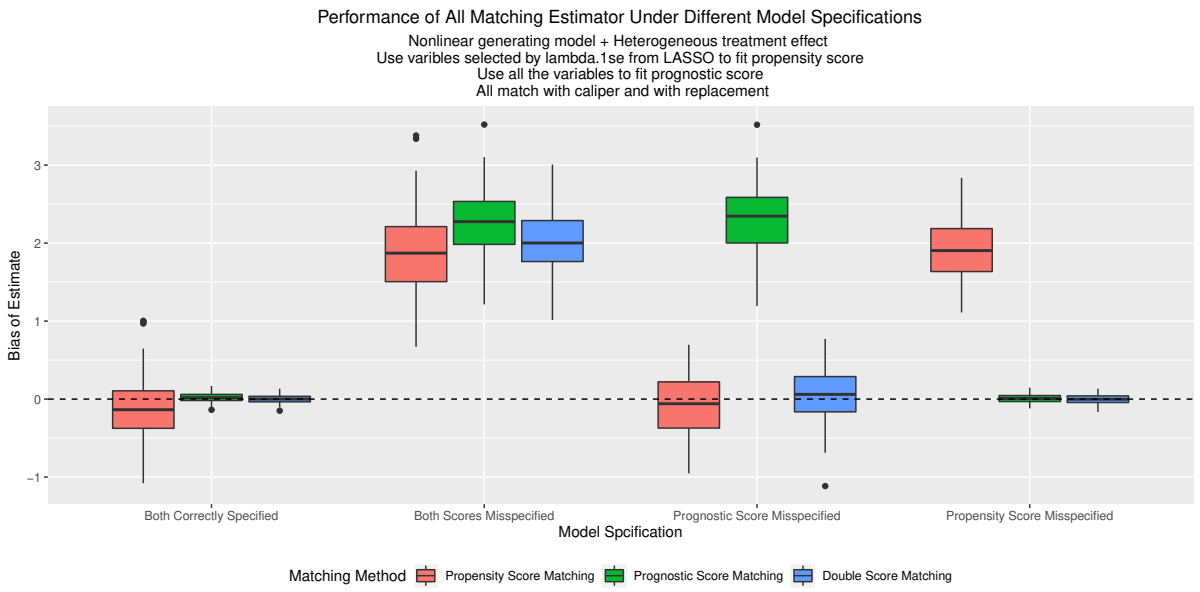


Figure 3.16: Comparison of PSM, PGM, and DSM when both models might be misspecified.

- Only prognostic score model is misspecified
- Only propensity score model is misspecified

In our simulation, model misspecification meant that a linear model was used in the fitting process while the true model was nonlinear. The set of variables included in the fitting model remained the same. Figure 3.16 presents the performance of the three matching estimators under different model specifications when treatment effect was heterogeneous. PSM and PGM suffered from huge bias if the related score was misspecified, while DSM still exhibited strong performance if only one score was misspecified. This is consistent with the double robustness property of DSM established in the literature. However, the estimators considered here take variable selection into consideration, indicating that variable selection strategy does not compromise the double robustness property of DSM.

3.6 Conclusions and discussion

In this article, we explore the statistical and numerical properties of PSM, PGM, and DSM via extensive simulation studies. Based on the simulation results, we summarize the findings

for each matching estimator with respect to choices of variable selection strategy, caliper, and replacement:

- PSM can be combined with the variable selection strategy of removing IVs from the propensity score model for efficiency consideration; matching can be done with a caliper and with replacement.
- PGM does not necessarily require variable selection in the sense that variable selection has little impact on efficiency, and matching can be done with replacement, while whether or not using caliper is immaterial.
- DSM can be combined with the variable selection strategy of removing IVs from the propensity score model; matching can be done with replacement, while whether or not using caliper is inconsequential.

Among the three matching schemes, DSM enjoys a double robustness property in that its consistency requires either the propensity score model or the prognostic score model to be correctly specified, in stark contrast to PSM and PGM, indicating that DSM is less vulnerable to model misspecification. Meanwhile, Section 3.5.1 showed that DSM was the most efficient estimator when models were correctly specified. As a result, DSM is the most robust and efficient matching estimator across different situations. Again, these findings may be restrictive to the simulation settings in the paper, where outcomes and covariates are generated from multivariate normal distributions, therefore our recommendations may not apply to other scenarios.

Intrinsically, DSM and PGM require outcome information to fit the prognostic score model, which violates Rubin's principle of designing studies without outcomeRubin (2007). The same issue arises when the variable selection strategy is incorporated to remove IVs in PSM. The separation of the design and analysis stages helps establish the credibility of the studies by avoiding the potential for selecting the balancing score based on the outcome results the researcher desires. However, our study showed that when outcome information is available or in retrospective studies, using the outcome information in matching can boost efficiency. Moreover, only the control arm outcome is used when estimating the ATT, and thus outcome for the treatment of interest can remain unobserved. This also safeguards the separation principle and prevents potential data dredging. In prospective studies, prior information from previous research can be used to construct the prognostic score when outcome information is not availableRubin (2001).

The variable selection result for PSM is consistent with the finding from Brookhart *et al.*Brookhart et al. (2006) and Myers *et al.*Myers et al. (2011) suggesting that including IVs into the propensity score model may amplify bias and variance of the causal effect estimator. However, different from the suggestion from Myers *et al.*Myers et al. (2011) that variable selection should not be used in consideration of confounding bias, we still recommend to use the variable selection procedure to increase accuracy but using DSM instead of PSM. The prognostic score works as a protection against confounding bias in DSM since all the variables are included in the prognostic score model, implying another important advantage of DSM. In other words, DSM changes a trade-off problem into a win-win situation. However, this may to some extent compromise the double robustness of the DSM estimator when prognostic variables are difficult to identify, such as in studies involving rare outcomes and high-dimensional covariates. It is likely that the variable selection algorithm may remove some important variables improperly, leading DSM to lose its double robustness. From this perspective, there is still a trade-off between efficiency and robustness. We plan to address this issue in our future work. Moreover, although we used LASSO with tuning parameter λ_{1se} to select variables in PSM, we believe that there should be better ways to remove IVs, such as the machine learning algorithms like random forest and neural network. Besides, soft variable selection strategies may also be used in matching methods. For example, Tang *et al.*Tang et al. (2020) used causal ball screening to assign higher weight to variables that are more related to outcome. Although this work was based on the weighting framework, the idea may be applied to matching estimators as well. Since all the variables are kept in the model, it is likely that the double robustness is always guaranteed even in difficult situations such as high-dimensional covariates and rare outcomes. The ball-covariance approach proposed by PanPan et al. (2019) was recently used by Zhao and YangZhao and Yang (2021) to select outcome predictors in the matching context. This may also be utilized in our setting. We hope to see more developed variable selection algorithms designated for the matching methods.

We suggest using all the variables in the prognostic score model in PGM and DSM. This conclusion is based on a simulation study with 10-100 predictors and 3000 observations. In some extreme scenarios where rare outcomes and high-dimensional covariates are involved, it is necessary to apply some variable selection algorithms since it is impossible to include all the variables in the prognostic score model. The performance may be improved by using more advanced variable selection methods than LASSO, which will be investigated in our future work.

The discussion of caliper was very limited in this paper: we only compared caliper of

0.25 standard deviation versus no caliper. Note that the latter case could be seen as a caliper with infinite width. As a result, we were simply considering two special values of caliper width. A more comprehensive study on the choice of caliper can be done in the future. Austin and Wang *et al.*'s studiesAustin (2011); Wang et al. (2013) are good examples, but a proper variable selection strategy should be applied before matching. Its effect on PGM and DSM should also be studied thoroughly. As far as we observed in our simulations, PGM and DSM were not sensitive to the choice of caliper, and we are interested that whether this will hold for a larger range of caliper widths. For the choice of replacement, the bias observed in matching without replacement under heterogeneous treatment effects may come from the estimation of ATT. When we change the target population from the treatment group to the matched population defined by the matching weights or overlap weights based on the propensity score, matching without replacement may not be biased anymoreLi and Greene (2013); Li et al. (2018); Matsouaka and Zhou (2020). We also acknowledge that the dichotomy in the paper oversimplified the comparison of the existing matching methods. Besides matching with and without replacement that are commonly used, algorithms such as full matching and full matching with structural restrictions are available in some specialized softwareHansen and Klopfer (2006). In light of this, researchers may choose from a wider range of matching methods. More work can be done in the future to provide practical recommendations for a wider spectrum of matching methods.

Although our final recommendation is DSM, PSM can be still useful when the outcome information is unavailable. In contrast to the usual recommendation that matching should be done with a caliper and without replacement to increase accuracy, we suggest matching with replacement to reduce potential bias from heterogeneous treatment effects. This is a bias-and-variance trade-off, but our simulation showed that the negative impact of bias can be more significant than than the problem of variance. Using variable selection strategies may compensate for the loss of accuracy. Despite the fact that the propensity score model is misspecified after removing IVs, the estimator remains consisten. This can be explained by the covariate scores proposed by WaernbaumWaernbaum (2012). What matters is the conditional independence between the potential outcomes and the treatment assignments, which can still hold despite the lack of balance in some covariates across treatment arms.

Our study is limited to the estimation of ATT. When estimating the ATE, multiple prognostic scores are necessary for PGM and DSM. Similarly to Corollary 2, Yang and ZhangYang and Zhang (2020) compared the three estimators for the ATE in their supplementary materials, but they only proved that DSM can be more efficient than PSM when estimating marginal means instead of the ATE. Future work will investigate the performance when

variable selection is considered in estimating treatment effects via DSM. However, the definition of IVs may be tricky here. For example, a variable may be irrelevant to outcome on the control group but important to outcome on the treated group. As a result, different sets of covariates should be used when estimating each prognostic score. It is not clear how much of a problem this would cause if one just used the prognostic score for the control arm. To deal with this problem, we recommend using the adapted DSM by Yang and ZhangYang and Zhang (2020). Instead of directly estimating the ATE, the marginal mean for each treatment level is estimated, where only the propensity score and the prognostic score for that particular level of the treatment need to be adjusted for. Thus, we can duplicate the procedure of ATT estimation considered in this paper, and different sets of IVs are excluded when estimating the averages of the potential outcomes for the treatment group and control group. We expect to see both theoretical and numerical results on this problem, including the effect of variable selection and the comparison among the three matching estimators. Another limitation is that we were not changing the overlap in propensity score or prognostic score in order to see how it is affecting the DSM performance. But it is worth noting that the prognostic score has good overlap insensitive to model specifications. Moreover, our simulation study is limited to the continuous outcomes. Binary outcomes as well as survival data can be investigated similarly in the future. Furthermore, presence of nomial covariates, especially with rare levels, may also change the conclusions in this paper since our simulations are based on continuous baseline covariates. These kinds of problems are worth investigating in future research. Researchers should be aware of these limitations and should be cautious when applying our recommendations to their settings. Nevertheless, we believe that our simulation study based on REFLECTIONS dataset, where variables come from non-normal distributions and exhibited complex correlation structures, is an improvement compared to the existing simulation studies that rely on independent normally distributed variables.

CHAPTER

4

ON THE INFLUENCE FUNCTION OF THE DOUBLY ROBUST REGRESSION ESTIMATOR

4.1 Introduction

Causal inference has received more and more attention in the last decade. In many areas such as biology, social science, and education, researchers are interested in observational studies because of the increasing accessibility of big data. However, due to the existence of confounders, the treatment assignment may be correlated to the potential outcomes, leading to a potential confounding bias of the treatment effect estimator. Thus, statisticians have been working on developing various estimators for the average treatment effect (ATE). Among them, weighting estimators, matching estimators, and regression estimators are widely used in practice. The inverse probability weighting (IPW) estimator is practically appealing by its simple construction: the weights are constructed by inverting the propensity scores, but a more popular choice is the augmented inverse probability weighting

(AIPW) because it is semi-parametrically efficient and doubly robust, i.e., the estimator is consistent if either one of the propensity score or outcome model is correctly specified (Robins et al. 1994; Scharfstein et al. 1999). However, weighting estimators may be unstable when propensity scores are close to zero or one (Kang and Schafer 2007; Guo and Fraser 2014). On the other hand, matching estimators as a hot dock imputation method is invulnerable to extreme propensity scores (Frölich 2004). More importantly, they are intuitively appealing by mimicking a randomized experiment (Dehejia and Wahba 1999; Heckman et al. 1997b). However, matching may suffer from the curse of dimensionality, and thus researchers choose to match on the balancing scores instead of matching on high-dimensional covariates directly. Rosenbaum and Rubin (1983b) proposed the popular propensity score matching (PSM) estimator, and Hansen (2008) proposed an alternative called the prognostic score matching (PGM) estimator. To increase robustness, researchers developed the double score matching (DSM) (Leacy and Stuart 2014; Antonelli et al. 2018; Yang and Zhang 2020) estimator by combining the two scores, and thus the estimator can be doubly robust similar to AIPW. Nevertheless, none of the matching estimators achieve the semiparametric efficiency bound. On the contrary, the outcome regression estimator is easy to implement and can be very efficient if the parametric model is correct, but the result may be biased if the outcome model is misspecified.

To improve the robustness of regression estimators, researchers have been working on incorporating the treatment information, i.e., the propensity score, into the outcome regression model. Little and An (2004) were the first in the missing data literature to propose the penalized spline of propensity prediction (PSPP) by using the idea of partial linear regression to obtain a doubly robust regression estimator. Zhang and Little (2009) later simplified PSPP by allowing non-centered regressors and extended it to the inference of condition means. The estimator is further extended to causal inference by Zhou et al. (2019) as the penalized spline of propensity methods for treatment comparison (PENCOMP) to estimate the average treatment effect. However, all of the existing results focused on the consistency, double robustness, and numerical performance of the estimators, and none of them have derived the asymptotic variance because partial linear regression brings great complexity.

Partial linear regression has been well studied in the nonparametric statistics literature. The response variable Y is assumed to be linear in X and nonlinear in T: $E(Y \mid X, T) = X^{\mathrm{T}}\beta + g(T)$. Engle et al. (1986) proposed a penalized likelihood estimator quantify the relationship between weather and electricity sales. Heckman (1986) further worked on this spline smoothing estimator and derived its asymptotic distribution when T is a known

scalar. Robinson (1988) generalized this result to where T is stochastic and of arbitrary dimension. He proposed a two-step estimator: the conditional expectations are nonparametrically estimated, for example, using kernel regression estimators, and beta can then be solved from a linear regression model constructed by the residuals of Y and X. Because T is considered stochastic, it may be related with X. Rice (1986) pointed out that when X and T are not independent, the estimate of the parametric component may be biased unless undersmoothing the nonparametric component. Chen (1988) further studied this correlation between X and T and gave explicit conditions to achieve the usual parametric rate for the least-squares estimator $\hat{\beta}$, though the result assumed a piecewise polynomial estimate of g. Chen and Shiau (1991) proposed a two-stage spline smoothing method without undersmoothing the nonparametric component and later derived its asymptotic results when the smoothing parameter is chosen by the generalized cross validation (Chen and Shiau 1994).

The unique challenge in applying partial linear regression to causal inference and missing data literature is that X and T are functionally dependent. T is not directly observed but estimated based on X, so-called the "generated regressor" problem. This has piqued the interest of statisticians and economists in recent years. Pagan (1984) studied the generated regressors in traditional parametric models. Heckman et al. (1998) suggested applying nonparametric regression on the estimated propensity scores, thus leading to a generated regressor problem in nonparametric settings. Hahn and Ridder (2013) denoted this as the HIT estimator and showed that it is as efficient as the imputation estimator, which is based on a nonparametric regression on the covariates directly. To illustrate this, he provided a series of general theorems to assess the effect of the estimated regressors on the three-step estimator that can be generated by the following steps: first, estimate the generated regressors either parametrically or nonparametrically; second, run a nonparametric regression on the regressors; third, estimate the finite-dimensional parameter of interest. The paper extensively applied the pathwise derivative approach from Newey (1994) and the orthogonality conditions for the conditional expectations to quantify the effects of the estimated regressors and nonparametric estimations. Similarly, Mammen et al. (2012) derived the stochastic expansion of the nonparametric regression estimator with a nonparametrically generated regressor. The authors further provided general theorems to verify needed conditions and derived the asymptotic variance for semiparametric M-estimator with generated covariates(Mammen et al. 2016). Lee (2018) later extended the asymptotic results for the kernel regression estimator on generated regressors to the continuous treatment scenario in the partial mean processes. These results are all based on the two-step or three-step

estimator, where the response variable is assumed to be a nonparametric function of the generated regressor. However, the partial linear regression does not directly fit in these scenarios because the outcome is assumed to include not only the nonparametric component but also a parametric term where the coefficients are estimated via an ordinal least square (OLS) regression on the residuals determined by the generated regressors. This additional step makes the partial-linear-regression-based estimator on the ATE being a four-step estimator and brings new difficulties to the analysis.

In this paper, we formalize the framework of the doubly robust regression (DReg) estimator by generalizing PENCOMP to a wider range of nonparametric methods, such as kernel estimators and smoothing splines. We show that the DReg estimator is doubly robust, i.e. consistent if either the propensity score or the outcome model is correctly specified. The key result in our paper is that we obtain the asymptotic results of the DReg estimator by deriving its influence function. The unique challenge compared to the existing work from Hahn and Mammen is the additional step in the generate regressor problem in partial linear regression because the OLS estimator in the linear component will be affected by both the propensity score parameter and the nonparametric estimation via a complicated process from M-estimation. We provide a few propositions to quantify this dependence by utilizing the orthogonality conditions, Hahn's decomposition, and Newey's pathwise derivative.

The remaining part of the paper proceeds as follows. Section 4.2 states the basic notations and assumptions. Section 4.3 introduces the framework of partial linear regression and describe the algorithm of DReg estimator. The challenges in theoretical analysis, key ideas, and important propositions are also illustrated in this section. Section 4.4 includes the main results of the paper. Section 4.5 compares the DReg estimator with other existing estimators in a simulation study, and Section 2.9 concludes.

4.2 Notation and assumptions

Let X_i be a p-dimensional vector of pre-treatment covariates, A_i the binary treatment, and Y_i the outcome for unit $i = 1, \ldots, n$. We follow the potential outcomes framework. Let $Y_i(a)$ be the potential outcome had unit i been given treatment a ($a = 0, 1$). Based on the potential outcomes, the ATE is $\tau = \mathbb{E}\{Y_i(1) - Y_i(0)\}$ and the ATT is $\tau_{\text{ATT}} = \mathbb{E}\{Y_i(1) - Y_i(0) \mid A_i = 1\}$. The observed outcome is $Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i) Y_i(0)$. We assume that $\{X_i, A_i, Y_i(0), Y_i(1)\}$, $i = 1, \ldots, n$, are independent and identically distributed. Thus, (X_i, A_i, Y_i), $i = 1, \ldots, n$,

are also independent and identically distributed.

For simplicity of exposition, for a generic variable V, denote

$$\mu_a(V)=\mathbb{E}\{Y(a)\mid V\},\ \sigma_a^2(V)=\mathbb{V}\{Y(a)\mid V\},\ e(V)=\mathbb{P}(A=1\mid V),$$

where $\mu_a(V)$ is an outcome mean function, $\sigma_a^2(V)$ is a variance function, and $e(V)$ is the propensity score.

To identify the causal effects, we make the standard positivity and treatment ignorability assumptions (Rosenbaum and Rubin 1983b).

Assumption 3 *There exist constants c_1 and c_2 such that $0<c_1\le e(X)\le c_2<1$ almost surely.*

Assumption 4 *$\{Y(0),Y(1)\}\perp\!\!\!\perp A\mid X$, where $\perp\!\!\!\perp$ means "independent of".*

Assumption 3 implies a sufficient overlap of the covariate distribution between the treatment groups. Generally, Assumption 4 can be made plausible by collecting detailed information on characteristics of the units that are related to treatment assignment and outcome. As a result, the dimension of X may be high.

Under Assumptions 3 and 4,

$$\tau=\mathbb{E}\left\{\mu_1(X)-\mu_0(X)\right\}, \tag{4.1}$$

are identifiable from the joint distribution of the observed data (A,X,Y).

The seminal paper (Rosenbaum and Rubin 1983b) showed the vital role of the propensity score as a balancing score.

Lemma 3 (Propensity score as a balancing score; Rosenbaum and Rubin, 1983) *Under Assumptions 3 and 4, $\{Y(0),Y(1)\}\perp\!\!\!\perp A\mid e(X)$.*

Lemma 3 implies that

$$\tau=\mathbb{E}\left[\mu_1\{e(X)\}-\mu_0\{e(X)\}\right]. \tag{4.2}$$

Then we can estimate τ through propensity score matching, subclassification or weighting (e.g., Imbens and Rubin 2015) or regression based on $e(X)$ (Hahn and Ridder 2013).

Combining (4.1) and (4.2) leads to

$$\tau=\mathbb{E}\left[\mu_1\{X,e(X)\}-\mu_0\{X,e(X)\}\right].$$

Based on this result, Zhou et al. (2019) proposed a regression estimator of the average treatment effects adjusting for X and $e(X)$. Because $e(X)$ is unknown in observational studies, a prior step to estimating $e(X)$ is required. They showed that the resulting estimator is doubly robust in that it is consistent if the relation of Y and X or the propensity score model of $e(X)$ is correctly specified. Because X and $\hat{e}(X)$ may be collinear, Zhou et al. (2019) proposed a penalized spline approach to fit $\mu_a\{X, \hat{e}(X)\}$ and a subsequent multiple imputation estimator for treatment effects. Promising numerical performance was shown in their paper; however, to the best of our knowledge, large sample properties have not been established in the literature. The essentiality of the asymptotic properties motivate our work in this paper.

4.3 Partial linear regression

4.3.1 Framework

We assume a semiparametric partially linear regression model

$$Y_i(a) = X_i^{\mathrm{T}}\beta_a + g_a\{e(X_i)\} + \epsilon_i(a), \tag{4.3}$$

where $\mathbb{E}\{\epsilon_i(a) \mid X_i\} = 0$, for $a = 0, 1$. In other words, the conditional mean of $Y_i(a)$ is linear in X_i and possibly nonlinear in $e(X_i)$. Although we specify a linear relationship between $Y_i(a)$ and X_i, any parametric relationship with a finite dimension of parameters can be imposed.

To estimate β_a and $g_a(\cdot)$, we resort to the typical procedure in partial linear regression. Define residuals $r_{V,a}\{e(X)\} = V - \mathbb{E}\{V \mid e(X), A = a\}$ and $\hat{r}_{V,a}\{e(X)\} = V - \hat{\mathbb{E}}\{V \mid e(X), A = a\}$. Here V can be either X or Y. The conditional expectations are estimated via nonparametric methods. Then, the linear regression holds for $r_{Y,a}\{e(X)\}$ and $r_{X,a}\{e(X)\}$,

$$r_{Y,a}\{e(X_i)\} = r_{X,a}\{e(X_i)\}\beta_a + \epsilon_i(a) \tag{4.4}$$

for all subjects receiving treatment a. Then, a least squared estimator $\hat{\beta}_a$ can be obtained based on $\{\hat{r}_{Y,a}\{\hat{e}(X_i)\}, \hat{r}_{X,a}\{\hat{e}(X_i)\}; A_i = a, i = 1, \ldots, n\}$, given by

$$\hat{\beta}_a = \left[\hat{r}_{X,a}\{\hat{e}(X)\}^{\mathrm{T}}\hat{r}_{X,a}\{\hat{e}(X)\}\right]^{-1}\hat{r}_{X,a}\{\hat{e}(X)\}^{\mathrm{T}}\hat{r}_{Y,a}\{\hat{e}(X)\}. \tag{4.5}$$

Upon obtaining $\hat{\beta}_a$, standard kernel regression can be used to estimate $g_a(\cdot)$ by smoothing

$Y_i - X_i^{\mathrm{T}}\hat{\beta}_a$ over $\hat{e}(X_i)$, restricted to $A_i = a$. However, because we have already estimated $\mathbb{E}\{X \mid \hat{e}(X), A = a\}$ and $\mathbb{E}\{Y(a) \mid \hat{e}(X), A = a\}$ nonparamatrically, we can obtain the estimate $\hat{g}_a(\cdot)$ directly from the two condition expectations. Take conditional expectations on $\hat{e}(X_i)$ and $A = a$ in (4.3),

$$\mathbb{E}\{Y \mid e(X_i), A = a\} = \mathbb{E}\{X \mid e(X_i), A = a\}^{\mathrm{T}} \beta_a + g_a\{e(X_i)\}.$$

Thus, $\hat{g}_a(\cdot)$ can be obtained by $\hat{\beta}_a$ and the two nonparametric estimates of the condition expectations, given by

$$\hat{g}_a\{\hat{e}(X_i)\} = \hat{\mathbb{E}}\{Y \mid \hat{e}(X_i), A = a\} - \hat{\mathbb{E}}\{X \mid \hat{e}(X_i), A = a\}^{\mathrm{T}} \hat{\beta}_a. \tag{4.6}$$

In summary, the DR regression estimator proceeds as follows.

Step 1. (*Propensity score estimation*) Fit a propensity score model $e(X;\alpha)$, where α is a vector of finite dimensional parameters, e.g., a logistic regression model, and obtain $\hat{e}(X_i) = e(X_i;\hat{\alpha})$. Typically, $\hat{\alpha}$ can be solved from the estimating equation

$$\sum_{i=1}^{n} \psi_e(X_i, A_i; \alpha) = \sum_{i=1}^{n} \frac{\partial e(X_i;\alpha)}{\partial \alpha} \cdot \frac{A_i - e(X_i;\alpha)}{e(X_i;\alpha)\{1 - e(X_i;\alpha)\}} = 0.$$

Step 2. (*Partial linear regression*) Fit the partial linear regression model (4.3) based on $\{Y_i, X_i, \hat{e}(X_i) : A_i = a, i = 1, \ldots, n\}$, separately for $a = 0, 1$. Estimate $\mathbb{E}\{X \mid \hat{e}(X), A = a\}$ and $\mathbb{E}\{Y \mid \hat{e}(X), A = a\}$ nonparamatrically (e.g., kernel estimator or spline smoother) and estimate β_a by (4.5) .

Step 3. (*ATE estimation*) Estimate $\mu_a = \mathbb{E}\{Y(a)\}$ by

$$\begin{aligned}
\hat{\mu}_a &= \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_a\{X_i, \hat{e}(X_i)\} = \frac{1}{n}\sum_{i=1}^{n}\left[X_i^{\mathrm{T}}\hat{\beta}_a + \hat{g}_a\{\hat{e}(X_i)\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[X_i^{\mathrm{T}}\hat{\beta}_a + \hat{\mathbb{E}}\{Y \mid \hat{e}(X_i), A = a\} - \hat{\mathbb{E}}\{X \mid \hat{e}(X_i), A = a\}^{\mathrm{T}}\hat{\beta}_a\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{r}_{X,a}\{\hat{e}(X_i)\}\hat{\beta}_a + \hat{\mathbb{E}}\{Y \mid \hat{e}(X_i), A = a\}\right] \qquad (4.7)
\end{aligned}$$

for $a = 0, 1$. The ATE estimator is $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$.

Remark 4 (Difference from PENCOMP) *The proposed DR regression estimator is highly related to the PENCOMP estimator proposed by Zhou et al. (2019). On one hand, the DR regression estimator can be treated as a special case of the PENCOMP estimator by setting the parametric component as linear, though any other finite-dimensional parametric model can be applied to the DR regression estimator as well. The specification of the linear component simplifies the analytical derivation of influence functions, albeit still complicated. On the other hand, the PENCOMP estimator uses the penalized spline model with the truncated polynomial basis to fit the conditional expectations, which is difficult to analyze theoretically to the best of our knowledge, while our DR regression estimator admits other consistent nonparametric methods that satisfy some regularity conditions, for example, kernel methods.*

4.3.2 Challenges in theoretical analysis and key ideas

To show the challenges in theoretical analysis, we focus on deriving the asymptotic results for $\hat{\mu}_1$ because the same theoretical analysis can be done for $\hat{\mu}_0$ and thus $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$. Denote $A_n \cong B_n$ as $A_n = B_n + o_{\mathbb{P}}(1)$. Denote $\hat{\mu}_1\{X_i, e(X_i)\}$ as the regression estimator of $\mu_1\{X_i, e(X_i)\}$ when $e(X_i)$ is known. Denote $\alpha^* = \text{plim}\hat{\alpha}$ and $e^*(X) = e(X;\alpha^*)$ as the probability limit of $\hat{e}(X)$, which may be different from $e(X)$ if the propensity score model is misspecified.

First, we decompose $n^{1/2}\left(\hat{\mu}_1 - \mu_1\right)$ into components that have different sources of errors as

$$
\begin{aligned}
n^{1/2}\left(\hat{\mu}_1 - \mu_1\right) \quad \cong \quad & \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[\mu_1\{X_i, \hat{e}(X_i)\} - \mu_1\{X_i, e^*(X_i)\}\right] && (4.8) \\
& + \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[\hat{\mu}_1\{X_i, e^*(X_i)\} - \mu_1\{X_i, e^*(X_i)\}\right] && (4.9) \\
& + \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[\mu_1\{X_i, e(X_i)\} - \mu_1\right], && (4.10)
\end{aligned}
$$

where $\cong$ follows by typical empirical process theory. The decomposition reveals three sources of errors: (4.8) accounts for the estimation of the propensity score in Step 1, (4.9) accounts for the estimation of the partial linear regression in Step 2, and (4.10) accounts for the variation of treatment effects in Step 3.

Analysis of (4.10) is standard. The major difficulty lies in the analysis of (4.8) and (4.9). For the variation from the propensity score, $\hat{e}(X_i)$ affects not only the argument of $\mu_1\{X_i, \hat{e}(X_i)\}$ but also implicitly its functional form, namely $\hat{\beta}_1$ and $\hat{g}_1(\cdot)$. To see this, we can further

decompose (4.8) by Taylor expansion:

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[\mu_1\{X_i,\hat{e}(X_i)\}-\mu_1\{X_i,e^*(X_i)\}\right]\cong\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}\,|_{\alpha=\alpha^*}\right]n^{1/2}(\hat{\alpha}-\alpha^*).$$

The influence function of $\hat{\alpha}$ can be easily derived from the standard M-estimation theory:

$$\hat{\alpha}-\alpha^*=\frac{1}{n}\sum_{i=1}^{n}M_e^{-1}\psi_e(X_iA_i;\alpha^*)+o_p\left(\frac{1}{n^{1/2}}\right),$$

where $M_e=\mathbb{E}\left\{-\partial\psi_e(X,A;\alpha)/\partial\alpha\,|_{\alpha=\alpha^*}\right\}$. Thus, we only need to derive the derivative of $\mathbb{E}\left[\partial\mu_1\{X,e(X;\alpha)\}/\partial\alpha\right]$ and evaluate at α^*. By (4.7), it can be written as

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}\,|_{\alpha=\alpha^*}\right] &= \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=1\}\right]|_{\alpha=\alpha^*}\\
&\quad+\mathbb{E}\left[\frac{\partial\, r_{X,1}\{e(X;\alpha)\}^{\mathrm{T}}}{\partial\alpha}\beta_1(\alpha^*)\right]|_{\alpha=\alpha^*}\\
&\quad+\mathbb{E}\left[r_{X,1}\{e^*(X)\}^{\mathrm{T}}\frac{\partial\beta_1(\alpha)}{\partial\alpha}\right]|_{\alpha=\alpha^*}\\
&= \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=1\}\,|_{\alpha=\alpha^*}\right] \qquad (4.11)\\
&\quad-\mathbb{E}\left[\frac{\partial\mathbb{E}\{X\mid e(X;\alpha),A=1\}^{\mathrm{T}}}{\partial\alpha}\right]|_{\alpha=\alpha^*}\beta_1(\alpha^*) \qquad (4.12)\\
&\quad+\mathbb{E}\left[r_{X,1}\{e^*(X)\}^{\mathrm{T}}\right]\frac{\partial\beta_1(\alpha)}{\partial\alpha}\,|_{\alpha=\alpha^*}, \qquad (4.13)
\end{aligned}$$

where $\beta_a(\alpha)$ is the plim of $\hat{\beta}_a(\alpha)$ as $n\to\infty$, which is a function of α, not X or Y. Then, we have

$$\begin{aligned}
\beta_a(\alpha) &= \mathbb{E}\left[r_{X,a}\{e(X;\alpha)\}\,r_{X,a}\{e(X;\alpha)\}^{\mathrm{T}}\mid A=a\right]^{-1}\mathbb{E}\left[r_{X,a}\{e(X;\alpha)\}\,r_{Y,a}\{e(X;\alpha)\}\mid A=a\right]\\
\beta_a(\alpha^*) &= \mathbb{E}\left[r_{X,a}\{e^*(X)\}\,r_{X,a}\{e^*(X)\}\mid A=a\right]^{-1}\mathbb{E}\left[r_{X,a}\{e^*(X)\}\,r_{X,a}\{e^*(X)\}\mid A=a\right].
\end{aligned}$$

For (4.11) and (4.12), the key is to characterize the dependence of $\mathbb{E}\{V\mid\hat{e}(X),A=1\}$ on $\hat{\alpha}$, which is a generated regressor problem that can be difficult to analyze(Mammen et al. 2016). Following Hahn and Ridder (2013), it can be decomposed into two parts by rewriting $\mathbb{E}\{V\mid e(X;\alpha),A=1\}$ as $\mathbb{E}\{V\mid e(X;\alpha_2)=e(X;\alpha_1),A=1\}$, where $\alpha_1=\alpha_2=\alpha$. But the roles of α_1 and α_2 are different: the change of α_1 reflects the value change of the estimated propensity score, and α_2 accounts for the variation of the projected space. Then we can

apply the chain rule, which leads to

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{V\mid e(X;\alpha),A=1\}\mid_{\alpha=\alpha^*}\right] &= \mathbb{E}\left[\frac{\partial}{\partial\alpha_1}\mathbb{E}\{V\mid e^*(X)=e(X;\alpha_1),A=1\}\mid_{\alpha_1=\alpha^*}\right]\\ &\quad+\mathbb{E}\left[\frac{\partial}{\partial\alpha_2}\mathbb{E}\{V\mid e(X;\alpha_2)=e^*(X),A=1\}\mid_{\alpha_2=\alpha^*}\right] \quad (4.14)\end{aligned}$$

The first part can be easily handled by applying the chain rule again. The analysis for the second part is more difficult. The key is the following orthogonality condition, which plays a crucial role throughout the paper:

$$\mathbb{E}(h\{e(X;\alpha)\}A[V-\mathbb{E}\{V\mid e(X;\alpha),A=1\}]) = 0. \quad (4.15)$$

By selecting $h=\pi\{e(X;\alpha)\}^{-1}$, where $\pi\{e(X;\alpha)\}=P\{A=1\mid e(X;\alpha)\}$, and taking partial derivatives, we obtain the following proposition. To simplify notations, we further denote $\lambda_{V,a}\{e^*(X)\}=\mathbb{E}\{V\mid e^*(X),A=a\}$ and $\hat{\lambda}_{V,a}\{e^*(X)\}=\mathbb{E}\{V\mid e^*(X),A=a\}$.

Proposition 1 *Under Assumptions 3 and 4, we have*

$$\begin{aligned}&\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{V\mid e(X;\alpha),A=1\}\mid_{\alpha=\alpha^*}\right]\\ =\;&\mathbb{E}\left[\frac{\pi\{e(X;\alpha^*)\}-e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\mathbb{E}\{V\mid v,A=1\}}{\partial v}\mid_{v=v^*}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\mid_{\alpha=\alpha^*}\right]\\ &+\mathbb{E}\left[\frac{e(X)\lambda_{V,1}\{e^*(X)\}}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right]\\ &-\mathbb{E}\left[\frac{e(X)\mathbb{E}(V\mid X,A=1)}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right]\end{aligned}$$

for V being either X or Y, where $v=e(X;\alpha)$, and $v^*=e^*(X)$.

The detailed proof of this and following propositions are included in the supplementary materials. The tricky part in the proof is that $\partial\mathbb{E}\{V\mid e(X;\alpha_2)=e^*(X),A=1\}/\partial\alpha_2\mid_{\alpha_2=\alpha^*}$ is a function of $e^*(X)$, but this is not true for $\partial\mathbb{E}\{V\mid e^*(X)=e(X;\alpha_1),A=1\}/\partial\alpha_1\mid_{\alpha_1=\alpha^*}$ or $\partial\mathbb{E}\{V\mid e(X;\alpha),A=1\}/\partial\alpha\mid_{\alpha=\alpha^*}$. Thus, $A/\pi\{e(X;\alpha)\}$ can be cancelled out when calculating $\mathbb{E}\left[\partial\mathbb{E}\{V\mid e(X;\alpha_2)=e^*(X),A=1\}/\partial\alpha_2\mid_{\alpha_2=\alpha^*}\right]$ by taking conditional expectations. On the other hand, $\mathbb{E}\left[\partial\mathbb{E}\{V\mid e^*(X)=e(X;\alpha_1),A=1\}/\partial\alpha_1\mid_{\alpha_1=\alpha^*}\right]$ can be directly derived by the chain rule. By (4.14), we are able to obtain the derivative of $\mathbb{E}\{V\mid e(X;\alpha),A=1\}$.

For (4.9), we would use the pathwise derivative to derive the adjustment term from the non-parametric estimates, following Newey (1994). Denote $\{F_\theta\}$ as a one-dimensional

subfamily of the distribution family of $\{X, A, Y\}$, i.e. a path in the general family, and it is equal to the true distribution when $\theta = 0$. Suppose we would like to estimate the variance formula for an estimator $\hat{\mu}$. It is equivalent to finding the semiparametric bound for the estimation of $\mu(F_\theta)$, which is the probability limit of $\hat{\mu}$ if the data follows the distribution F_θ, through its pathwise derivative. Theorem 2.1 in Newey (1994) illustrates this, and we restate this as a lemma in the supplementary materials. With the help of this lemma, we focus on deriving the pathwise derivative for our DR regression estimator. Specifically, we would like to decompose (4.9) for further analysis. To simplify notations, we use $\hat{\beta}_1^*$ to denote $\hat{\beta}_1(\alpha^*)$ and β_1^* to denote its limit in probability. Then, we have

$$
\begin{aligned}
\hat{\mu}_1\{X_i, e^*(X_i)\} - \mu_1\{X_i, e^*(X_i)\} &= \hat{\lambda}_{Y,1}\{e^*(X_i)\} + \left[X_i - \hat{\lambda}_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \hat{\beta}_1^* \\
&\quad - \lambda_{Y,1}\{e^*(X_i)\} - \left[X_i - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \beta_1^* \\
&= \hat{\lambda}_{Y,1}\{e^*(X_i)\} - \lambda_{Y,1}\{e^*(X_i)\} \\
&\quad + \left[X_i - \hat{\lambda}_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \hat{\beta}_1^* - \left[X_i - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \hat{\beta}_1^* \\
&\quad + \left[X_i - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \hat{\beta}_1^* - \left[X_i - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \beta_1^* \\
&= \hat{\lambda}_{Y,1}\{e^*(X_i)\} - \lambda_{Y,1}\{e^*(X_i)\} \\
&\quad - \left[\hat{\lambda}_{X,1}\{e^*(X_i)\} - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \hat{\beta}_1^* \\
&\quad + \left[X_i - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \left(\hat{\beta}_1^* - \beta_1^*\right) \\
&\cong \hat{\lambda}_{Y,1}\{e^*(X_i)\} - \lambda_{Y,1}\{e^*(X_i)\} \qquad (4.16) \\
&\quad - \left[\hat{\lambda}_{X,1}\{e^*(X_i)\} - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \beta_1^* \qquad (4.17) \\
&\quad + \left[X_i - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \left(\hat{\beta}_1^* - \beta_1^*\right), \qquad (4.18)
\end{aligned}
$$

where $\cong$ is because $\left[\hat{\lambda}_{X,1}\{e^*(X_i)\} - \lambda_{X,1}\{e^*(X_i)\}\right]^{\mathrm{T}} \left(\hat{\beta}_1^* - \beta_1^*\right)$ as a higher order term can be negligible. (4.16) stands for the adjustment term from the nonparametric estimator $\hat{\lambda}_{Y,1}\{e^*(X)\}$, and (4.17) is the adjustment term from $\hat{\lambda}_{X,1}\{e^*(X)\}$ multiplied by β_1^*. (4.18) accounts for the variability from estimating $\hat{\beta}_1^*$, which comes from a M-estimating equation with nonpametric terms involved. The orthogonality condition (4.15) will also be frequently used because the key components of our estimator are conditional expectations, resulting the following proposition. To apply the pathwise derivative techniques, we need to further assume some regularity conditions for the nonparametric estimators listed in the Supplementary Materials.

Proposition 2 *Under Assumptions 3, 4, and the regularity conditions in the Supplementary*

Matrials, the adjustment term for $\hat{\lambda}_{V,1}\{e^*(X)\}$ *is*

$$\alpha_{V,1}(X,A,Y)=\frac{A\big[V-\lambda_{V,1}\{e^*(X)\}\big]}{\pi\{e^*(X)\}},$$

for V *being either* X *or* Y*, i.e.,*

$$\frac{1}{n}\sum_{i=1}^{n}\big[\hat{\lambda}_{V,1}\{e^*(X_i)\}-\lambda_{V,1}\{e^*(X_i)\}\big]=\frac{1}{n}\sum_{i=1}^{n}\alpha_{V,1}(X_i,A_i,Y_i)+o_p(1).$$

The derivation here is similar to Hahn and Ridder (2013)'s works on the asymptotic result for the HIT estimator. However, the difference is that the conditional expectations take effects not only from the explicit terms $\hat{\lambda}_{Y,1}\{e^*(X)\}$ and $\hat{\lambda}_{X,1}\{e^*(X)\}$ in (4.16) and (4.17), but also indirectly through the ordinal least square estimator $\hat{\beta}_1$ in (4.18), which is solved from an M-estimating equation constructed by the conditional expectations. This problem also appears in the analysis for (4.8) as shown in (4.13), leading to a big challenge in our estimator. It is difficult becuase, from the nature of ordinal least square estimator, we have to deal with the derivatives of conditional expectations of the products of residuals: $\mathbb{E}\big[r_{X,1}\{e(X;\alpha)\}\,r_{Y,1}\{e(X;\alpha)\}\mid A=1\big]$ and $\mathbb{E}\big[r_{X,1}\{e(X;\alpha)\}\,r_{X,1}^{\mathrm{T}}\{e(X;\alpha)\}\mid A=1\big]$, with respect to both the propensity score parameter α and the pathwise parameter θ, which is more challenging than deriving derivatives for $\lambda_{V,1}\{e^*(X)\}$. For (4.8), we apply the orgonality conditions as well as the properties of residuals to derive the variation from the propensity score parameter, and the result is summarized in the following proposition.

Proposition 3 *Under Assumptions 3 and 4, we have*

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}\mathbb{E}\big[r_{X,1}\{e(X;\alpha)\}\,r_{Y,1}\{e(X;\alpha)\}\mid A=1\big]\big|_{\alpha=\alpha^*}\\
=\;&-\mathbb{E}\left(\frac{\partial\mathbb{E}\{X\mid v,A=1\}}{\partial v}\big|_{v=v^*}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\big|_{\alpha=\alpha^*}\cdot r_{Y,1}\{e^*(X)\}\mid A=1\right)\\
&-\mathbb{E}\left(r_{X,1}\{e^*(X)\}\cdot\frac{\partial\mathbb{E}\{Y\mid v,A=1\}}{\partial v}\big|_{v=v^*}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\big|_{\alpha=\alpha^*}\mid A=1\right),
\end{aligned}$$

$$
\begin{aligned}
& \frac{\partial}{\partial \alpha_j}\mathbb{E}\left[r_{X,1}\{e(X;\alpha)\}\, r_{X,1}^{\mathrm{T}}\{e(X;\alpha)\} \mid A=1\right]|_{\alpha=\alpha^*} \\
= \; & -\mathbb{E}\left(\frac{\partial \mathbb{E}\{X \mid v, A=1\}}{\partial v}|_{v=v^*} \cdot \frac{\partial e(X;\alpha)}{\partial \alpha_j}|_{\alpha=\alpha^*} \cdot r_{X,1}^{\mathrm{T}}\{e^*(X)\} \mid A=1 \right) \\
& -\mathbb{E}\left(\frac{\partial \mathbb{E}\{X \mid v, A=1\}}{\partial v}|_{v=v^*} \cdot \frac{\partial e(X;\alpha)}{\partial \alpha_j}|_{\alpha=\alpha^*} \cdot r_{X,1}^{\mathrm{T}}\{e^*(X)\} \mid A=1 \right)^{\mathrm{T}},
\end{aligned}
$$

where $v = e(X;\alpha)$*, and* $v^* = e^*(X)$*.*

Based on this proposition and matrix calculus, we are able to derive the derivative for $\beta_1(\alpha)$.

Proposition 4 *Under Assumptions 3 and 4, we have*

$$
\begin{aligned}
\frac{\partial \beta_1(\alpha)}{\partial \alpha}|_{\alpha=\alpha^*} = \; & \mathbb{E}\left(r_{X,1}\{e^*(X)\}\, r_{X,1}^{\mathrm{T}}\{e^*(X)\} \mid A=1\right)^{-1} \\
& \cdot \Big(\frac{\partial}{\partial \alpha}\mathbb{E}\left[r_{X,1}\{e(X;\alpha)\}\, r_{Y,1}\{e(X;\alpha)\} \mid A=1\right]|_{\alpha=\alpha^*} \\
& -\sum_{j=1}^{p} \frac{\partial}{\partial \alpha_j}\mathbb{E}\left[r_{X,1}\{e(X;\alpha)\}\, r_{X,1}^{\mathrm{T}}\{e(X;\alpha)\} \mid A=1\right]|_{\alpha=\alpha^*} \cdot B_j \Big),
\end{aligned}
$$

where the partial derivatives are provided in Proposition (3) *and* $B_j = (\underbrace{\mathbf{0},...,\mathbf{0}}_{j-1}, \beta_1(\alpha^*), \mathbf{0}, ..., \mathbf{0})$ *is a* $p \times p$ *matrix with the* j*-th column as* $\beta_1(\alpha^*)$ *and other columns as zero vectors.*

For (4.9), we use Newey (1994)'s pathwise derivative technique for the M-estimator, summarized in the lemma in the supplementary materials. However, we cannot directly derive the adjustment terms for the product of residuals due to its complicated structure. Fortunately, we are able to obtain the influence function of $\hat{\beta}_1^*$ as a whole term by heavily applying the orthogonality conditions.

Proposition 5 *Under Assumptions 3, 4, and the regularity conditions in the Supplementary Matrials, the influence function of* $\hat{\beta}_1^*$ *is*

$$
\begin{aligned}
\psi_{\beta,1}(X,A,Y) = \; & \mathbb{E}\left[A r_{X,1}\{e^*(X)\}\, r_{X,1}^{\mathrm{T}}\{e^*(X)\}\right]^{-1} \\
& \cdot \left(A\left[X - \lambda_{X,1}\{e^*(X)\}\right]\left[Y - X^{\mathrm{T}}\beta_1^* - \frac{A\varphi_1\{e^*(X)\}}{\pi\{e^*(X)\}}\right]\right),
\end{aligned}
$$

and the influence function of $\hat{\beta}_0^*$ *is*

$$\begin{aligned}\psi_{\beta,0}(X,A,Y) &= \mathbb{E}\big[(1-A)\, r_{X,0}\{e^*(X)\}\, r_{X,0}^{\mathrm{T}}\{e^*(X)\}\big]^{-1} \\ &\quad \cdot\left((1-A)\big[X-\lambda_{X,0}\{e^*(X)\}\big]\left[Y-X^{\mathrm{T}}\beta_0^*-\frac{(1-A)\,\varphi_0\{e^*(X)\}}{\pi\{e^*(X)\}}\right]\right),\end{aligned}$$

where $\varphi_a\{e^*(X)\}=\mathbb{E}\big[e(X)\big\{\mathbb{E}(Y\mid X,A=a)-X^{\mathrm{T}}\beta_a^*\big\}\mid e^*(X)\big]$.

Having all these propositions, we are ready to derive the asymptotic results for our DR regression estimator in the following section.

4.4 Main results

In this section, we establish the asymptotic results for the DR regression estimator. We first show that the estimator is consistent with the double robust property, which means that the estimator is consistent with the ATE if either the propensity score or outcome model is correctly specified. By saying that the propensity score model is correctly specified, $e(X)=e(X;\alpha_0)$, where α_0 is the true propensity parameter. On the other hand, the outcome model is correctly specified if $\mathbb{E}\{Y\mid X,A=a\}=X^{\mathrm{T}}\beta_a$ for $a=0,1$. In addition to the assumptions stated in Section 4.2, we further impose some conditions for the nonparametric estimators. To obtain consistency for our DR regression estimator, we require all the nonparametric estimators included to be consistent:

Assumption 5 *The nonparametric estimators for the conditional expectations are consistent, i.e.,* $\mathbb{E}\{Y\mid e(X;\alpha),A=a\}\xrightarrow{p}\mathbb{E}\{Y\mid e(X;\alpha),A=a\}$ *and* $\mathbb{E}\{X\mid e(X;\alpha),A=a\}\xrightarrow{p}\mathbb{E}\{X\mid e(X;\alpha),A=a\}$ *for any* α*, where* $a=0,1$.

In the following theorem, we prove that the DR regression estimator is consistent with the ATE with the double robustness property.

Theorem 5 (Consistency) *Under Assumptions 3-5, if either the propensity score or outcome model is correctly specified, the DR regression estimator is a consistent estimator of the ATE.*

Remark 5 *From Theorem 5, the consistency of the DR regression estimator is guaranteed if any model for the propensity score or outcome is correctly specified.*

The key result of this paper is presented in the next following theorem.

Theorem 6 (Influence function) *Under Assumptions 3-5, and the regularity conditions in the Supplementary Matrials, the influence function of the DR regression estimator is*

$$\left(\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}|_{\alpha=\alpha^*}\right]-\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_0\{X,e(X;\alpha)\}|_{\alpha=\alpha^*}\right]\right)M_e^{-1}\psi_e(A,X;\alpha^*) \tag{4.19}$$

$$+\quad \frac{A\left[Y-\lambda_{Y,1}\{e^*(X)\}\right]}{\pi\{e^*(X)\}}+\frac{\pi\{e^*(X)\}-A}{\pi\{e^*(X)\}}\left[X-\lambda_{X,1}\{e^*(X)\}\right]^{\mathrm{T}}\beta_1^* \tag{4.20}$$

$$-\quad \frac{(1-A)\left[Y-\lambda_{Y,0}\{e^*(X)\}\right]}{1-\pi\{e^*(X)\}}+\frac{\pi\{e^*(X)\}-A}{1-\pi\{e^*(X)\}}\left[X-\lambda_{X,0}\{e^*(X)\}\right]^{\mathrm{T}}\beta_0^* \tag{4.21}$$

$$+\quad \mathbb{E}\left(r_{X,1}\{e^*(X)\}\right)^{\mathrm{T}}\psi_{\beta,1}(X,A,Y)-\mathbb{E}\left(r_{X,0}\{e^*(X)\}\right)^{\mathrm{T}}\psi_{\beta,0}(X,A,Y), \tag{4.22}$$

$$+\quad \lambda_{Y,1}\{e^*(X)\}-\lambda_{Y,0}\{e^*(X)\}-\tau \tag{4.23}$$

where

$$\begin{aligned}
\psi_e(A,X;\alpha) &= \frac{\partial e(X;\alpha)}{\partial\alpha}\cdot\frac{A-e(X;\alpha)}{e(X;\alpha)\{1-e(X;\alpha)\}},\\
M_e &= \mathbb{E}\left\{-\frac{\partial}{\partial\alpha}\psi_e(A,X;\alpha)|_{\alpha=\alpha^*}\right\},\\
\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_a\{X,e(X;\alpha)\}|_{\alpha=\alpha^*}\right] &= \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=a\}|_{\alpha=\alpha^*}\right] \qquad (4.24)\\
&\quad -\mathbb{E}\left[\frac{\partial\mathbb{E}\{X\mid e(X;\alpha),A=a\}^{\mathrm{T}}}{\partial\alpha}\right]|_{\alpha=\alpha^*}\beta_a^* \qquad (4.25)\\
&\quad +\mathbb{E}\left[r_{X,a}\{e^*(X)\}^{\mathrm{T}}\right]\frac{\partial\beta_a(\alpha)}{\partial\alpha}|_{\alpha=\alpha^*}, \qquad (4.26)
\end{aligned}$$

and the partial derivatives are included in Propositions (1)-(5).

Remark 6 (Explanation of the terms in the influence function) (4.19) *accounts for the effect of the estimated propensity score. It can be further decomposed into three parts, where* (4.24) *and* (4.25) *stand for the direct effect on the conditional expectations from the propensity score parameter, and* (4.26) *represents the indirect effect through the least square estimator* $\hat{\beta}_a$. (4.20), (4.21), *and* (4.22) *correspond to* (4.9) *in Section* (4.3.2). *Similarly,* (4.20) *and* (4.21) *stand for the direct effect from the nonparametrically estimated conditional expectations, while* (4.22) *represent the indirect effect through the least square estimates.* (4.23) *correspond to* (4.10).

Remark 7 (Comparison to HIT) *The influence function of the HIT estimator derived by*

Hahn and Ridder (2013) is a part of the influence function of the DR regression estimator because of their common structure: the HIT estimator fits a nonparametric model of the propensity score for the outcome, and the DR regression estimator adds an additional linear component. Thus, there are multiple terms related to β_a in the influence function of the DR regression estimator. Besides, because we allow the propensity score model to be misspecified, $\pi\{e^*(X)\}$ are involved in the influence function to account for this difference.

Remark 8 (Generality) *Theorem 5 always holds even if both the propensity score and outcome models are misspecified. $e^*(X)$ is the probability limit of the estimated propensity score from the model, not necessarily the true propensity score. The conditional expectations and term $\pi\{e(X;\alpha^*)\}$ depend on the misspecified propensity score and make no assumption on the generating model. The non-reliance on the correctness of the nuisance models indicates better generality of the theorem.*

Remark 9 (Variance estimation) *Theorem 5 holds generally; however, it is impossible to construct a plug-in variance estimator based on the asymptotic result due to not only the great complexity of the variance formula but also some unattainable terms such as $\partial\mathbb{E}\{Y \mid v, A=a\}/\partial v \mid_{v=v^*}$. They cannot be further reduced unless making strong distributional assumptions. Even though some numerical methods may be applied to estimate them, the numerical error may be non-negligible in practice. Instead, we recommend using a variation of the bootstrap algorithm to estimate the variance. Following Newey (1997), we require the tuning parameters of the nonparametric estimators in each bootstrap replicate remain the same: we directly use the tuning parameters selected from cross-validation based on the original full data. Simulations in Section 4.5 support the validity of the variance estimator.*

Remark 10 (Efficiency) *Although the influence function is complicated, it can be reduced to the efficient influence function when both the propensity score and outcome models are correctly specified. The following corollary illustrates the semi-parametric efficiency of the estimator.*

Corollary 1 *Under Assumptions 3-5, and the regularity conditions in the Supplementary Materials, if both the propensity score and outcome models are correctly specified, the DR regression estimator is semi-parametrically efficient.*

4.5 Simulation

We design a simulation study to investigate the finite sample properties of the proposed DR regression estimator, including the double robustness property and its efficiency compared to other existing estimators. Quadratic polynomials are used for the propensity score and outcome models to allow possible misspecifications. We also consider two different settings regarding the propensity score distribution. According to Kang and Schafer (2007), weighting estimators maybe perform unstably when propensity score distribution is extreme, i.e., the probabilities of receiving treatment are close to zero or one. Zhou et al. (2019) showed that the PENCOMP estimator outperformed AIPW under misspecified outcome models and highly variable weights. On the other hand, ensemble machine learning algorithms, such as Targeted Maximum Likelihood Estimator (TMLE) and AIPW with nuisance functions estimated by Super Learner (AIPWsl), are widely used in practice because they are robust to model misspecification. To the best of our knowledge, there has not been a comprehensive study comparing these methods with doubly robust regression and weighting estimators, motivating us to conduct a simulation study with complex data generative models and multiple candidate estimators.

Specifically, we generate confounders $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i = 1,...,n$ and $j = 1,...,p$, where sample size $n = 1000$ and number of dimensions $p = 4$. Here X_{i1} stands for the confounder; X_{i2} represents the instrumental variable; X_{i3} is the predictor for the outcome but not for the treatment assignment; X_{i4} is the noise variable. The treatment indicator A follows a second-order logistic model, where the propensity score $e(X_i) = \text{logit}^{-1}\{\gamma(1 + X_{i1} - 2X_{i2} - X_{i1}^2 + 0.5 X_{i2}^2 + 0.5X_{i1}X_{i2})\}$. γ controls the distribution of the propensity score: we use $\gamma = 1$ in the extreme propensity score setting and $\gamma = 0.25$ in the non-extreme setting. Visualization of the propensity score distributions in the two settings is provided in Figure 4.1. The potential outcomes are generated from a second-order regression model: $Y_i(0) = -1 - 2X_{i1} + X_{i3} - X_{i1}^2 + 2X_{i3}^2 - 3X_{i1}X_{i3} + \epsilon_i$ and $Y_i(1) = Y_i(0) + 2X_{i1} + X_{i3}$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Thus, the ATE is 0 under this data generating model.

To illustrate the doubly robust property of the proposed DR regression estimator, we consider two different models to fit the propensity score: A second-order logistic model including interactions is used to obtain the correct propensity score, while a misspecified first-order logistic model generates the biased propensity score. Similarly, we use two different models to fit the outcomes: a correctly specified second-order regression or partial regression model is used to fit both the treated and controlled outcomes, while a standard linear regression or partial regression model is included as a misspecified model.

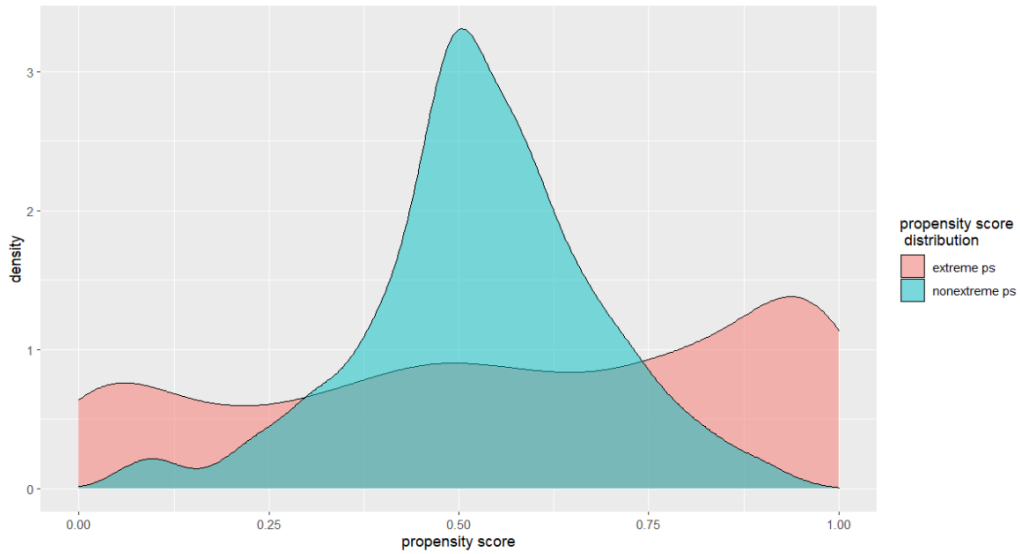


Figure 4.1: Density plot of the propensity score distributions in the extreme and non-extreme settings.

We compare the following estimators:

a) Regression estimators, including DR regression estimators (DReg), standard regression estimators (Reg), and HIT estimators;

b) AIPW estimators, including the parametric and nonparametric (Super Learner) versions;

c) TMLE estimator.

Each regression and estimator is combined with a one-digit or two-digit number, indicating how the nuisance functions are estimated. We use "1" to represent the model correctly specified and "0" to denote that the model is misspecified. For example, "Reg1" is the regression estimator based on second-order polynomials, and "HIT0" stands for the HIT estimator with misspecified propensity scores. For DR regression and AIPW estimators with a two-digit number, the first number indicates the propensity score model and the second number indicates the outcome model. For example, "DReg10" is the DR regression estimator with the propensity score correctly specified but the outcome model misspecified. A special case for AIPW is "AIPWsl", where propensity score and outcome are estimated by Super Learner. In our simulation, we include the generalized linear model (SL.glm), LASSO (SL.glmnet), and Bayesian regression tree (tmle.SL.dbarts2) in the library. The same ensemble methods are used in TMLE as well. For DR regression estimators, the nonparametric

Table 4.1: Monte Carlo bias and standard error of the estimators when models are correctly specified from the simulation results based on 1000 repeated sampling. The mean square error (MSE) ratio is calculated by using the MSE for Dreg11 divided by MSE for each estimator. Ratio smaller than 1 reflects less efficient estimator.

(a) Extreme propensity score scenario

Estimator	Dreg11	AIPW11	AIPWsl	TMLE	Reg1	HIT1
Bias	0.001	-0.001	0.038	0.046	0.001	0.246
Std. error	0.137	0.182	0.138	0.141	0.122	0.563
MSE ratio	1.000	0.560	0.914	0.845	1.254	0.049

(b) Non-extreme propensity score scenario

Estimator	Dreg11	AIPW11	AIPWsl	TMLE	Reg1	HIT1
Bias	0.001	0.001	0.013	0.020	0.001	0.021
Std. error	0.098	0.099	0.108	0.115	0.098	0.162
MSE ratio	1.000	0.988	0.810	0.706	1.008	0.360

components are estimated via smoothing spline methods, and the variance is estimated by bootstrap with fixed tuning parameters in the nonparametric estimation as mentioned in Remark 9.

Figure 4.2 shows the distributions of the various estimators based on 1000 repeated sampling. As illustrated from the results, the DR regression estimator is doubly robust, i.e., consistent unless both the propensity score and outcome models are misspecified. The DR regression estimator is also efficient when both models are correctly specified because "DReg11" has a similar variance as the asymptotically efficient estimator "AIPW11" in the non-extreme propensity score scenario. When the distribution of the propensity score is extreme, the DR regression estimator is more stable than the AIPW estimator, i.e., lighter tail in the distribution. Note that "DRreg10" is not invulnerable to extreme propensity scores even though the propensity score model is correctly specified. The reason is that the nonparametric component may be estimated poorly when some propensity scores are closed to zero or one. This circumstance can be reversed by the correctly specified outcome model, for example, "Dreg11" and "Dreg01". Nevertheless, the performance of "DReg10" is still better than "AIPW10".

Table 4.1 includes the summary statistics of the estimators when models are correctly specified. We also calculate the mean square error (MSE) ratio by using the MSE for Dreg11 divided by MSE for each estimator. Ratio smaller than 1 reflects less efficient estimator. The standard regression estimator achieves the minimum error across different settings, but it can be significantly biased if the outcome model is misspecified. Although the HIT

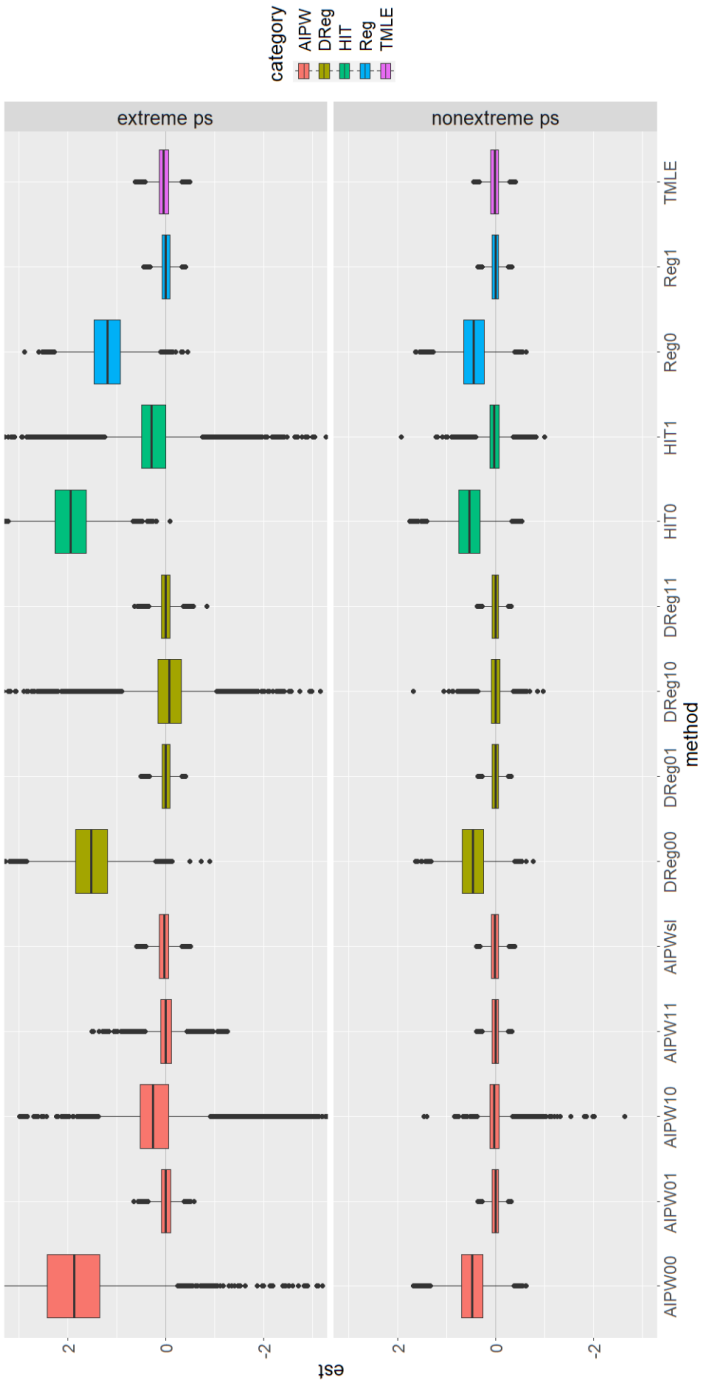


Figure 4.2: Boxplot of the various estimators from the simulation results based on 1000 repeated sampling. The upper panel is for the extreme propensity score scenario, and the lower panel is for the non-extreme propensity score scenario. Some extreme outliers are not shown in the plot.

Table 4.2: Coverage rates of the DReg, AIPWsl, and TMLE estimators (empirical coverage rate ± 1.96×Monte Carlo standard error) from the simulation results based on 1000 repeated sampling. The variances of the DR regression estimators are estimated based on 100 bootstrap replicates.

Estimator	Dreg11	Dreg10	Dreg01	Dreg00	AIPWsl	TMLE
Extreme propensity score	(93.3,94.2)	(91.8,92.8)	(94.1,95.0)	(10.3,11.5)	(78.3,79.9)	(86.4,87.7)
Non-extreme propensity score	(94.3,96.9)	(97.1,98.9)	(94.6,97.0)	(79.7,84.5)	(87.9,89.2)	(93.6,94.5)

estimator has been proved to be asymptotically efficient(Hahn and Ridder 2013), its finite sample performance is worse than DR regression and AIPW estimator, and it is even more vulnerable to extreme propensity scores as shown in Table 4.1(a). Ensemble methods such as "AIPWsl" and "TMLE" are consistent, efficient, and stable. However, due to the numerical errors from the various assembled algorithms, their bias and variance may be larger than those of the DR regression estimator to some extent.

Table 4.2 summarizes the coverage rates of the various DR regression estimators using the bootstrap method in Remark 9. The coverage rates are closed to the nominal level if at least one of the propensity score or outcome model is correctly specified. Note that the coverage rate of "Dreg10" is slightly away from 95%, implying the greater importance of the outcome information. The coverage rates of the ensemble methods are also included in the table. The variance estimators are based on their influence functions and are provided in the corresponding package. The "AIPWsl" estimator always tends to underestimate the variance. The TMLE estimator provides reasonable estimate in the non-extreme propensity score scenario, but the corresponding confidence interval also suffers from undercoverage.

CHAPTER

5

SENSITIVITY ANALYSIS IN CAUSAL INFERENCE WITH INFORMATIVE SAMPLING

5.1 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a large group of nearly 5,000 different synthetic chemicals that are resistant to heat, water, and oil (Grandjean and Clapp 2015). A few of the most studied PFAS, such as perfluorooctanoic acid (PFOA) and perfluorooctane sulfonic acid (PFOS), have been found to be hazardous to human health (Al Amin et al. 2020; Sunderland et al. 2019; Blake and Fenton 2020). However, it is important to consider the informative nature of the sampling process when selecting locations for measuring PFAS. In particular, locations with higher PFAS concentrations are more likely to be sampled (Zhang et al. 2020). While certain covariates such as the distance to military bases can be controlled, it is essential to acknowledge the presence of unmeasured confounders that can influence both the sampling process and PFAS levels. Failure to account for this sensitivity

analysis could result in sampling bias. Moreover, unmeasured confounders between PFAS exposure and the desired health outcome may exist, necessitating the use of multiple sensitivity analysis models to accurately estimate the causal effect. Furthermore, the unique structure of PFAS datasets offers an additional characteristic distinct from other sampling problems. Although PFAS measurements may not always be available, covariates such as geographical and population data are typically collected. Additionally, health outcome data for these locations may be accessible. The distinctive attributes of this PFAS dataset inspire the development of innovative methods to investigate the causal effect of PFAS on the desired health outcome.

Causal inference is a field that focuses on methodologies to estimate causal effects. In observational studies, there is often an inherent imbalance between the treatment and control groups. To address this imbalance, researchers have developed various estimators to quantify causal effects. However, most of these estimators rely on the assumption of no unmeasured confounders. When unmeasured confounders are present, the estimators may become inconsistent and biased. To address this issue, a sensitivity analysis framework has been developed to relax this assumption. One approach is to adopt a parametric model for the sensitivity framework, such as the logistic selection specification. However, selecting an appropriate sensitivity parameter can be challenging and lacks clear guidelines. Additionally, as far as we are aware, there is currently no existing model that simultaneously handles both unmeasured confounders and informative sampling.

In this paper, we present the derivation of the semiparametric efficient estimator for the average treatment effect (ATE) within the context of a sensitivity analysis framework with informative sampling, building upon the semiparametric theory (Tsiatis 2006). Drawing inspiration from the studies by Franks et al. (2020); Scharfstein et al. (2021), we employ the logistic selection specification to model the impact of unmeasured confounders. By utilizing this parametric model, we first establish the identification formula for the marginal mean and subsequently derive the efficient influence function (EIF) based on semiparametric theory. The semiparametric efficient estimator of the ATE is then constructed using the EIF, along with its asymptotic distribution and variance estimator. Recognizing that the EIF-based estimator necessitates knowledge of the true sensitivity parameters, we propose an approach to calibrate these parameters, building upon the work of Franks et al. (2020) with a modification to address an error. To assess the performance of the EIF-based estimator, we conduct a simulation study comparing it to other existing estimators. Finally, we apply our estimator to the PFAS dataset to investigate the health effects of PFAS.

The subsequent sections of this article are organized as follows: In Section (5.2), we

introduce the notations and assumptions required to construct the sensitivity analysis framework with informative sampling, particularly emphasizing the logistic selection specification. Section (5.3) presents the identification formula and efficient influence function for the ATE, along with the semiparametric efficient estimator and its asymptotic properties. The calibration of sensitivity parameters is discussed in Section (5.4). In Section (5.5), we present a simulation study to demonstrate the numerical properties of the EIF-based estimator. The application of the semiparametric efficient estimator to the PFAS dataset is described in Section (5.6), and finally, we conclude in Section (5.7).

5.2 Notation and assumptions

Let X_i be a p-dimensional vector of pre-treatment covariates, A_i the binary treatment indicator, S_i the binary sampling indicator, and Y_i the outcome for unit $i = 1, \ldots, n$. We follow the potential outcomes framework. Let $Y_i(a)$ be the potential outcome had unit i been given treatment a ($a = 0, 1$). Based on the potential outcomes, the ATE is $\tau = \mathbb{E}\{Y_i(1) - Y_i(0)\}$ and the ATT is $\tau_{\text{ATT}} = \mathbb{E}\{Y_i(1) - Y_i(0) \mid A_i = 1\}$. The observed outcome is $Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i) Y_i(0)$. We assume that $\{X_i, A_i, S_i, Y_i(0), Y_i(1)\}$, $i = 1, \ldots, n$, are independent and identically distributed. Thus, (X_i, A_i, S_i, Y_i), $i = 1, \ldots, n$, are also independent and identically distributed.

Motivated by the PFAS dataset, we do not assume all the information are missing for $S_i = 0$. For locations where PFAS chemicals are not collected, the covariates such as the population and landcover data and even the health outcome data may be available. Thus, we only assume A_i's are missing for locations with $S_i = 0$. For non-sampled locations, we need the covariates information X_i to remove the selection bias for estimating the ATE.

For simplicity of exposition, denote

$$\mu_a\left\{g(Y); X\right\} = \mathbb{E}\left\{g(Y) \mid X, A = a\right\}, \omega_s(X) = \mathbb{P}(S = s \mid X), \pi_a(X) = \mathbb{P}(A = a \mid X, S = 1),$$

where $\mu_a\left\{g(Y); X\right\}$ is the outcome mean function, $\omega_s(X)$ is the sampling probability for $s = 0, 1$, and $\pi_a(X)$ is the treatment probability for sampled subjects for $a = 0, 1$.

To identify the causal effects, we make the standard positivity assumption following Rosenbaum and Rubin (1983b).

Assumption 6 *There exist constants c_1 such that $\omega_1(X) \geq c_1 > 0$ almost surely, and there exist constants d_1 and d_2 such that $0 < d_1 \leq \pi_1(X) \leq d_2 < 1$ almost surely.*

The first part of the assumption is on the sampling mechanism, and the second part is about the treatment probability. The sampling probability needs to be positive so that enough information is collected. Similarly, positive $\pi_1(X)$ is necessary to identify $\mathbb{E}\{Y(1)\}$, and positive $\pi_0(X)$ is required to identify $\mathbb{E}\{Y(0)\}$.

Besides, we make a further assumption on the conditional independence between sampling and potential outcomes given covariates and treatment.

Assumption 7 $S \perp\!\!\!\perp Y(a) \mid X, A$.

This assumption cannot be justified by the observed data. However, it should hold if we belive that the mechanisms of sampling and potential outcomes follow separate independent processes given the covariates and treatment.

The following statement is the key assumption for the sensitivity framework between treatment and potential outcomes, following Franks et al. (2020); Scharfstein et al. (2021).

Assumption 8 $\mathbb{P}(A = 1-a \mid X, Y(a)) = \text{logit}^{-1}\left[\beta_a(x) + \gamma_a s_a\{Y(a)\}\right]$ *for* $a = 0, 1$.

This assumption can also be rewritten using Bayes rule in the following form:

$$dF\{Y(a) \mid A = 1-a, X\} = dF\{Y(a) \mid A = a, X\} \cdot \frac{exp\left[\gamma_a s_a\{Y(a)\}\right]}{\mu_a\left\{\gamma_a s_a(Y); X\right\}},$$

for $a = 0, 1$, where dF is the probability density, γ_a is the sensitivity parameter between treatment and potential outcomes, and $s_a\{Y(a)\}$ is some sufficient statistics of the potential outcomes. The assumption refers to a logistic selection specification of the sensitivity framework, which implies that the unobserved quantity can be estimated via some "tilting" function. Similarly, the conditional expectation for the observed and unobserved outcomes can also be related via this "tilting function"

For the sensitivity framework between sampling and treatment, we adopt the analogical logistic selection specification in the following assumption.

Assumption 9 $\mathbb{P}(S = 1 \mid X, A = a) = \text{logit}^{-1}\left\{\alpha(x) + \gamma_s a\right\}$ *for* $a = 0, 1$.

Because A is binary, we do not need to assume a sufficient statistics as we did in the previous assumption. Similarly, this assumption can be rewritten in the following form:

$$\mathbb{P}(A = a \mid X, S = 0) = \mathbb{P}(A = a \mid X, S = 1) \cdot \frac{exp\left(-\gamma_s a\right)}{\eta\left(\gamma_s, X\right)},$$

where $\eta\left(\gamma_s, X\right) = \mathbb{E}\left\{exp\left(-\gamma_s A\right) \mid X, S = 1\right\} = exp\left(-\gamma_s\right)\pi_1(X) + \pi_0(X)$. The assumption makes the unobserved density estimable.

5.3 Efficient influence function

In this section, we will obtain the semiparametrically efficient estimator for the ATE by deriving its efficient influence function under the assumed sensitivity framework. To achieve this, it is sufficient to focus on the target estimand $\psi_a(P) = \mathbb{E}\{Y(a)\}$, which is the marginal mean of the potential outcome. We first derive the identification formula for our target estimand in the following theorem.

Theorem 7 (Identification) *Given the target estimand $\psi_a(P) = \mathbb{E}\{Y(a)\}$, under assumptions 6-9, the identification formula is*

$$\begin{aligned}\psi_a(P) = \mathbb{E}\{Y(a)\} \quad = \quad & \mathbb{E}\big[\mu_a(Y;X)\pi_a(X)\omega_1(X)\big] \\ & + \mathbb{E}\left[\frac{\mu_a\big[Y exp\{\gamma_a s_a(Y)\};X\big]}{\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]}\pi_{1-a}(X)\omega_1(X)\right] \\ & + \mathbb{E}\left[\mu_a(Y;X)\pi_a(X)\frac{exp\big(-\gamma_s a\big)}{\eta\big(\gamma_s,X\big)}\omega_0(X)\right] \\ & + \mathbb{E}\left[\frac{\mu_a\big[Y exp\{\gamma_a s_a(Y)\};X\big]}{\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]}\pi_{1-a}(X)\frac{exp\big(-\gamma_s a\big)}{\eta\big(\gamma_s,X\big)}\omega_0(X)\right].\end{aligned}$$

The identification formula for the ATE is therefore $\psi_1(P) - \psi_0(P)$.

Proof 1 *The key idea is to first decompose our target using the law of iterated expectation and then use the assumptions 9 and 8 to estimate the unobserved quantities:*

$$
\begin{aligned}
\psi_a(P)=\mathbb{E}\{Y(a)\} &= \mathbb{E}[\mathbb{E}\{Y(a)\mid X,S\}] \\
&= \mathbb{E}[\mathbb{E}\{Y(a)\mid A=a,X,S=1\}\mathbb{P}(A=a\mid X,S=1)\mathbb{P}(S=1\mid X)] \\
&\quad +\mathbb{E}[\mathbb{E}\{Y(a)\mid A=1-a,X,S=1\}\mathbb{P}(A=1-a\mid X,S=1)\mathbb{P}(S=1\mid X)] \\
&\quad +\mathbb{E}[\mathbb{E}\{Y(a)\mid A=a,X,S=0\}\mathbb{P}(A=a\mid X,S=0)\mathbb{P}(S=0\mid X)] \\
&\quad +\mathbb{E}[\mathbb{E}\{Y(a)\mid A=a,X,S=0\}\mathbb{P}(A=a\mid X,S=0)\mathbb{P}(S=0\mid X)] \\
&= \mathbb{E}\big[\mu_a(Y;X)\pi_a(X)\omega_1(X)\big] \\
&\quad +\mathbb{E}\left[\frac{\mu_a\big[Yexp\{\gamma_a s_a(Y)\};X\big]}{\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]}\pi_{1-a}(X)\omega_1(X)\right] \\
&\quad +\mathbb{E}\left[\mu_a(Y;X)\pi_a(X)\frac{exp\big(-\gamma_s a\big)}{\eta\big(\gamma_s,X\big)}\omega_0(X)\right] \\
&\quad +\mathbb{E}\left[\frac{\mu_a\big[Yexp\{\gamma_a s_a(Y)\};X\big]}{\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]}\pi_{1-a}(X)\frac{exp\big(-\gamma_s a\big)}{\eta\big(\gamma_s,X\big)}\omega_0(X)\right].
\end{aligned}
$$

All the terms in the identification formula are estimable. Thus, we can easily design an estimator from this formula.

$$
\begin{aligned}
\hat{\psi}_{a,\mathrm{id}}=\hat{\mathbb{E}}\{Y(a)\} &= \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mu}_a(Y;X_i)\hat{\pi}_a(X_i)\hat{\omega}_1(X_i)\big] \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{\mu}_a\big[Yexp\{\gamma_a s_a(Y)\};X_i\big]}{\hat{\mu}_a\big[exp\{\gamma_a s_a(Y)\};X_i\big]}\hat{\pi}_{1-a}(X_i)\hat{\omega}_1(X_i)\right] \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_a(Y;X_i)\hat{\pi}_a(X_i)\frac{exp\big(-\gamma_s a\big)}{\hat{\eta}\big(\gamma_s,X_i\big)}\hat{\omega}_0(X_i)\right] \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{\mu}_a\big[Yexp\{\gamma_a s_a(Y)\};X_i\big]}{\hat{\mu}_a\big[exp\{\gamma_a s_a(Y)\};X_i\big]}\hat{\pi}_{1-a}(X)\frac{exp\big(-\gamma_s a\big)}{\hat{\eta}\big(\gamma_s,X_i\big)}\hat{\omega}_0(X_i)\right] \quad (5.1)
\end{aligned}
$$

where the nuisance functions are replaced by their consistent estimates. For example, the conditional expectations $\mu_a(Y;X)$, $\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]$, $\mu_a\big[Yexp\{\gamma_a s_a(Y)\};X\big]$ can be estimated via different regression models, or we can first estimate the conditional density function $dF\{Y\mid A=a,X\}$ and then calculate the expectations by numerical integrals. The probabilities $\pi_a(X)$ and $\omega_s(X)$ can be estimated using a logistic model. More complicated models such as Super Learner (Van der Laan et al. 2007) can also be utilized to increase

prediction accuracy. Sample splitting, or double machine learning (Chernozhukov et al. 2016), may also be applied to remove correlation between nuisance function estimation and causal effect estimation. However, this method will result in the loss of sample size and thus we will not dig into this issue in this paper.

The above estimator $\hat{\psi}_{a,id}$ is not semiparametric efficient because it is only derived from the identification formula. In the following theorem, we derive the efficient influence function in this sensitivity and informative sampling scenario. The proof is provided in the appendix.

Theorem 8 (EIF) *Given the target estimand $\psi_a(P)=\mathbb{E}\{Y(a)\}$, under assumptions 6-9, the efficient influence function is*

$$
\begin{aligned}
\phi_a(P)(O) \;=\;& SI(A=a)(\{Y-\mu_a(Y;X)\}\left\{1+\frac{exp(-\gamma_s a)}{\eta(\gamma_s,X)}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right\} \\
&+\left(\frac{Yexp\{\gamma_a s_a(Y)\}}{\mu_a[exp\{\gamma_a s_a(Y)\};X]}-\frac{exp\{\gamma_a s_a(Y)\}\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]^2}\right) \\
&\cdot\frac{\pi_{1-a}(X)}{\pi_a(X)}\cdot\left[1+\frac{exp\{-\gamma_s(1-a)\}}{\eta(\gamma_s,X)}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right]) && (5.2)\\
&+S\left(\mu_a(Y;X)\pi_a(X)+\frac{\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]}\pi_{1-a}(X)\right)-\psi_a(P) && (5.3)\\
&+S(A-\pi_1(X))(-1)^{a+1}\left(\mu_a(Y;X)-\frac{\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]}\right)\{1+ && (5.4)\\
&\frac{exp(-\gamma_s)}{\eta(\gamma_s,X)^2}\cdot\frac{\omega_0(X)}{\omega_1(X)}\} && (5.5)\\
&+(1-S)\{\mu_a(Y;X)\pi_a(X)\cdot\frac{exp(-\gamma_s a)}{\eta(\gamma_s,X)}+\frac{\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]} && (5.6)\\
&\pi_{1-a}(X)\cdot\frac{exp\{-\gamma_s(1-a)\}}{\eta(\gamma_s,X)}\}. && (5.7)
\end{aligned}
$$

The efficient influence function for the ATE is therefore $\phi_1(P)(O)-\phi_0(P)(O)$.

Similarly, all the terms in the theorem are estimable. We provide the semiparametric efficient estimator for the ATE and its variance formula in the following corollary.

Corollary 2 (semiparametric efficient estimator) *Under assumptions 1-9, the semiparametric efficient estimator for the ATE is*

$$\hat{\tau}_{eif} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_{1,i} - \hat{\phi}_{0,i}\right),$$

where

$$\begin{aligned}
\hat{\phi}_{a,i} \;=\;& S_i I(A_i = a)(\{Y_i - \hat{\mu}_a(Y;X_i)\}\left\{1 + \frac{exp(-\gamma_s a)}{\hat{\eta}(\gamma_s, X_i)} \cdot \frac{\hat{\omega}_0(X_i)}{\hat{\omega}_1(X_i)}\right\} \\
&+\left(\frac{Y_i exp\{\gamma_a s_a(Y_i)\}}{\hat{\mu}_a[exp\{\gamma_a s_a(Y)\};X_i]} - \frac{exp\{\gamma_a s_a(Y_i)\}\hat{\mu}_a[Y exp\{\gamma_a s_a(Y)\};X_i]}{\hat{\mu}_a[exp\{\gamma_a s_a(Y)\};X_i]^2}\right) \\
&\cdot\frac{\hat{\pi}_{1-a}(X_i)}{\hat{\pi}_a(X_i)} \cdot \left[1 + \frac{exp\{-\gamma_s(1-a)\}}{\hat{\eta}(\gamma_s, X_i)} \cdot \frac{\hat{\omega}_0(X_i)}{\hat{\omega}_1(X_i)}\right]) \\
&+S_i\left(\hat{\mu}_a(Y;X_i)\hat{\pi}_a(X_i) + \frac{\hat{\mu}_a[Y exp\{\gamma_a s_a(Y)\};X_i]}{\hat{\mu}_a[exp\{\gamma_a s_a(Y)\};X_i]}\hat{\pi}_{1-a}(X_i)\right) \\
&+S_i(A_i - \hat{\pi}_1(X_i))(-1)^{a+1}\left(\hat{\mu}_a(Y;X_i) - \frac{\hat{\mu}_a[Y exp\{\gamma_a s_a(Y)\};X_i]}{\hat{\mu}_a[exp\{\gamma_a s_a(Y)\};X_i]}\right)\{1 + \frac{exp(-\gamma_s)}{\hat{\eta}(\gamma_s, X_i)^2} \cdot \\
&\frac{\hat{\omega}_0(X_i)}{\hat{\omega}_1(X_i)}\} \\
&+(1-S_i)\{\hat{\mu}_a(Y;X_i)\hat{\pi}_a(X_i) \cdot \frac{exp(-\gamma_s a)}{\hat{\eta}(\gamma_s, X_i)} + \frac{\hat{\mu}_a[Y exp\{\gamma_a s_a(Y)\};X_i]}{\hat{\mu}_a[exp\{\gamma_a s_a(Y)\};X_i]}\hat{\pi}_{1-a}(X_i) \cdot \\
&\frac{exp\{-\gamma_s(1-a)\}}{\hat{\eta}(\gamma_s, X_i)}\}.
\end{aligned}$$

The estimator $\hat{\tau}_{eif}$ is asymptotically normal with asymptotic variance $\sigma^2_{eif} = \mathbb{E}[\{\phi_1(P)(O) - \phi_0(P)(O)\}^2]$, and thus the variance of $\hat{\tau}_{eif}$ can be estimated by

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{\phi}_{1,i} - \hat{\phi}_{0,i} - \hat{\tau}_{eif}\right)^2. \tag{5.8}$$

The corollary is trivial from the asymptotic theory of the influence function.

We explain the meanings of terms in Theorem (8). (5.2) is the main term for collecting direct information from sampled subjects in the corresponding treatment group. This is an inverse probability weighting form adapted by the sensitivity parameters. To see this, if we set $\gamma_s = \gamma_1 = \gamma_0 = 0$, (5.2) will reduce to $SI(A=a)\{Y - \mu_a(Y;X)\}/\{\pi_a(X)\omega_1(X)\}$. (5.3) is the main augmentation term from sampled subjects, which basically estimates $\mathbb{P}(S=1 \mid X) \cdot \mathbb{E}\{Y(a) \mid X\}$. (5.5) is the minor augmentation term from sampled subjects, and

it will vanish when $\gamma_a = 0$. (5.7) is the augmentation term from the non-sampled subjects by utilizing their covariate information to remove the selection bias. $exp\left(-\gamma_s a\right)/\eta\left(\gamma_s, X\right)$ and $exp\left\{-\gamma_s(1-a)\right\}/\eta\left(\gamma_s, X\right)$ can be treated as weights determined by γ_s. When $\gamma_s = 0$, the two weights are both equal to 1, and (5.7) regards an estimate of $\mathbb{P}(S=0 \mid X)\cdot\mathbb{E}\{Y(a) \mid X\}$. Combining (5.3) and (5.7) will lead to the estimate of $\mathbb{E}\{Y(a) \mid X\}$.

Following the above analysis, the next corollary focuses on the special case when no unmeasured confounder assumption is met. When the sensitivity parameters are all zero, the above estimator is equivalent to the standard augmented inverse probability weighting (AIPW) estimator.

Corollary 3 (EIF) *When $\gamma_s = \gamma_1 = \gamma_0 = 0$, given the target estimand $\psi_a(P) = \mathbb{E}\{Y(a)\}$, under assumptions 6-9, the efficient influence function reduces to the standard efficient influence function following the form of AIPW estimator.*

$$\phi_a(P)(O) = \frac{SI(A=a)}{\pi_a(X)\,\omega_1(X)}\left\{Y - \mu_a(Y;X)\right\} + \mu_a(Y;X) - \psi_a(P).$$

The proof for the corollary is trivial.

5.4 Calibration of sensitivity parameters

A key problem in the logistic selection specification of the sensitivity framework is the value determination of the sensitivity parameters. Even though we are willing to assume these parameters to be independent of covariates (Jung et al. 2018), the sensitivity parameters cannot be directly identified from the observed data. On one hand, prior field-specific knowledge may help us on the parameter selection. On the other hand, we offer a data-driven method to calibrate sensitivity parameters following Franks et al. (2020). The key idea is to quantify the maximum effect using the partial variance explained by the covariate. To illustrate the idea, we focus on the sensitivity model between sampling and treatment. By Assumption (9), we reformulate the sampling process using the latent variable model:

$$\begin{aligned} Z_{s_1} &= m(X,A) + \epsilon_1 \\ S &= \begin{cases} 1 & \text{if } Z_{s_1} \geq 0 \\ 0 & \text{if } Z_{s_1} < 0, \end{cases} \end{aligned} \tag{5.9}$$

where ϵ_1 follows the standard logistic distribution and $m(X,A) = \text{logit}\mathbb{P}(S=1 \mid X,A) = \alpha(X)+\gamma_s A$ is the linear predictor. We define the partial variance explained by X and A as

$$\rho^2_{X,A} = \frac{\text{var}\{m(X,A)\}}{\text{var}\{m(X,A)\}+\pi^2/3}, \tag{5.10}$$

where $\pi^2/3$ comes from the variance of a standard logistic random variable. On the other hand, if we ignore A and only use X to construct the sampling model, the latent variable is

$$\begin{aligned} Z_{s_2} &= m(X)+\epsilon_2 \\ S &= \begin{cases} 1 & \text{if } Z_{s_2} \geq 0 \\ 0 & \text{if } Z_{s_2} < 0, \end{cases} \end{aligned} \tag{5.11}$$

where ϵ_2 follows the standard logistic distribution and $m(X) = \text{logit}\mathbb{P}(S=1 \mid X) = \text{logit}\ \omega_1(X)$. Similarly, the partial variance explained by X is

$$\rho^2_X = \frac{\text{var}\{m(X)\}}{\text{var}\{m(X)\}+\pi^2/3}. \tag{5.12}$$

Then, the partial variance explained by A given X is

$$\rho^2_{A|X} = \frac{\rho^2_{X,A}-\rho^2_X}{1-\rho^2_X}. \tag{5.13}$$

Because A is missing for $S=0$, the sampling process cannot be fitted by X and A from the observed data. Thus, $\rho^2_{A|X}$ cannot be estimated in practice. To deal with this problem, we assume the information obtained by A given X will not exceed the information obtained by any covariate X_j given all the other covariates X_{-j}: $\rho^2_{A|X} \leq \rho^2_{X_j|X_{-j}}$ for $j=1,\ldots,p$, where $\rho^2_{X_j|X_{-j}}$ can be estimated using observed data following (5.10), (5.12), and (5.13). The next theorem connect the sensitivity parameter γ_s with $\rho^2_{A|X}$.

Theorem 9 (Calibrate γ_s) *The relationship between sensitivity parameter γ_s with partial variance explained by A given X is*

$$\begin{aligned} \rho^2_{A|X} &= 1-\frac{var\{m(X)\}+\pi^2/3}{var\{m^*(X)\}+\pi^2/3+\sigma^2_A\gamma^2_s} \\ \left|\gamma_s\right| &= \frac{1}{\sigma_A}\sqrt{\frac{var\{m(X)\}+\pi^2/3}{1-\rho^2_{A|X}}-var\{m^*(X)\}-\pi^2/3}, \end{aligned}$$

where $\sigma_A^2 = \mathbb{E}\{var(A \mid X)\} = \mathbb{E}[e(X)\{1-e(X)\}]$, $m^*(X) = \alpha(X) + \gamma_s e(X)$, *and* $e(X) = \mathbb{P}(A = 1 \mid X)$ *is the propensity score. If we assume* $m^*(X) \approx m(X)$,

$$\rho_{A|X}^2 \approx \frac{\sigma_A^2 \gamma_s^2}{var\{m(X)\} + \pi^2/3 + \sigma_A^2 \gamma_s^2} \tag{5.14}$$

$$\left|\gamma_s\right| \approx \frac{1}{\sigma_A} \sqrt{\frac{\rho_{A|X}^2}{1-\rho_{A|X}^2}[var\{m(X)\} + \pi^2/3]}. \tag{5.15}$$

The proof of the theorem is included in the appendix. We would like to emphasize that the theorem is related to but different from the Proposition 3 in Franks et al. (2020), where the authors treat $m^*(X)$ and $m(X)$ as the same. $m^*(X)$ comes from a reformulation of (5.9):

$$\begin{aligned} m(X,A) &= \alpha(X) + \gamma_s A \\ &= \alpha(X) + \gamma_s e(X) + \gamma_s \{A - e(X)\} \\ &= m^*(X) + \widetilde{\gamma}_s \widetilde{R}_A(X), \end{aligned}$$

where $\widetilde{\gamma}_s = \sigma_A \gamma_s$ and $\widetilde{R}_A(X) = \{A - e(X)\}/\sigma_A$. We can see that $m^*(X) = \mathbb{E}\{m(X,A) \mid X\} = \mathbb{E}\left\{\text{logit}\mathbb{E}(S \mid X, A) \mid X\right\} \neq \text{logit}\mathbb{E}\{\mathbb{E}(S \mid X, A) \mid X\} = \text{logit}\mathbb{E}(S \mid X) = m(X)$ because logit function is not linear. However, $m^*(X)$ is not obtainable without knowing γ_s. Consider that we only need a rough estimate of γ_s, we can approximate $m^*(X)$ by the obtainable term $m(X)$ to make the calibration formula (5.15) applicable. Meanwhile, it is worth pointing out that

$$\begin{aligned} e(X) &= \mathbb{E}(A \mid X) = \mathbb{E}(\mathbb{E}(A \mid X, S) \mid X) \\ &= \mathbb{P}(A = 1 \mid X, S = 1)\omega_1(X) \\ &\quad + \mathbb{P}(A = 1 \mid X, S = 0)\omega_0(X) \\ &= \pi_1(X)\left\{\omega_1(X) + \frac{exp(-\gamma_s)\omega_0(X)}{\eta(\gamma_s, X)}\right\} \end{aligned}$$

also depends on γ_s. Thus, (5.15) is in fact an equation of γ_s, and its calibration can be obtained by solving this equation. Similarly, the following theorem gives the calibration of γ_a. Note that we need to assume the linear predictor of A is linear in $Y(a)$, i.e., $s\{Y(a)\} = Y(a)$, following Franks et al. (2020).

Theorem 10 (Calibrate γ_a) *Suppose logit$\mathbb{P}(A=1-a \mid X, Y(a))$ is linear in $Y(a)$. The relationship between sensitivity parameter γ_a with partial variance explained by $Y(a)$ given X is*

$$\begin{aligned}\rho^2_{Y(a)|X} &= 1-\frac{var\{n(X)\}+\pi^2/3}{var\{n^*(X)\}+\pi^2/3+\sigma^2_{Y(a)}\gamma_a^2} \\ \left|\gamma_a\right| &= \frac{1}{\sigma_A}\sqrt{\frac{var\{n(X)\}+\pi^2/3}{1-\rho^2_{Y(a)|X}}-var\{n^*(X)\}-\pi^2/3},\end{aligned}$$

where $\sigma^2_{Y(a)}=\mathbb{E}[var\{Y(a) \mid X, S=1\}]$, $n(X)=$ logit$\mathbb{P}(A=1-a \mid X, S=1)=\pi_{1-a}(X)$, $n^*(X)=\beta_a(X)+\gamma_a\left[s_a\{Y(a)\}-\mu_a(X)\right]$, and $\mu_a(X)=\mathbb{E}\{Y(a) \mid X, S=1\}$. If we assume $n^*(X)\approx n(X)$,

$$\rho^2_{Y(a)|X} \approx \frac{\sigma^2_{Y(a)}\gamma_a^2}{var\{n(X)\}+\pi^2/3+\sigma^2_{Y(a)}\gamma_a^2} \tag{5.16}$$

$$\left|\gamma_a\right| \approx \frac{1}{\sigma_{Y(a)}}\sqrt{\frac{\rho^2_{Y(a)|X}}{1-\rho^2_{Y(a)|X}}[var\{n(X)\}+\pi^2/3]}. \tag{5.17}$$

We omit the proof of this theorem because it is mostly the same as the proof for the previous theorem. Note the regression model is applied only on the sampled subjects because A is missing for non-sampled subjects. Also, $\sigma_{Y(a)}$ depends on the sensitivity parameter γ_a, and so γ_a is calibrated by solving (5.17). To estimate $\sigma_{Y(a)}$, note that

$$\text{var}\{Y(a) \mid X, S=1\}=\mathbb{E}\left\{Y(a)^2 \mid X, S=1\right\}-\mathbb{E}\{Y(a) \mid X, S=1\}^2,$$

where

$$\begin{aligned}\mathbb{E}\left\{Y(a)^2 \mid X, S=1\right\} &= \mathbb{E}\left\{Y^2 \mid A=a, X, S=1\right\}\mathbb{P}(A=a \mid X, S=1) \\ &\quad +\frac{\mathbb{E}\left[Y^2 exp\left\{\gamma_a s_a(Y)\right\} \mid A=a, X, S=1\right]}{\mathbb{E}\left[exp\left\{\gamma_a s_a(Y)\right\} \mid A=a, X, S=1\right]}\mathbb{P}(A=1-a \mid X, S=1) \\ \mathbb{E}\{Y(a) \mid X, S=1\} &= \mathbb{E}\{Y \mid A=a, X, S=1\}\mathbb{P}(A=a \mid X, S=1) \\ &\quad +\frac{\mathbb{E}\left[Y exp\left\{\gamma_a s_a(Y)\right\} \mid A=a, X, S=1\right]}{\mathbb{E}\left[exp\left\{\gamma_a s_a(Y)\right\} \mid A=a, X, S=1\right]}\mathbb{P}(A=1-a \mid X, S=1).\end{aligned}$$

The conditional expectations $\mathbb{E}\left[exp\left\{\gamma_a s_a(Y)\right\} \mid A=a, X, S=1\right]$, $\mathbb{E}[Y exp\left\{\gamma_a s_a(Y)\right\} \mid A=$

$a, X, S=1]$, $\mathbb{E}\left[Y^2 exp\left\{\gamma_a s_a(Y)\right\} \mid A=a, X, S=1\right]$ need to be reevaluated when γ_a changes. To alleviate computational burden, we suggest first estimate the conditional probability density function $dF\{Y(a) \mid A=a, X, S=1\}$, for example, using the semi-parametric quantile regression (SPQR) method (Xu and Reich 2023), and calculate the conditional expectations by numerical integrals.

5.5 Simulation

In this section, we design a simulation study to compare the numerical performance of EIF-based estimator with other existing estimators. We include the estimator $\hat{\tau}_{id} = \hat{\psi}_{1,\text{id}} - \hat{\psi}_{0,\text{id}}$, where $\hat{\psi}_{a,\text{id}}$ is defined in (5.1) and derived from the identification formula. Similar to the EIF-based estimator, $\hat{\tau}_{id}$ also takes account of the sensitivity analysis framework, but it may not achieve the semiparametric efficiency bound. We also compare our estimator to the standard weighting estimators without taking unmeasured confounders into consideration, including the inverse probability weighting (IPW) estimator and the AIPW estimator (Robins et al. 1994):

$$\begin{aligned}
\hat{\tau}_{ipw} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{SI(A=a)Y}{\pi_a(X)\omega_1(X)} - \frac{SI(A=1-a)Y}{\pi_{1-a}(X)\omega_1(X)}\right\} \\
\hat{\tau}_{aipw} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{SI(A=a)}{\pi_a(X)\omega_1(X)}\left\{Y-\mu_a(Y;X)\right\}+\mu_a(Y;X)\right] \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{SI(A=1-a)}{\pi_{1-a}(X)\omega_1(X)}\left\{Y-\mu_{1-a}(Y;X)\right\}+\mu_{1-a}(Y;X)\right].
\end{aligned}$$

We use Super Learner to estimate the nuisance functions, where the base algorithms include penalized regression, generalized additive model, and neural network.

The data generating process of the simulation study is introduced as follows. We first generate the covariates $X=\left(X_1,...,X_p\right)$ following independent standard normal distributions, where $p=5$ is the number of covariates and $n=3000$ is the sample size. We then generate the sampling indicator S with sampling probability

$$\omega_1(X)=\mathbb{P}(S=1 \mid X)=\text{logit}^{-1}\left\{0.2\cdot(1+X_1-X_2+2X_3-2X_4+0.5X_5)\right\}.$$

The probability to receive treatment for sampled subjects is

$$\pi_1(X) = \mathbb{P}(A=1 \mid X, S=1) = \text{logit}^{-1}\{0.2 \cdot (1 + X_1 + X_2 - 2X_3 - 0.5X_4 + 2X_5)\},$$

and the treatment probability for non-sampled subjects can be derived using assumption (9):

$$\mathbb{P}(A=0 \mid X, S=1) = \text{logit}^{-1}\left\{0.2 \cdot (1 + X_1 + X_2 - 2X_3 - 0.5X_4 + 2X_5) - \gamma_s\right\},$$

where $\gamma_s = 2$ in our simulation. Although we do not the treatment indicators for non-sampled subjects, these are essential to calculate the true ATE. The potential outcomes are similarly generated by groups: for treatment group $A=1$,

$$Y(1) = 0.25 \cdot (-1 - X_1 + 2X_2 + 0.5X_3 - X_4 + 2X_5) + \epsilon_1,$$

where $\epsilon_1 \sim \mathcal{N}\left(0, \sigma_1^2\right)$ and $\sigma_1 = 1$. While for control group $A=0$,

$$Y(1) = 0.25 \cdot (-1 - X_1 + 2X_2 + 0.5X_3 - X_4 + 2X_5) + \sigma_1^2\gamma_1 + \epsilon_1,$$

where γ_1 is set to be 0.2. It is easy to verify that the above formulation satisfies the assumption (8). Similarly, we first generate controlled outcome for control group $A=0$:

$$Y(0) = 0.25 \cdot (-1 - X_1 + 2X_2 + 0.5X_3 - X_4 + 2X_5) + \epsilon_0,$$

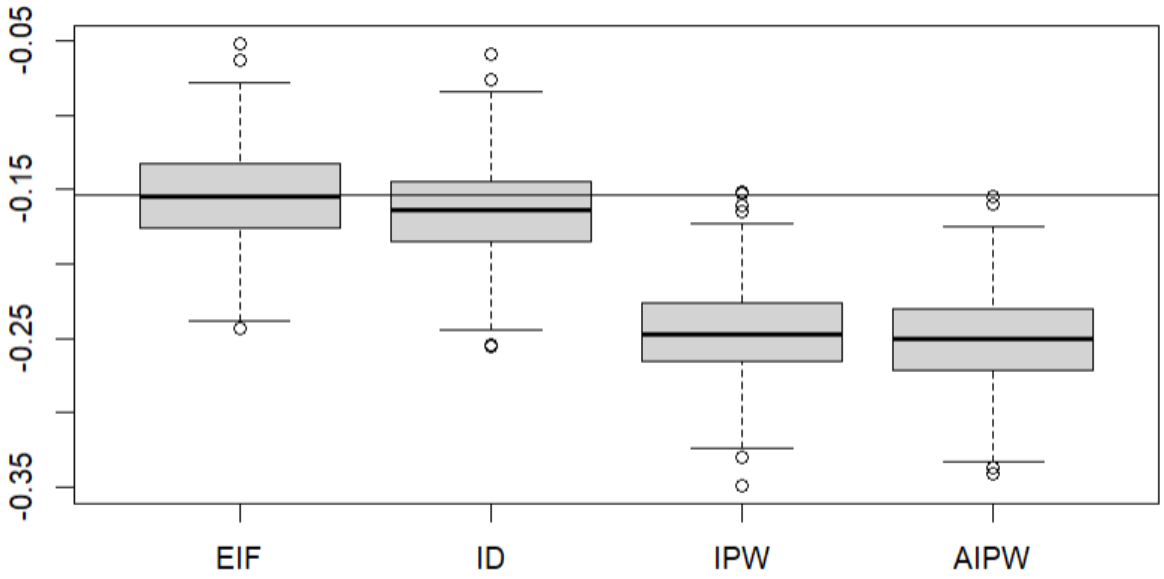
where $\epsilon_0 \sim \mathcal{N}\left(0, \sigma_0^2\right)$ and $\sigma_0 = 1$. While for treatment group $A=1$,

$$Y(0) = 0.25 \cdot (-1 - X_1 + 2X_2 + 0.5X_3 - X_4 + 2X_5) + \sigma_0^2\gamma_0 + \epsilon_0,$$

where $\gamma_0 = 0.2$. We replicate the simulation for 300 times, and the true ATE is -0.15.

Figure shows the distribution of the estimators. We can see that our semiparametric efficient estimator $\hat{\tau}_{\text{eif}}$ achieves the smallest variance without significant bias. The identification-formula-based estimator $\hat{\tau}_{id}$ is biased and has slightly larger variance. On the other hand, both the IPW and AIPW estimator are significantly biased because they do not consider the sensitivity framework in the dataset. We also construct the 95% confidence interval for $\hat{\tau}_{\text{eif}}$ with variance estimated from (5.8). The coverage rate is 95.7% with monte carlo standard error 1.2%, showing that our variance estimator is valid.

Figure 5.1: Simulation results of various estimators.



5.6 Real data application

Per- and polyfluoroalkyl substances (PFAS) encompass a large group of nearly 5,000 synthetic chemicals known for their resistance to heat, water, and oil. Several extensively studied PFAS have been identified as hazardous to human health (Al Amin et al. 2020). In this study, we utilize data from the Groundwater Ambient Monitoring and Assessment Program (GAMA) available in the Groundwater Information System (GIS). Specifically, we analyze three GAMA datasets: Division of Drinking Water (DDW), Local Groundwater Projects (LOCALGW), and GAMA US Geological Survey (USGS). These datasets provide information on PFAS and copollutant concentrations in wells used for drinking water across California from 2010 to 2022. However, it is important to note that the sampling process for selecting locations for PFAS measurement exhibits informativeness. Locations with higher concentrations of PFAS have a greater probability of being sampled. Despite efforts to control certain covariates, such as distance to military bases, unmeasured confounders that influence both the sampling process and PFAS levels still exist. Neglecting this sensitivity analysis may result in sampling bias. Moreover, there may also be unmeasured confounders between PFAS exposure and the health outcome of interest. Our focus is specifically on estimating the causal effect of PFAS on the hypertension incidence rate in zip codes with more than 10 beneficiaries during the 7-year period. The incidence rate is calculated as

the 7-year incidence divided by person-time multiplied by the number of people, and the rates are expressed per 100,000 individuals. We aim to estimate this causal effect within the framework of sensitivity analysis.

Given that health outcome data is collected at the zip code level, we define the binary treatment indicator as the detection of PFAS. It is assigned a value of 1 if at least one PFAS chemical exceeds the detection limit in any well within the zip code area. In cases where none of the wells in a zip code are selected for PFAS detection, we consider those locations as non-sampled. The proportion of sampled zip code locations is 35.9%, which is 420 out of 1171 locations. Among the locations where PFAS is measured, approximately 59.8% exhibit the detection of at least one PFAS chemical. Covariate data are collected from population and landcover data within a 1000m radius of each well, and the average values are aggregated at the zip code level. The proportion of missing covariates is less than 1%, allowing us to impute missing values using the "mice" package in R.

We first calibrate sensitivity parameters following Section 5.4. We remove each covariate from the variable list and calculate its partial variance explained given all the other covariates. Super Learner is used to estimate the nuisance functions, where the base algorithms include penalized regression, random forest, generalized additive model, extreme gradient boosting, and neural network. The maximum absolute value of the sensitivity parameter between sampling and PFAS γ_s is 1.98. On the other hand, 21.74 is the maximum absolute value of the sensitivity parameter between PFAS and health outcome under treated γ_1 and 16.90 is the maximum absolute value for γ_0 respectively. Thus, we set $\gamma_s = 2$ and $\gamma_1 = \gamma_0 = 20$. The EIF-based estimator is applied to estimate the ATE. The estimated ATE is 0.0179, and its estimated standard error is 0.0023. The 95% confidence interval of ATE is $(0.0133, 0.0224)$, which means the ATE is significantly larger than zero. Thus, the existence of PFAS chemical will increase the hypertension incidence rate.

On the other hand, we compare our result to the standard AIPW estimator. The estimated ATE is -0.00015, and its estimated standard error is 0.0009. The corresponding 95% confidence interval $(-0.0019, 0.0016)$ covers zero, leading to the non-significant result. This is different from the result we get from the EIF-based estimator, implying that ignoring the sensitivity framework may lead to the opposite conlusion.

5.7 Discussion

In this paper, our primary focus lies on the sensitivity analysis framework under informative sampling. The semiparametric efficient estimator overcomes selection bias caused by unmeasured confounders while simultaneously achieving the minimum asymptotic variance. Furthermore, it leverages information from non-sampled subjects to enhance estimator efficiency. However, it is important to acknowledge that in practical applications, the estimator's stability may be compromised by numerical errors incurred during the fitting of nuisance functions. The inverse probability weighting formulation employed in the estimator can be susceptible to outliers. To address this issue, one potential approach is to truncate the denominator, excluding extreme values (Yang and Ding 2018). Alternatively, employing more sophisticated algorithms can enhance prediction accuracy, and the application of double machine learning can help control the convergence rate (Chernozhukov et al. 2016).

Additionally, it is important to consider the limitations of this work. The semiparametric efficient estimator is based on the logistic selection specification, which assumes a parametric model that may not accurately reflect real-world scenarios. While the sensitivity analysis framework is adopted to alleviate the assumption of no unmeasured confounders, our intention is not to replace one unrealistic assumption with another. Recently, researchers have been using negative control exposures or outcomes to assess the bias caused by unmeasured confounding (Lipsitch et al. 2010). We anticipate future work that utilizes negative control methods to address the sensitivity analysis problem in the context of informative sampling. Moreover, if we have confidence in the logistic selection specification, negative control methods can also assist in calibrating the sensitivity parameters.

CHAPTER

6

CONCLUSIONS

The assumption of no unmeasured confounders is the cornerstone of causal inference in observational studies. However, its verification using observed data has always been a challenge, raising questions about its reliability. One approach to enhance the validity of this assumption is to collect a comprehensive set of variables to ensure the inclusion of important confounders. Nevertheless, in high-dimensional settings, dimension reduction techniques become crucial, typically through the use of propensity and prognostic scores. In addition, existing estimators often rely on the correct specification of the model, underscoring the need for robustness.

In this dissertation, our focus is on developing doubly robust estimators for observational studies, aiming to address the theoretical gaps in doubly robust matching and regression estimators by establishing their asymptotic properties. Furthermore, we improve the existing double score matching estimator to achieve multiple robustness, enhancing its performance in practical applications.To enhance the finite sample performance of matching estimators, we propose the utilization of variable selection and matching with replacement, particularly valuable in high-dimensional settings. By incorporating treatment information into the outcome model using a partial linear regression model, our doubly robust regression estimator not only achieves doubly robustness but also attains

the semiparametric efficiency bound when both models are correctly specified.

Recognizing the limitations of solely increasing the number of covariates, we also acknowledge the potential existence of unmeasured confounders and advocate for their assessment within the sensitivity analysis framework. Furthermore, we extend this framework to incorporate informative sampling, resulting in the derivation of a semiparametric efficient estimator. Considering the ongoing controversy surrounding the selection of sensitivity parameters, we adapt existing methods to calibrate these parameters, ensuring a robust and reliable estimation process.

Overall, this dissertation offers novel approaches to address the challenges associated with the no unmeasured confounders assumption from both theoretical and practical perspectives. By incorporating multiple robustness, dimension reduction techniques, and sensitivity analysis, we aim to enhance the validity and reliability of causal inference in observational studies.

The work presented in this dissertation provides valuable insights into causal inference in a basic setting, laying the foundation for future extensions to more complex scenarios. For instance, in practical applications, treatments are often administered at varying levels of intensity or continuously, necessitating the development of novel estimators to capture the effects of such treatments (Yang et al. 2016). Exploring these new definitions of treatment effects will undoubtedly contribute to the advancement of causal inference methods.

Furthermore, the assignment of treatment may occur at multiple time points, and the collection of covariate information may span a long period, giving rise to longitudinal studies and dynamic treatment regimes (Murphy 2003). These settings pose unique challenges and offer opportunities for further research and methodological development in causal inference. Additionally, the concept of precision medicine has gained prominence, wherein treatments are tailored to individuals based on their covariate profiles (Tsiatis et al. 2019). The development of optimal individualized treatment regimes becomes crucial in this context, and it represents an important area for future investigation.

Moreover, the sensitivity analysis framework can be enhanced by leveraging information from negative control exposures and outcomes (Lipsitch et al. 2010). By incorporating these additional sources of information, we can better relax the strong parametric assumptions and gain a deeper understanding of the impact of unmeasured confounders.

We anticipate that future research in causal inference will continue to explore these avenues and tackle the complexities of real-world settings. It is our hope that these advancements will lead to more interesting and impactful work, further improving our ability to draw reliable causal conclusions and inform decision-making processes in various domains.

REFERENCES

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

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76:1537–1557.

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

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *J Am Stat Assoc*, 107:833–843.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Abadie, A. and Spiess, J. (2016). Robust post-matching inference. In *Unpublished Paper. MIT and Harvard University. Retrieved from https://editorialexpress. com/cgi-bin/conference/down load.*

Aikens, R. C., Greaves, D., and Baiocchi, M. (2020). A pilot design for observational studies: Using abundant data thoughtfully. *Statistics in Medicine*, 39(30):4821–4840.

Al Amin, M., Sobhani, Z., Liu, Y., Dharmaraja, R., Chadalavada, S., Naidu, R., Chalker, J. M., and Fang, C. (2020). Recent advances in the analysis of per-and polyfluoroalkyl substances (pfas)—a review. *Environmental technology & innovation*, 19:100879.

Andreou, E. and Werker, B. J. (2012). An alternative asymptotic analysis of residual-based statistics. *Rev Econ Stat*, 94:88–99.

Antonelli, J., Cefalu, M., Palmer, N., and Agniel, D. (2018). Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179.

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.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973.

Bickel, P. J., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Billingsley, P. (1995). *Probability and Measure*. Wiley: New York, 3 edition.

Blake, B. E. and Fenton, S. E. (2020). Early life exposure to per-and polyfluoroalkyl substances (pfas) and latent health outcomes: A review including the placenta as a target tissue and possible driver of peri-and postnatal effects. *Toxicology*, 443:152565.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, pages 136–146.

Chen, H. and Shiau, J.-J. H. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference*, 27(2):187–201.

Chen, H. and Shiau, J.-J. H. (1994). Data-driven efficient estimators for a partially linear model. *The Annals of Statistics*, pages 211–237.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *J. Offic. Stat.*, 16:113–131.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *J Am Stat Assoc*, 96:260–269.

Chen, S. and Haziza, D. (2017a). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104:439–453.

Chen, S. and Haziza, D. (2017b). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Stat. Sin*, 29:2035–2053.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. (2016). Double machine learning for treatment and causal parameters. Technical report, Cemmap Working Paper, Centre for Microdata Methods and Practice.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81:535–580.

Chiang, C.-T. and Huang, M.-Y. (2012). New estimation and inference procedures for a single-index conditional distribution model. *Journal of Multivariate Analysis*, 111:271–285.

Ciociola, A. A., Cohen, L. B., Kulkarni, P., Kefalas, C., Buchman, A., Burke, C., Cain, T., Connor, J., Ehrenpreis, E. D., Fang, J., et al. (2014). How drugs are developed and approved by the fda: current process and future directions. *Official journal of the American College of Gastroenterology| ACG*, 109(5):620–623.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(4):417–446.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Rev Econ Stat*, 84:151–161.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26.

Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394):310–320.

Faries, D., Zhang, X., Kadziola, Z., Siebert, U., Kuehne, F., Obenchain, R. L., and Haro, J. M. (2020). *Real World Health Care Data Analysis: Causal Methods and Implementation Using SAS*. SAS Institute.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75:259–276.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923):554.

Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete data in sample surveys*, 2(Part IV):185–207.

Foresi, S. and Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90:451–466.

Francisco, C. A. and Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19:454–469.

Franks, A., D'Amour, A., and Feller, A. (2020). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115(532):1730–1746.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Rev Econ Stat*, 86:77–90.

Ghosh, J. K. (1971). A new proof of the bahadur representation of quantiles and an application. *The Annals of Mathematical Statistics*, 42:1957–1961.

Grandjean, P. and Clapp, R. (2015). Perfluorinated alkyl substances: emerging insights into health risks. *New solutions: a journal of environmental and occupational health policy*, 25(2):147–163.

Guo, S. and Fraser, M. W. (2014). *Propensity Score Analysis: Statistical Methods and Applications*, volume 11. Thousand Oaks, CA: SAGE.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331.

Hahn, J. and Ridder, G. (2013). Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81:315–340.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *J Am Stat Assoc*, 109(507):1159–1173.

Han, P., Kong, L., Zhao, J., and Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:305–333.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100:417–430.

Hansen, B. B. (2006). Bias reduction in observational studies via prognosis scores. Technical report, Technical Report 441, University of Michigan, Statistics Department.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.

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

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997a). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64:605–654.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997b). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64:605–654.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):244–248.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Stat Assoc*, 96:440–448.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*, 47:663–685.

Hu, Z., Follmann, D. A., and Qin, J. (2012). Semiparametric double balancing score estimation for incomplete data with ignorable missingness. *Journal of the American Statistical Association*, 107(497):247–257.

Huang, M.-Y. and Yang, S. (2021). Robust inference of conditional average treatment effects using dimension reduction. *Statistica Sinica*, page doi:10.5705/ss.202020.0409.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev Econ Stat*, 86:4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge UK.

Jung, J., Shroff, R., Feller, A., and Goel, S. (2018). Algorithmic decision making in the presence of unmeasured confounding. *arXiv preprint arXiv:1805.01868*.

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

King, G. and Nielsen, R. A. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.

Kumamaru, H., Schneeweiss, S., Glynn, R. J., Setoguchi, S., and Gagne, J. J. (2016). Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. *Emerging Themes in Epidemiology*, 13:Article 5.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76:604–620.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer: Berlin.

Leacy, F. P. and Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine*, 33(20):3488–3508.

Lee, Y.-Y. (2018). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. *arXiv preprint arXiv:1811.00157*.

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

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*, 9(2):215–234.

Li, W., Yang, S., and Han, P. (2020). Robust estimation for moment condition models with data missing not at random. *Journal of Statistical Planning and Inference*, 207:246–254.

Lin, Z., Ding, P., and Han, F. (2021). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *arXiv preprint arXiv:2112.13506*.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, pages 949–968.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken.

Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14:570–580.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*, 23:2937–2960.

Mammen, E., Rothe, C., and Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2):1132–1170.

Mammen, E., Rothe, C., and Schienle, M. (2016). Semiparametric estimation with generated covariates. *Econometric Theory*, 32(5):1140–1177.

Matsouaka, R. A. and Zhou, Y. (2020). A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights. *arXiv preprint arXiv:2011.01388*.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11):1213–1222.

Naik, C., McCoy, E. J., and Graham, D. J. (2016). Multiply robust dose-response estimation for multivalued causal inference problems. *arXiv preprint arXiv:1611.02433*.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62:1349–1382.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics*, 79:147–168.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T. *Statistical Science*, 5:465–472.

Nguyen, T.-L. and Debray, T. P. (2019). The use of prognostic scores for causal inference with general treatment regimes. *Statistics in medicine*, 38(11):2013–2029.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *J Am Stat Assoc*, 112:1720–1732.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, pages 221–247.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2019). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*.

Pearl, J. (2011). Invited commentary: understanding bias amplification. *American Journal of Epidemiology*, 174(11):1223–1227.

Peng, X., Robinson, R. L., Mease, P., Kroenke, K., Williams, D. A., Chen, Y., Faries, D., Wohlreich, M., McCarberg, B., and Hann, D. (2015). Long-term evaluation of opioid treatment in fibromyalgia. *The Clinical Journal of Pain*, 31(1):7–13.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics & probability letters*, 4(4):203–208.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94. Springer, New York: Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Robinson, R. L., Kroenke, K., Mease, P., Williams, D. A., Chen, Y., D'Souza, D., Wohlreich, M., and McCarberg, B. (2012). Burden of illness and treatment patterns for patients with fibromyalgia. *Pain Medicine*, 13(10):1366–1376.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *J Am Stat Assoc*, 84:1024–1032.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York, 2 edition.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 45:212–218.

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

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

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*, 66:688–701.

Rubin, D. B. (1978a). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

Rubin, D. B. (1978b). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the Section on Survey Research Methods*, page 20. The Association.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *J Am Stat Assoc*, 75:591–593.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292.

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):169–188.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, England.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*, 26:20–36.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52(1):249–264.

Scharfstein, D. O., Nabi, R., Kennedy, E. H., Huang, M.-Y., Bonvini, M., and Smid, M. (2021). Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc*, 94:1096–1120.

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

Stuart, E. A., Lee, B. K., and Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8):S84–S90.

Sunderland, E. M., Hu, X. C., Dassuncao, C., Tokranov, A. K., Wagner, C. C., and Allen, J. G. (2019). A review of the pathways of human exposure to poly-and perfluoroalkyl substances (pfass) and present understanding of health effects. *Journal of exposure science & environmental epidemiology*, 29(2):131–147.

Tang, D., Kong, D., Pan, W., and Wang, L. (2020). Outcome model free causal inference with ultra-high dimensional covariates.

Tang, S., Yang, S., Wang, T., Cui, Z., Li, L., and Faries, D. E. (2019). Causal inference of hazard ratio based on propensity score matching. *arXiv preprint arXiv:1911.12430*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. CRC press.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Stat Appl Genet Mol Biol*, 6(1).

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge, MA.

Waernbaum, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in medicine*, 31(15):1572–1581.

Wang, L. (2019). Multiple robustness estimation in causal inference. *Communications in Statistics-Theory and Methods*, 48:5701–5718.

Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., and Xia, J. (2013). Optimal caliper width for propensity score matching of three treatment groups: a monte carlo study. *PloS ONE*, 8(12):e81045.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer, New York, 2 edition.

Wyss, R., Ellis, A. R., Brookhart, M. A., Jonsson Funk, M., Girman, C. J., Simpson Jr, R. J., and Stürmer, T. (2015). Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety*, 24(9):951–961.

Wyss, R., Glynn, R. J., and Gagne, J. J. (2016). A review of disease risk scores and their application in pharmacoepidemiology. *Current Epidemiology Reports*, 3(4):277–284.

Xu, S. G. and Reich, B. J. (2023). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, 79(1):151–164.

Yang, S. (2018). Propensity score weighting for causal inference with clustered data. *Journal of Causal Inference*, pages doi.org/10.1515/jci–2017–0027.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105:487–493.

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

Yang, S. and Kim, J. K. (2018). Nearest neighbor imputation for general parameter estimation in survey sampling. *Advances in Econometrics*, 39:211–236.

Yang, S. and Kim, J. K. (2019). Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scand J Stat*, page doi.org/10.1111/sjos.12429.

Yang, S., Kim, J. K., and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):445–465.

Yang, S. and Lok, J. J. (2017). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, 28:1703–1723.

Yang, S. and Zhang, Y. (2020). Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*.

Zhang, B., He, Y., Huang, Y., Hong, D., Yao, Y., Wang, L., Sun, W., Yang, B., Huang, X., Song, S., et al. (2020). Novel and legacy poly-and perfluoroalkyl substances (pfass) in indoor dust from urban, industrial, and e-waste dismantling areas: the emergence of pfas alternatives in china. *Environmental Pollution*, 263:114461.

Zhang, G. and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3):911–918.

Zhang, Z., Chen, Z., Troendle, J. F., and Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics*, 68:697–706.

Zhao, H. and Yang, S. (2021). Outcome-adjusted balance measure for generalized propensity score model selection.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Rev Econ Stat*, 86:91–107.

Zhou, T., Elliott, M. R., and Little, R. J. (2019). Penalized spline of propensity methods for treatment comparison. *J Am Stat Assoc*, 114:1–19.

APPENDICES

APPENDIX

A

APPENDIX FOR "MULTIPLY ROBUST MATCHING ESTIMATORS FOR AVERAGE AND QUANTILE TREATMENT EFFECTS"

Sections A.1, A.3, A.4, and A.5 present the proofs of Theorems 1, 2, 3, and 4, respectively. Section A.2 presents Le Cam's third lemma. Section A.7 presents a figure for the application.

A.1 Proof of Theorem 1

Before presenting the asymptotic properties of $\hat{\tau}_{\text{dsm}}(\theta^*)$, we require technical conditions. For simplicity, let $S_a = S_a(\theta_a^*)$ and let $f_1(S_a)$ and $f_0(S_a)$ be the conditional density of S_a given $A=1$ and $A=0$, respectively.

Assumption 10 *For $a=0,1$, (i) the matching variable S_a has a compact and convex support $\mathscr{S}$, with a continuous density bounded and bounded away from zero: there exist constants C_{1L} and C_{1U} such that $C_{1L} \leq f_1(S_a)/f_0(S_a) \leq C_{1U}$ for any $S_a \in \mathscr{S}$; (ii) $\mu_a(S_a)$ and $\sigma_a^2(S_a)$ satisfy*

Lipschitz continuity conditions: there exists a constant C_2 such that $|\mu_a(S_{a,i})-\mu_a(S_{a,j})| < C_2||S_{a,i}-S_{a,j}||$ for any $S_{a,i}$ and $S_{a,j}$, and similarly for $\sigma_a^2(S_a)$; (iii) there exists $\delta>0$ such that $\mathbb{E}\left\{|Y(a)|^{2+\delta} \mid S_a\right\}$ is uniformly bounded for any $S_a \in \mathscr{S}$; and (iv) $\hat{B}_n = B_n + o_{\mathrm{P}}(1)$.

Assumption 10 has been considered by Abadie and Imbens (2006) and Abadie and Imbens (2016) for matching estimators based on the covariates and the propensity score. Assumption 10 (i) is a convenient regularity condition. Assumption 10 (ii) imposes smoothness conditions for the outcome mean function $\mu_a(S_a)$ and the variance function $\sigma_a^2(S_a)$. Assumption 10 (iii) is a moment condition for establishing the central limit theorem. Assumption 10 (iv) requires regularity conditions on $\mu_a(S_a)$ $(a=0,1)$ and the nonparametric estimators; detailed discussions have appeared in Abadie and Imbens (2011), Abadie and Imbens (2012) and Otsu and Rai (2017).

Under Assumption 10, similar to the proof of Theorem 2 in Abadie and Imbens (2011), $\hat{\tau}_{\mathrm{dsm}}(\theta^*)$ has the following asymptotic linear form:

$$
\begin{aligned}
n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta^*)-\tau\} &= n^{-1/2}\sum_{i=1}^{n}\left\{\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau\right\} && \text{(A.1)}\\
&\quad +n^{-1/2}\sum_{i=1}^{n}A_i\left(1+M^{-1}K_{S_1,i}\right)\left\{Y_i-\mu_1(S_{1,i})\right\} && \text{(A.2)}\\
&\quad -n^{-1/2}\sum_{i=1}^{n}(1-A_i)\left(1+M^{-1}K_{S_0,i}\right)\left\{Y_i-\mu_0(S_{0,i})\right\}+o_{\mathrm{P}}(1). && \text{(A.3)}
\end{aligned}
$$

If any model of the propensity score or prognostic score is correctly specified, by Lemma 2, we have $\mathbb{E}\{\mu_1(S_{1,i})-\mu_0(S_{0,i})\}=\tau$ and therefore the asymptotic bias of $n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta^*)-\tau\}$ is zero.

Let the three terms in (A.1) be

$$
\begin{aligned}
T_{1n} &= n^{-1/2}\sum_{i=1}^{n}\left\{\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau\right\},\\
T_{2n} &= n^{-1/2}\sum_{i=1}^{n}A_i\left(1+M^{-1}K_{S_1,i}\right)\left\{Y_i-\mu_1(S_{1,i})\right\},\\
T_{3n} &= -n^{-1/2}\sum_{i=1}^{n}(1-A_i)\left(1+M^{-1}K_{S_0,i}\right)\left\{Y_i-\mu_0(S_{0,i})\right\}.
\end{aligned}
$$

We show the covariances of the three terms are zero:

$$\begin{aligned}
\operatorname{cov}(T_{1n}, T_{2n}) &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}\left[\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau, A_j\left(1+M^{-1}K_{S_{1,j}}\right)\left\{Y_j-\mu_1(S_{1,j})\right\}\right] \\
&= n^{-1}\sum_{i=1}^{n}\operatorname{cov}\left[\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau, A_i\left(1+M^{-1}K_{S_{1,i}}\right)\left\{Y_i-\mu_1(S_{1,i})\right\}\right] \\
&= n^{-1}\sum_{i=1}^{n}\operatorname{cov}\Big(\mathbb{E}\left\{\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau \mid S_{1,i}, S_{0,i}\right\}, \\
&\quad \mathbb{E}\left[A_i\left(1+M^{-1}K_{S_{1,i}}\right)\left\{Y_i-\mu_1(S_{1,i})\right\} \mid S_{1,i}, S_{0,i}\right]\Big) \\
&\quad +n^{-1}\sum_{i=1}^{n}\mathbb{E}\Big(\operatorname{cov}\left\{\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau \mid S_{1,i}, S_{0,i},\right. \\
&\quad \left. A_i\left(1+M^{-1}K_{S_{1,i}}\right)\left\{Y_i-\mu_1(S_{1,i})\right\} \mid S_{1,i}, S_{0,i}\right\}\Big) \\
&= 0,
\end{aligned}$$

similarly, $\operatorname{cov}(T_{1n}, T_{3n}) = 0$, and by construction, $\operatorname{cov}(T_{2n}, T_{3n}) = 0$. Thus, the asymptotic variance of $n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta^*)-\tau\}$ is

$$\begin{aligned}
\mathbb{V}\left[n^{-1/2}\sum_{i=1}^{n}\left\{\mu_1(S_{1,i})-\mu_0(S_{0,i})-\tau\right\}\right] &+ \mathbb{V}\left[n^{-1/2}\sum_{i=1}^{n}A_i\left(1+M^{-1}K_{S_{1,i}}\right)\left\{Y_i-\mu_1(S_{1,i})\right\}\right] \\
&+ \mathbb{V}\left[n^{-1/2}\sum_{i=1}^{n}(1-A_i)\left(1+M^{-1}K_{S_{0,i}}\right)\left\{Y_i-\mu_0(S_{0,i})\right\}\right].
\end{aligned}$$

The first term becomes $\mathbb{E}\left[\{\mu_1(S_1)-\mu_0(S_0)-\tau\}^2\right]$. Following Abadie and Imbens (2006), the second and third term, as $n \to \infty$, becomes

$$\begin{aligned}
&\operatorname{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}A_i\left(1+M^{-1}K_{S_{1,i}}\right)^2\mathbb{V}(Y_i \mid S_{1,i})\right] \\
&+\operatorname{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}(1-A_i)\left(1+M^{-1}K_{S_{0,i}}\right)^2\mathbb{V}(Y_i \mid S_{0,i})\right] \\
= \;&\mathbb{E}\left(\sigma_1^2(S_1)\left[\frac{1}{e(S_1)}+\frac{1}{2M}\left\{\frac{1}{e(S_1)}-e(S_1)\right\}\right]\right) \\
&+\mathbb{E}\left(\sigma_0^2(S_0)\left[\frac{1}{1-e(S_0)}+\frac{1}{2M}\left\{\frac{1}{1-e(S_0)}-1+e(S_0)\right\}\right]\right).
\end{aligned}$$

This completes the proof of Theorem 1.

A.2 Le Cam's third Lemma

Consider two sequences of probability measures $(\mathbb{Q}^{(n)})_{n=1}^{\infty}$ and $(\mathbb{P}^{(n)})_{n=1}^{\infty}$. Assume that under $\mathbb{P}^{(n)}$, a statistic T_n and the likelihood ratios $\mathrm{d}\mathbb{Q}^{(n)}/\mathrm{d}\mathbb{P}^{(n)}$ satisfy

$$\begin{pmatrix} T_n \\ \log(\mathrm{d}\mathbb{Q}^{(n)}/\mathrm{d}\mathbb{P}^{(n)}) \end{pmatrix} \to \mathcal{N}\left\{ \begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} \tau^2 & c \\ c & \sigma^2 \end{pmatrix} \right\}$$

in distribution, as $n \to \infty$. Then, under $\mathbb{Q}^{(n)}$,

$$T_n \to \mathcal{N}(c, \tau^2)$$

in distribution, as $n \to \infty$. See Le Cam and Yang (1990), Bickel et al. (1993), and van der Vaart (2000) for textbook discussions.

A.3 Proof of Theorem 2

We follow the technique in Andreou and Werker (2012) and Abadie and Imbens (2016). In Abadie and Imbens (2016), the PSM estimators rely on the nuisance parameter estimator under a correct specification of the propensity score model. In our setting, the nuisance parameters include both parameters in the propensity score model and the prognostic score model, and require only one of the models to be correctly specified. Without loss of generality, we assume one working model $e(X;\alpha)$ for the propensity score and one working model $\Psi(X;\beta) = \{\Psi_0(X;\beta_0), \Psi_1(X;\beta_1)\}$ for the prognostic score. The proof for the case with more than two working models for each score is similar at the expense of heavier notation. Let $\mathbb{P}$ be the distribution of $\{(A_i, X_i, Y_i) : i = 1, \ldots, n\}$. Consider $\mathbb{P} = \mathbb{P}^{\theta^*}$ to be indexed by $\theta^* = (\alpha^{*\mathrm{T}}, \beta_0^{*\mathrm{T}}, \beta_1^{*\mathrm{T}})^{\mathrm{T}}$, which satisfies

$$\mathbb{E}\{U(A, X, Y; \theta^*)\} = \mathbb{E}\left\{ \begin{pmatrix} U_1(A, X; \alpha^*) \\ U_2(A, X, Y; \beta_0^*) \\ U_3(A, X, Y; \beta_1^*) \end{pmatrix} \right\} = 0. \tag{A.4}$$

We invoke standard regularity conditions on Z-estimation (van der Vaart 2000) as follows.

Assumption 11 *(i) Under* $\mathbb{P}^{\theta^*}$, $\mathscr{U}_n(\theta^*) \to \mathcal{N}(0, \Sigma_U)$ *in distribution, as* $n \to \infty$, *where* $\Sigma_U = \mathbb{E}\{U(A, X, Y; \theta^*)\, U(A, X, Y; \theta^*)^{\mathrm{T}}\}$; *(ii)* $\Gamma_\theta = \mathbb{E}\{\partial U(A, X, Y; \theta)/\partial \theta^{\mathrm{T}}\}$ *is nonsingular around* θ^*; *and (iii) for any vector of constant* h, $\exp\{n^{1/2} h^{\mathrm{T}} \Gamma_{\theta^*} \Sigma_U^{-1} \mathscr{U}_n(\theta^*)\}$ *is uniformly integrable.*

Under Assumption 11,

$$n^{1/2}(\hat{\theta}-\theta^*) = -\Gamma_{\theta^*}^{-1}\mathscr{U}_n(\theta^*) + o_{\mathrm{P}}(1) \to \mathscr{N}(0,\Sigma_{\theta^*}), \tag{A.5}$$

in distribution, as $n\to\infty$, where $\Sigma_{\theta^*} = \Gamma_{\theta^*}^{-1}\Sigma_U(\Gamma_{\theta^*}^{-1})^{\mathrm{T}}$.

To derive the large sample distribution of $\hat{\tau}_{\mathrm{dsm}}(\hat{\theta})$, following Abadie and Imbens (2016), we impose the following regularity conditions.

Assumption 12 *There exists a neighborhood of θ^*, such that for any θ in this region, the following conditions hold: for $a=0,1$, (i) the matching variable $S_a(\theta)$ has a compact and convex support $\mathscr{S}$, with a continuous density bounded and bounded away from zero; (ii) $\mu_a\{S_a(\theta)\}$ and $\sigma_a^2\{S_a(\theta)\}$ satisfy the Lipschitz continuity condition; and (iii) there exists $\delta>0$ such that $\mathbb{E}\left\{|Y(a)|^{2+\delta} \mid S_a(\theta)\right\}$ is uniformly bounded for any $S_a(\theta)\in\mathscr{S}$.*

Following Andreou and Werker (2012), because we consider a semiparametric model for θ^*, to invoke the Le Cam's lemma, we specify an auxiliary parametric model $\mathbb{P}^{\theta_n}$ defined locally though θ^*, $\theta_n = \theta^* + n^{-1/2}h$, with a density

$$\frac{\exp\left\{n^{1/2}(\theta_n-\theta^*)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathscr{U}_n(\theta^*) - 2^{-1}n(\theta_n-\theta^*)^{\mathrm{T}}\Sigma_{\theta^*}^{-1}(\theta_n-\theta^*)\right\}}{\mathbb{E}\left[\exp\left\{n^{1/2}(\theta_n-\theta^*)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathscr{U}_n(\theta^*) - 2^{-1}n(\theta_n-\theta^*)^{\mathrm{T}}\Sigma_{\theta^*}^{-1}(\theta_n-\theta^*)\right\}\right]}. \tag{A.6}$$

By Assumption 11 (iii), $\exp\{n^{1/2}(\theta_n-\theta^*)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathscr{U}_n(\theta^*)\}$ is uniformly integrable, and thus model (A.6) is uniformly locally asymptotically normal. Because under $\mathbb{P}^{\theta^*}$, $\mathscr{U}_n(\theta^*)\to\mathscr{N}(0,\Sigma_U)$ in distribution, the normalizing constant in the denominator converges to one as $n\to\infty$. The Fisher information under the parametric model (A.6) is $n\Sigma_{\theta^*}^{-1}$. Therefore, $\hat{\theta}$ is efficient under model (A.6).

Now consider (A_i, X_i, Y_i), for $i=1,\ldots,n$, with the local shift $\mathbb{P}^{\theta_n}$ (Bickel et al. 1993). Under model (A.6), the likelihood ratio under $\mathbb{P}^{\theta_n}$ is

$$\begin{aligned}\log(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n}) &= -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathscr{U}_n(\theta^*) + \frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h + o_{\mathrm{P}}(1)\\ &= -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathscr{U}_n(\theta_n) - \frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h + o_{\mathrm{P}}(1),\end{aligned} \tag{A.7}$$

where the second equality follows by the Taylor expansion of $\mathscr{U}_n(\theta^*)$ at θ_n. Moreover, under $\mathbb{P}^{\theta_n}$: $\mathscr{U}_n(\theta_n)\to\mathscr{N}(0,\Sigma_U)$ in distribution, as $n\to\infty$, and

$$n^{1/2}(\hat{\theta}-\theta_n) = \Gamma_{\theta^*}^{-1}\mathscr{U}_n(\theta_n) + o_{\mathrm{P}}(1). \tag{A.8}$$

We also assume the following regularity condition.

Assumption 13 *For all bounded continuous functions $h(A,X,Y)$, the conditional expectation $\mathbb{E}_{\theta_n}\{h(A,X,Y)\}$ converges in distribution to $\mathbb{E}\{h(A,X,Y)\}$, where $\mathbb{E}_{\theta_n}(\cdot)$ is the expectation taken with respect to P^{θ_n}.*

We derive the results in Theorem 2 in two steps.

In the first step, under $\mathbb{P}^{\theta_n}$, we write $\tau=\tau(\theta_n)$ to reflect its dependence on θ_n; to be specific, we have

$$\tau(\theta_n)=\mathbb{E}\big[\mu_1\{S_1(\theta_n)\}-\mu_0\{S_0(\theta_n)\}\big].$$

We derive that under $\mathbb{P}^{\theta_n}$,

$$\begin{pmatrix} n^{1/2}\{\hat\tau_{\mathrm{dsm}}(\theta_n)-\tau(\theta_n)\} \\ n^{1/2}(\hat\theta-\theta_n) \\ \log(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n}) \end{pmatrix} \to \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} & -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} & -h \\ -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\gamma_1 & -h^{\mathrm{T}} & h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h \end{pmatrix} \right\} \tag{A.9}$$

in distribution, as $n\to\infty$. We then express $\tau(\theta_n)=\tau(\theta^*)+\gamma_2^{\mathrm{T}}(n^{-1/2}h)+o(n^{-1/2})$, where

$$\gamma_2=\left.\frac{\partial\tau(\theta)}{\partial\theta}\right|_{\theta=\theta^*}=\mathbb{E}\left[\left.\frac{\partial\mu_1\{S_1(\theta)\}-\mu_0\{S_0(\theta)\}}{\partial\theta}\right|_{\theta=\theta^*}\right]. \tag{A.10}$$

By Le Cam's third lemma, under $\mathbb{P}^{\theta^*}$,

$$\begin{pmatrix} n^{1/2}\{\hat\tau_{\mathrm{dsm}}(\theta_n)-\tau\} \\ n^{1/2}(\hat\theta-\theta_n) \end{pmatrix} \to \mathcal{N}\left\{ \begin{pmatrix} -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h-\gamma_2^{\mathrm{T}}h \\ -h \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} \end{pmatrix} \right\}$$

in distribution, as $n\to\infty$. Replacing θ_n by $\theta^*+n^{-1/2}h$ yields that under $\mathbb{P}^{\theta^*}$,

$$\begin{pmatrix} n^{1/2}\{\hat\tau_{\mathrm{dsm}}(\theta^*+n^{-1/2}h)-\tau\} \\ n^{1/2}(\hat\theta-\theta^*) \end{pmatrix} \to \mathcal{N}\left\{ \begin{pmatrix} -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h-\gamma_2^{\mathrm{T}}h \\ 0 \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} \end{pmatrix} \right\} \tag{A.11}$$

in distribution, as $n\to\infty$.

In the second step, we provide a heuristic derivation for (A.11) to obtain the approximate distribution (2.8). If the Normal distribution were exact, then

$$n^{1/2}\{\hat\tau_{\mathrm{dsm}}(\theta^*+n^{-1/2}h)-\tau\}\mid n^{1/2}(\hat\theta-\theta^*)=h\sim\mathcal{N}\big(-\gamma_2^{\mathrm{T}}h,V_\tau-\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1\big). \tag{A.12}$$

Given that $n^{1/2}(\hat\theta-\theta^*)=h$, we have $\theta^*+n^{-1/2}h=\hat\theta$, and hence $\hat\tau_{\mathrm{dsm}}(\theta^*+n^{-1/2}h)=\hat\tau_{\mathrm{dsm}}(\hat\theta)$.

Marginalizing (A.12) over the asymptotic distribution of $n^{1/2}(\hat{\theta}-\theta^*)$, we derive (2.8). The formal technique to derive (2.8) can be find in Andreou and Werker (2012) and Abadie and Imbens (2016). To avoid repetition, we omit this step.

In the following, we provide the proof to (A.9) in the first step of the proof. Asymptotic normality of $n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta_n)-\tau(\theta_n)\}$ under $\mathbb{P}^{\theta_n}$ follows from Theorem 1 and the uniform local asymptotic normality of model (A.6). Asymptotic joint normality of $\log(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n})$ and $n^{1/2}(\hat{\theta}-\theta_n)$ follows from (A.7) and (A.8). Also, $n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta_n)-\tau(\theta_n)\}=D_n(\theta_n)+o_{\text{P}}(1)$, where

$$\begin{aligned} D_n(\theta_n) &= n^{-1/2}\sum_{i=1}^{n}\big[\mu_1\{S_{1,i}(\theta_n)\}-\mu_0\{S_{0,i}(\theta_n)\}-\tau(\theta_n)\big] \\ &\quad +n^{-1/2}\sum_{i=1}^{n}A_i\left\{1+M^{-1}K_{S_1(\theta_n),i}\right\}\big[Y_i-\mu_1\{S_{1,i}(\theta_n)\}\big] \\ &\quad -n^{-1/2}\sum_{i=1}^{n}(1-A_i)\left\{1+M^{-1}K_{S_0(\theta_n),i}\right\}\big[Y_i-\mu_0\{S_{0,i}(\theta_n)\}\big]+o_{\text{P}}(1). \end{aligned}$$

Therefore, the remaining is to show that, under $\mathbb{P}^{\theta_n}$:

$$\begin{pmatrix} D_n(\theta_n) \\ \mathscr{U}_n(\theta_n) \end{pmatrix} \to \mathscr{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix},\begin{pmatrix} V_\tau & \gamma_1^{\text{T}} \\ \gamma_1 & \Sigma_U \end{pmatrix}\right\} \tag{A.13}$$

in distribution, as $n\to\infty$. To prove (A.13), consider the linear combination

$$\begin{aligned} T_n &= c_0 D_n(\theta_n)+c^{\text{T}}\mathscr{U}_n(\theta_n) \\ &= c_0 n^{-1/2}\sum_{i=1}^{n}\big[\mu_1\{S_{1,i}(\theta_n)\}-\mu_0\{S_{0,i}(\theta_n)\}-\tau(\theta_n)\big] \\ &\quad +c_0 n^{-1/2}\sum_{i=1}^{n}(2A_i-1)\left\{1+M^{-1}K_{S_{A_i}(\theta_n),i}\right\}\big[Y_i-\mu_{A_i}\{S_{A_i,i}(\theta_n)\}\big] \\ &\quad +c_1^{\text{T}} n^{-1/2}\sum_{i=1}^{n}\left[\frac{\partial e(X_i;\alpha_n)}{\partial\alpha}\frac{A_i-e(X_i;\alpha_n)}{e(X_i;\alpha_n)\{1-e(X_i;\alpha_n)\}}\right] \\ &\quad +c_2^{\text{T}} n^{-1/2}\sum_{i=1}^{n}\left[(1-A_i)\frac{\partial\mu_0(X_i;\beta_{0,n})}{\partial\beta_0}\{Y_i-\mu_0(X_i;\beta_{0,n})\}\right] \\ &\quad +c_3^{\text{T}} n^{-1/2}\sum_{i=1}^{n}\left[A_i\frac{\partial\mu_1(X_i;\beta_{1,n})}{\partial\beta_1}\{Y_i-\mu_1(X_i;\beta_{1,n})\}\right]+o_{\text{P}}(1), \end{aligned}$$

where $c=(c_1^{\text{T}},c_2^{\text{T}},c_3^{\text{T}})^{\text{T}}$. We analyze T_n using the martingale theory. We rewrite $T_n=\sum_{k=1}^{2n}\xi_{n,k}$,

where

$$\xi_{n,k} = \begin{cases} \sum_{j=1}^{8} \xi_{n,k}^{(j)}, & 1 \le k \le n, \\ \sum_{j=9}^{11} \xi_{n,k}^{(j)}, & n+1 \le k \le 2n, \end{cases}$$

$$\begin{aligned}
\xi_{n,k}^{(1)} &= c_0 n^{-1/2}\big[\mu_1\{S_{1,k}(\theta_n)\} - \mu_0\{S_{0,k}(\theta_n)\} - \tau(\theta_n)\big], \\
\xi_{n,k}^{(2)} &= c_0 n^{-1/2}(2A_k - 1)\Big\{1 + M^{-1} K_{S_{A_k}(\theta_n),k}\Big\}\big[\mu_{A_k}(X_k) - \mu_{A_k}\{S_{A_k,k}(\theta_n)\}\big] \\
\xi_{n,k}^{(3)} &= c_1^{\mathrm{T}} n^{-1/2}\left[\frac{\partial e(X_k;\alpha_n)}{\partial \alpha}\frac{e(X_k) - e(X_k;\alpha_n)}{e(X_k;\alpha_n)\{1 - e(X_k;\alpha_n)\}}\right], \\
\xi_{n,k}^{(4)} &= c_2^{\mathrm{T}} n^{-1/2}\{1 - e(X_k)\}\frac{\partial \mu_0(X_k;\beta_{0,n})}{\partial \beta_0}\{\mu_0(X_k) - \mu_0(X_k;\beta_{0,n})\}, \\
\xi_{n,k}^{(5)} &= c_3^{\mathrm{T}} n^{-1/2} e(X_k)\frac{\partial \mu_1(X_k;\beta_{1,n})}{\partial \beta_1}\{\mu_1(X_k) - \mu_1(X_k;\beta_{1,n})\}, \\
\xi_{n,k}^{(6)} &= c_1^{\mathrm{T}} n^{-1/2}\left[\frac{\partial e(X_k;\alpha_n)}{\partial \alpha}\frac{A_k - e(X_k)}{e(X_k;\alpha_n)\{1 - e(X_k;\alpha_n)\}}\right], \\
\xi_{n,k}^{(7)} &= -c_2^{\mathrm{T}} n^{-1/2}\left[\{A_k - e(X_k)\}\frac{\partial \mu_0(X_k;\beta_{0,n})}{\partial \beta_0}\{\mu_0(X_k) - \mu_0(X_k;\beta_{0,n})\}\right], \\
\xi_{n,k}^{(8)} &= c_3^{\mathrm{T}} n^{-1/2}\left[\{A_k - e(X_k)\}\frac{\partial \mu_1(X_k;\beta_{1,n})}{\partial \beta_1}\{\mu_1(X_k) - \mu_1(X_k;\beta_{1,n})\}\right], \\
\xi_{n,k}^{(9)} &= c_0 n^{-1/2}(2A_{k-n} - 1)\Big\{1 + M^{-1} K_{S_{A_{k-n}}(\theta_n),k-n}\Big\}\big\{Y_{k-n} - \mu_{A_{k-n}}(X_{k-n})\big\} \\
\xi_{n,k}^{(10)} &= c_2^{\mathrm{T}} n^{-1/2}(1 - A_{k-n})\frac{\partial \mu_0(X_{k-n};\beta_{0,n})}{\partial \beta_0}\{Y_{k-n} - \mu_0(X_{k-n})\}, \\
\xi_{n,k}^{(11)} &= c_3^{\mathrm{T}} n^{-1/2} A_{k-n}\frac{\partial \mu_1(X_{k-n};\beta_{1,n})}{\partial \beta_1}\{Y_{k-n} - \mu_1(X_{k-n})\}.
\end{aligned}$$

Consider the σ-fields

$$\mathscr{F}_{n,k} = \begin{cases} \sigma(A_1, \ldots, A_k, X_1, \ldots, X_k), & 1 \le k \le n, \\ \sigma(A_1, \ldots, A_n, X_1, \ldots, X_n, Y_{k-1}, \ldots, Y_{k-n}), & 2n+1 \le k \le 3n. \end{cases}$$

Then, we have $\Big\{\sum_{k=1}^{i} \xi_{n,i}, \mathscr{F}_{n,i}, 1 \le i \le 2n\Big\}$ is a martingale for each $n \ge 1$, which follows by the following reasons:

(i) because $S_{a,k}(\theta_n)$ is a double balancing score,

$$\mathbb{E}_{\theta_n}(\xi_{n,k}^{(1)} \mid \mathscr{F}_{n,k-1}) = \mathbb{E}\big(c_0 n^{-1/2}\big[\mu_1\{S_{1,k}(\theta_n)\} - \mu_0\{S_{0,k}(\theta_n)\} - \tau(\theta_n)\big] \mid \mathscr{F}_{n,k-1}\big) = 0;$$

(ii) let $\mathscr{F}^0_{n,k} = \sigma\{A_1, \ldots, A_k, S_1(\theta_n), \ldots, S_k(\theta_n)\}$ for $1 \le k \le n$, then

$$\begin{aligned}
\mathbb{E}_{\theta_n}(\xi^{(2)}_{n,k} \mid \mathscr{F}_{n,k-1}) &= \mathbb{E}_{\theta_n}\{\mathbb{E}_{\theta_n}(\xi^{(2)}_{n,k} \mid \mathscr{F}^0_{n,k-1}) \mid \mathscr{F}_{n,k-1}\} \\
&= c_0 n^{-1/2} \mathbb{E}_{\theta_n}\Big((2A_k - 1)\Big\{1 + M^{-1} K_{S_{A_k}(\theta_n),k}\Big\} \\
&\quad \times \mathbb{E}_{\theta_n}\big[\mu_{A_k}(X_k) - \mu_{A_k}\{S_{A_k,k}(\theta_n)\} \mid \mathscr{F}^0_{n,k-1}\big] \mid \mathscr{F}_{n,k-1}\Big) \\
&= c_0 n^{-1/2} \mathbb{E}_{\theta_n}\Big[(2A_k - 1)\Big\{1 + M^{-1} K_{S_{A_k}(\theta_n),k}\Big\} \times 0 \mid \mathscr{F}_{n,k-1}\Big] \\
&= 0;
\end{aligned}$$

(iii) $\mathbb{E}_{\theta_n}(\xi^{(3)}_{n,k} \mid \mathscr{F}_{n,k-1}) = \mathbb{E}_{\theta_n}(\xi^{(4)}_{n,k} \mid \mathscr{F}_{n,k-1}) = \mathbb{E}_{\theta_n}(\xi^{(5)}_{n,k} \mid \mathscr{F}_{n,k-1}) = 0$ because $\mathbb{E}_{\theta_n}\{U(\theta_n)\} = 0$;

(iv) by the conditioning argument,

$$\mathbb{E}_{\theta_n}(\xi^{(6)}_{n,k} \mid \mathscr{F}_{n,k-1}) = \mathbb{E}_{\theta_n}\left[c_1^{\mathrm{T}} n^{-1/2} \frac{\partial e(X_k;\alpha_n)}{\partial \alpha} \frac{\mathbb{E}\{A_k - e(X_k) \mid \mathscr{F}_{n,k-1}, X_k\}}{e(X_k;\alpha_n)\{1 - e(X_k;\alpha_n)\}} \mid \mathscr{F}_{n,k-1}\right] = 0;$$

(v) $\mathbb{E}_{\theta_n}(\xi^{(7)}_{n,k} \mid \mathscr{F}_{n,k-1}) = 0$ and $\mathbb{E}_{\theta_n}(\xi^{(8)}_{n,k} \mid \mathscr{F}_{n,k-1}) = 0$ due to that fact that $A_k - e(X_k)$ is unbiased conditional on X_k;

(vi) $\mathbb{E}_{\theta_n}(\xi^{(9)}_{n,k} \mid \mathscr{F}_{n,k-1}) = 0$ because $(1 - A_{k-n})\{Y_{k-n} - \mu_0(X_{k-n})\}$ is unbiased given $\mathscr{F}_{n,k-1}$;

(vii) $\mathbb{E}_{\theta_n}(\xi^{(10)}_{n,k} \mid \mathscr{F}_{n,k-1}) = 0$ because $A_{k-n}\{Y_{k-n} - \mu_1(X_{k-n})\}$ is unbiased given $\mathscr{F}_{n,k-1}$.

Therefore, we can apply the martingale central limit theorem (Billingsley 1995) to derive the limiting distribution of T_n. Under Assumption 12, we can verify the conditions for the martingale central limit theorem hold. It follows that under $\mathbb{P}^{\theta_n}$, $T_n \to \mathscr{N}(0, \sigma^2)$ in distribution, as $n \to \infty$, where $\sigma^2 = \operatorname{plim} \sum_{k=1}^{2n} \mathbb{E}_{\theta_n}(\xi^2_{n,k} \mid \mathscr{F}_{n,k-1})$. Under Assumption 13, we thus derive the expression of σ^2 and specify the components in (A.13) with

$$\gamma_1 = (\gamma_{11}^{\mathrm{T}}, \gamma_{12}^{\mathrm{T}}, \gamma_{13}^{\mathrm{T}})^{\mathrm{T}}, \tag{A.14}$$

$$\begin{aligned}
\gamma_{11} &= \mathbb{E}\left(\big[\mu_1\{S_1(\theta^*)\} - \mu_0\{S_0(\theta^*)\} - \tau\big] \frac{\partial e(X;\alpha^*)}{\partial \alpha} \frac{A - e(X;\alpha^*)}{e(X;\alpha^*)\{1 - e(X;\alpha^*)\}}\right) \\
&\quad + \mathbb{E}\left(\big[\mu_1(X) - \mu_1\{S_1(\theta^*)\}\big] \frac{\partial e(X;\alpha^*)}{\partial \alpha} \frac{1 - e(X;\alpha^*)}{e(X;\alpha^*)\{1 - e(X;\alpha^*)\}}\right) \\
&\quad - \mathbb{E}\left(\big[\mu_0(X) - \mu_0\{S_0(\theta^*)\}\big] \frac{\partial e(X;\alpha^*)}{\partial \alpha} \frac{-e(X;\alpha^*)}{e(X;\alpha^*)\{1 - e(X;\alpha^*)\}}\right),
\end{aligned}$$

$$\begin{aligned}\gamma_{12} &= -\mathbb{E}\left(\left[\mu_1\{S_1(\theta^*)\}-\mu_0\{S_0(\theta^*)\}-\tau\right](1-A)\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\{\mu_0(X)-\mu_0(X;\beta_0^*)\}\right)\\ &\quad -\mathbb{E}\left(\left[\mu_0(X)-\mu_0\{S_0(\theta^*)\}\right]\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\{\mu_0(X)-\mu_0(X;\beta_0^*)\}\right)-\mathbb{E}\left\{\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\sigma_0^2(X)\right\},\end{aligned}$$

and

$$\begin{aligned}\gamma_{13} &= -\mathbb{E}\left(\left[\mu_1\{S_1(\theta^*)\}-\mu_0\{S_0(\theta^*)\}-\tau\right]A\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\{\mu_1(X)-\mu_1(X;\beta_1^*)\}\right)\\ &\quad -\mathbb{E}\left(\left[\mu_1(X)-\mu_1\{S_1(\theta^*)\}\right]\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\{\mu_1(X)-\mu_1(X;\beta_1^*)\}\right)-\mathbb{E}\left\{\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\sigma_1^2(X)\right\}.\end{aligned}$$

A.4 Proof of Theorem 3

To establish the multiple robustness and asymptotic distributions of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)$ and $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$, we require further technical conditions.

Assumption 14 *For $a=0,1$, the following conditions hold for $q_{a,\xi}$, $F_a(q)$ and $\hat{F}^{(0)}_{a,\mathrm{dsm}}(q)$: (i) $q_{a,\xi}$ lies in a closed interval $\mathscr{I}$; (ii) the estimating equation $F_a(q)=\xi$ has a unique root in the interior of $\mathscr{I}$; (iii) $F_a(q)$ is strictly increasing and continuous with finite first derivative in $\mathscr{I}$, and the derivative $f_a(q)=\mathrm{d}F_a(q)/\mathrm{d}q$ is continuous and bounded away from 0 for all q in $\mathscr{I}$; (iv) $F_a(q;S_a)$ satisfies a Lipschitz continuity condition: there exists a constant C_3 such that $|F_a(q;S_{a,i})-F_a(q;S_{a,j})|<C_3||S_{a,i}-S_{a,j}||$ for any $S_{a,i}$ and $S_{a,j}$; (v) there exists C_4 such that $\mathrm{var}\{\hat{F}_{a,\mathrm{dsm}}(q+\delta)-\hat{F}_{a,\mathrm{dsm}}(q)\}\le C_4 n^{-1}|\delta|$ for any $q, q+\delta$ in $\mathscr{I}$; and (vi) $\hat{B}_{a,n}(q)=B_{a,n}(q)+o_{\mathrm{P}}(1)$ for q in $\mathscr{I}$.*

Assumption 14 extends from the regularity conditions in Assumption 10 to the distribution function and the conditional distribution function, together with some typical regularity conditions for quantile estimation (see e. g., Ghosh 1971; Francisco and Fuller 1991, for detialed discussion).

Now, we give proof for Theorem 3. We obtain the following decomposition

$$\begin{aligned}
n^{1/2}\left\{\hat{F}^{(0)}_{a,\mathrm{dsm}}(q)-F_a(q)\right\} &= n^{-1/2}\sum_{i=1}^{n}\left\{\mathbf{1}(A_i=a)\left(1+M^{-1}K_{S_a,i}\right)\mathbf{1}(Y_i\le q)-F_a(q)\right\}\\
&= n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(A_i=a)\left(1+M^{-1}K_{S_a,i}\right)\left\{\mathbf{1}(Y_i\le q)-F_a(q;S_{a,i})\right\}\\
&\quad +n^{-1/2}\sum_{i=1}^{n}\left\{\mathbf{1}(A_i=a)\left(1+M^{-1}K_{S_a,i}\right)F_a(q;S_{a,i})-F_a(q)\right\}\\
&= n^{-1/2}\sum_{i=1}^{n}\left\{F_a(q;S_{a,i})-F_a(q)\right\}\\
&\quad +n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(A_i=a)\left(1+M^{-1}K_{S_a,i}\right)\left\{\mathbf{1}(Y_i\le q)-F_a(q;S_{a,i})\right\}\\
&\quad -n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(A_i=1-a)M^{-1}\sum_{j\in\mathscr{J}_{S_a,i}}\left\{F_a(q;S_{a,i})-F_a(q;S_{a,j})\right\}
\end{aligned}$$

and denote

$$\begin{aligned}
C_{a,n}(q) &= n^{-1/2}\sum_{i=1}^{n}\left\{F_a(q;S_{a,i})-F_a(q)\right\}\\
&\quad +n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(A_i=a)\left(1+M^{-1}K_{S_a,i}\right)\left\{\mathbf{1}(Y_i\le q)-F_a(q;S_{a,i})\right\},\\
B_{a,n}(q) &= -n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(A_i=1-a)M^{-1}\sum_{j\in\mathscr{J}_{S_a,i}}\left\{F_a(q;S_{a,i})-F_a(q;S_{a,j})\right\}. \qquad \text{(A.15)}
\end{aligned}$$

By Lemma 2, $\mathbb{E}\{F_a(q;S_{a,i})\}=F_a(q)$, so $\mathbb{E}\{C_{a,n}(q)\}=0$. The difference $F_a(q;S_{a,i})-F_a(q;S_{a,j})$ in (A.15) accounts for the matching discrepancy, and therefore $B_{a,n}(q)$ contributes to the asymptotic bias of the matching estimator. To correct for the bias due to matching discrepancy, let $\hat{F}_a(q;S)$ be a nonparametric estimator of $F_a(q;S)$, for $a=0,1$. We propose a de-biasing DSM estimator $\hat{F}_{a,\mathrm{dsm}}(q)$ of $F_a(q)$ in (2.10). Thus, under Assumption 1, regularity conditions in Assumptions 10 (i) and 14, $\hat{F}_{a,\mathrm{dsm}}(q)-F_a(q)=\hat{F}^{(0)}_{a,\mathrm{dsm}}(q)-F_a(q)-n^{-1/2}\hat{B}_{a,n}(q)$, is asymptotically Normal with zero mean and finite variance for all $q\in\mathscr{I}$.

To establish the Bahadur representation (2.12), we follow the strategy of Ghosh (1971). For simplicity, introduce the following notation for $a=0,1$:

$$L^{(n)}_{a,\xi}=n^{1/2}\left(\hat{q}_{a,\xi,\mathrm{dsm}}-q_{a,\xi}\right),\ R^{(n)}_{a,\xi}=-n^{1/2}\frac{\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi})-F_a(q_{a,\xi})}{f_a(q_{a,\xi})}.$$

Writing $r_{a,\xi}^{(n)} = n^{-1/2}(L_{a,\xi}^{(n)} - R_{a,\xi}^{(n)})$, the Bahadur representation (2.12) is equivalent to $r_{a,\xi}^{(n)} = o_{\mathrm{P}}(n^{-1/2})$. We now show that $r_{a,\xi}^{(n)} = o_{\mathrm{P}}(n^{-1/2})$ holds uniformly for $\xi \in \widetilde{\mathscr{I}}$ by verifying the conditions of Lemma 1 of Ghosh (1971).

For a constant c and any $\epsilon > 0$, we show $\lim_{n\to\infty} \mathbb{P}(L_{a,\xi}^{(n)} \le c\,, R_{a,\xi}^{(n)} \ge c+\epsilon) = 0$ for $\xi \in \widetilde{\mathscr{I}}$. Note that $L_{a,\xi}^{(n)} \le c$ is equivalent to

$$-n^{1/2}\frac{\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}+c\,n^{-1/2}) - F_a(q_{a,\xi}+c\,n^{-1/2})}{f_a(q_{a,\xi})} \le -n^{1/2}\frac{\hat{F}_{a,\mathrm{dsm}}(\hat{q}_{a,\xi,\mathrm{dsm}}) - F_a(q_{a,\xi}+c\,n^{-1/2})}{f_a(q_{a,\xi})}.$$

Thus, it suffices to show

$$\lim_{n\to\infty} \mathbb{P}\big(S_{a,\xi}^{(n)} \le c_{a,n}(\xi), R_{a,\xi}^{(n)} \ge c+\epsilon\big) = 0 \tag{A.16}$$

uniformly for all $\gamma \in H_1$, where

$$\begin{aligned} S_{a,\xi}^{(n)} &= -n^{1/2}\frac{\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}+c\,n^{-1/2}) - F_a(q_{a,\xi}+c\,n^{-1/2})}{f_a(q_{a,\xi})}, \\ c_{a,\xi}^{(n)} &= -n^{1/2}\frac{\hat{F}_{a,\mathrm{dsm}}(\hat{q}_{a,\xi,\mathrm{dsm}}) - F_a(q_{a,\xi}+c\,n^{-1/2})}{f_a(q_{a,\xi})}. \end{aligned}$$

By Assumption 14 (iii), we have $F_a(q_{a,\xi}+c\,n^{-1/2}) - F_a(q_{a,\xi}) = c\,n^{-1/2} f_a(q_{a,\xi}) + o(n^{-1/2})$ and thus $c_{a,\xi}^{(n)} \xrightarrow{p} c$, as $n \to \infty$, uniformly for all $\gamma \in \widetilde{\mathscr{I}}$. By Assumption 14 (i)–(iii), we have $\hat{F}_{a,\mathrm{dsm}}(\hat{q}_{a,\xi,\mathrm{dsm}}) - F_a(q_{a,\xi}) = o_{\mathrm{P}}(n^{-1/2})$. By Assumption 14 (v), we have

$$\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}+c\,n^{-1/2}) - \hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}) = F_a(q_{a,\xi}+c\,n^{-1/2}) - F_a(q_{a,\xi}) + O_{\mathrm{P}}(n^{-3/4}),$$

uniformly for all $\xi \in \widetilde{\mathscr{I}}$. Then, we have

$$\begin{aligned} S_{a,\xi}^{(n)} - R_{a,\xi}^{(n)} &= n^{1/2}\frac{F_a(q_{a,\xi}+c\,n^{-1/2}) - F_a(q_{a,\xi})}{f_a(q_{a,\xi})} \\ &\quad -n^{1/2}\frac{\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}+c\,n^{-1/2}) - \hat{F}_{a,\mathrm{dsm}}(q_{a,\xi})}{f_a(q_{a,\xi})} = O_{\mathrm{P}}(n^{-1/4}) \xrightarrow{p} 0, \end{aligned}$$

for all $\xi \in \widetilde{\mathscr{I}}$. This establishes (A.16) and thus the first condition of Lemma 1 of Ghosh (1971). The second condition of Lemma 1 follows similarly.

Therefore, by the Bahadur-type representation for $\hat{q}_{a,\xi,\mathrm{dsm}}$, we have

$$
\begin{aligned}
& n^{1/2}\left\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)-\Delta_\xi\right\} \\
&= n^{1/2}\left\{-\frac{\hat{F}_{1,\mathrm{dsm}}(q_{1,\xi})}{f_1(q_{1,\xi})}+\frac{\hat{F}_{0,\mathrm{dsm}}(q_{0,\xi})}{f_0(q_{0,\xi})}-\Delta_\xi\right\}+o_{\mathrm{P}}(1) \\
&= n^{-1/2}\sum_{i=1}^{n}\left\{-\frac{F_1(q_{1,\xi};S_{1,i})}{f_1(q_{1,\xi})}+\frac{F_0(q_{0,\xi};S_{0,i})}{f_0(q_{0,\xi})}-\Delta_\xi\right\} \\
&\quad -n^{-1/2}\sum_{i=1}^{n}(2A_i-1)\left\{1+M^{-1}K_{S_{A_i},i}\right\}\left\{\frac{\mathbf{1}(Y_i\le q_{A_i,\xi})-F_{A_i}(q_{A_i,\xi};S_{A_i,i})}{f_{A_i}(q_{A_i,\xi})}\right\}+o_{\mathrm{P}}(1).
\end{aligned}
$$

Following a similar derivation in the proof for Theorem 1, the asymptotic variance of $n^{1/2}\left\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)-\Delta_\xi\right\}$ is then given by

$$
\begin{aligned}
V_\Delta = {} & \mathbb{E}\left[\left\{\frac{F_1(q_{1,\xi};S_1)}{f_1(q_{1,\xi})}-\frac{F_0(q_{0,\xi};S_0)}{f_0(q_{0,\xi})}\right\}^2\right] \\
& +\{f_1(q_{1,\xi})\}^{-2}\mathbb{E}\left(F_1(q_{1,\xi};S_1)\{1-F_1(q_{1,\xi};S_1)\}\left[\frac{1}{e(S_1)}+\frac{1}{2M}\left\{\frac{1}{e(S_1)}-e(S_1)\right\}\right]\right) \\
& +\{f_0(q_{0,\xi})\}^{-2}\mathbb{E}\left(F_0(q_{0,\xi};S_0)\{1-F_0(q_{0,\xi};S_0)\}\left[\frac{1}{1-e(S_0)}+\frac{1}{2M}\left\{\frac{1}{1-e(S_0)}-1+e(S_0)\right\}\right]\right).
\end{aligned}
\tag{A.17}
$$

A.5 Proof of Theorem 4

The proof of Theorem 4 is similar to that of Theorem 2. It suffices to replace $n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta_n)-\tau\}$ in (A.11) by $n^{1/2}\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta_n)-\Delta_\xi\}$ and derive the corresponding γ_1 and γ_2, denoted by γ_3 and γ_4. Toward that end, the key is to write

$$
\hat{q}_{1,\xi,\mathrm{dsm}}-\hat{q}_{0,\xi,\mathrm{dsm}}-\Delta_\xi(\theta^*) \;=\; -\frac{\hat{F}_{1,\mathrm{dsm}}(q_{1,\xi})-F_1(q_{1,\xi})}{f_1(q_{1,\xi})}+\frac{\hat{F}_{0,\mathrm{dsm}}(q_{0,\xi})-F_0(q_{0,\xi})}{f_0(q_{0,\xi})}+o_{\mathrm{P}}(n^{-1/2}),
$$

where

$$
\begin{aligned}
\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi}) \;=\; & n^{-1}\sum_{i=1}^{n}\left[F_a\{q_{a,\xi};S_a(\theta^*)\}-\xi\right] \\
& +n^{-1}\sum_{i=1}^{n}\mathbf{1}(A_i=a)\left(1+M^{-1}K_{S_a,i}\right)\left[\mathbf{1}(Y_i\le q_{a,\xi})-F_a\{q_{a,\xi};S_a(\theta^*)\}\right],
\end{aligned}
$$

$$\begin{aligned}\Delta_\xi(\theta^*) &= \mathbb{E}[\frac{A}{e(X;\alpha^*)}\frac{\mathbf{1}(Y\le q_{1,\xi})-F_1\{q_{1,\xi};S_1(\theta^*)\}}{f_1(q_{1,\xi})}-\frac{1-A}{1-e(X;\alpha)}\\ &\quad \frac{\mathbf{1}(Y\le q_{0,\xi})-F_0\{q_{0,\xi};S_0(\theta^*)\}}{f_0(q_{0,\xi})}]+\mathbb{E}\left[\frac{F_1\{q_{1,\xi};S_1(\theta^*)\}}{f_1(q_{1,\xi})}-\frac{F_0\{q_{0,\xi};S_0(\theta^*)\}}{f_0(q_{0,\xi})}\right],\end{aligned}$$

and repeat a similar analysis in Section A.3 with the following changes to $\xi^{(1)}_{n,k}$, $\xi^{(2)}_{n,k}$ and $\xi^{(9)}_{n,k}$:

$$\begin{aligned}\xi^{(1)}_{n,k} &= c_0 n^{-1/2}\left[f_1(q_{1,\xi})^{-1}F_1\{q_{1,\xi};S_{1,k}(\theta_n)\}-f_0(q_{0,\xi})^{-1}F_0\{q_{0,\xi};S_{0,k}(\theta_n)\}-\Delta_\xi(\theta_n)\right],\\ \xi^{(2)}_{n,k} &= c_0 n^{-1/2}(2A_k-1)\left\{1+M^{-1}K_{S_{A_k}(\theta_n),k}\right\}f_{A_k}(q_{A_k,\xi})^{-1}\\ &\quad \left[F_{A_k}(q_{A_k,\xi};X_k)-F_{A_k}\{q_{A_k,\xi};S_{A_k,k}(\theta_n)\}\right],\\ \xi^{(9)}_{n,k} &= c_0 n^{-1/2}(2A_{k-n}-1)\left\{1+M^{-1}K_{S_{A_k}(\theta_n),k-n}\right\}f_{A_{k-n}}(q_{A_{k-n},\xi})^{-1}\\ &\quad \times\left\{\mathbf{1}(Y_{k-n}\le q_{A_{k-n},\xi})-F_{A_{k-n}}(q_{A_{k-n},\xi};X_{k-n})\right\}.\end{aligned}$$

Then, we can derive

$$\gamma_3^{\mathrm{T}}=(\gamma_{31}^{\mathrm{T}},\gamma_{32}^{\mathrm{T}},\gamma_{33}^{\mathrm{T}}), \tag{A.18}$$

where

$$\begin{aligned}\gamma_{31} &= \mathbb{E}\left(\left[\frac{F_1\{q_{1,\xi};S_1(\theta^*)\}}{f_1(q_{1,\xi})}-\frac{F_0\{q_{0,\xi};S_0(\theta^*)\}}{f_0(q_{0,\xi})}-\Delta_\xi\right]\frac{\partial e(X;\alpha^*)}{\partial\alpha}\frac{A-e(X;\alpha^*)}{e(X;\alpha^*)\{1-e(X;\alpha^*)\}}\right)\\ &\quad +\mathbb{E}\left(f_1(q_{1,\xi})^{-1}\left[F_1(q_{1,\xi};X)-F_1\{q_{1,\xi};S_1(\theta^*)\}\right]\frac{\partial e(X;\alpha^*)}{\partial\alpha}\frac{1-e(X;\alpha^*)}{e(X;\alpha^*)\{1-e(X;\alpha^*)\}}\right)\\ &\quad -\mathbb{E}\left(f_0(q_{0,\xi})^{-1}\left[F_0(q_{0,\xi};X)-F_0\{q_{0,\xi};S_0(\theta^*)\}\right]\frac{\partial e(X;\alpha^*)}{\partial\alpha}\frac{-e(X;\alpha^*)}{e(X;\alpha^*)\{1-e(X;\alpha^*)\}}\right),\end{aligned}$$

$$\begin{aligned}\gamma_{32} &= -\mathbb{E}\left(\left[\frac{F_1\{q_{1,\xi};S_1(\theta^*)\}}{f_1(q_{1,\xi})}-\frac{F_0\{q_{0,\xi};S_0(\theta^*)\}}{f_0(q_{0,\xi})}-\Delta_\xi\right](1-A)\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\{\mu_0(X)-\mu_0(X;\beta_0^*)\}\right)\\ &\quad -\mathbb{E}\left(f_0(q_{0,\xi})^{-1}\left[F_0(q_{0,\xi};X)-F_0\{q_{0,\xi};S_0(\theta^*)\}\right]\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\{\mu_0(X)-\mu_0(X;\beta_0^*)\}\right)\\ &\quad -\mathbb{E}\left\{\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}f_0(q_{0,\xi})^{-1}\left[I\{Y(0)\le q_{0,\xi}\}-F_0(q_{0,\xi};X)\right]\{Y(0)-\mu_0(X)\}\right\},\end{aligned}$$

$$
\begin{aligned}
\gamma_{33} &= -\mathbb{E}\left(\left[\frac{F_1\{q_{1,\xi};S_1(\theta^*)\}}{f_1(q_{1,\xi})}-\frac{F_0\{q_{0,\xi};S_0(\theta^*)\}}{f_0(q_{0,\xi})}-\Delta_\xi\right]A\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\{\mu_1(X)-\mu_1(X;\beta_1^*)\}\right)\\
&\quad -\mathbb{E}\left(f_1(q_{1,\xi})^{-1}\left[F_1(q_{1,\xi};X)-F_1\{q_{1,\xi};S_1(\theta^*)\}\right]\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\{\mu_1(X)-\mu_1(X;\beta_1^*)\}\right)\\
&\quad -\mathbb{E}\left\{\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}f_1(q_{1,\xi})^{-1}\left[\mathbf{1}\{Y(1)\le q_{1,\xi}\}-F_1(q_{1,\xi};X)\right]\{Y(1)-\mu_1(X)\}\right\},
\end{aligned}
$$

and

$$
\gamma_4=-\mathbb{E}\left[\frac{1}{f_1(q_{1,\xi})}\frac{\partial F_1\{q_{1,\xi};S_1(\theta^*)\}}{\partial\theta}-\frac{1}{f_0(q_{0,\xi})}\frac{\partial F_0\{q_{0,\xi};S_0(\theta^*)\}}{\partial\theta}\right]. \tag{A.19}
$$

A.6 An additional simulation

We conduct an additional simulation with heterogeneous treatment effects and log-normal errors in the models. The covariate distributions and the propensity score models remain the same as in the main text. The potential outcomes are $Y(0)=\beta_0^{\mathrm{T}}Z+\epsilon(0)$ and $Y(1)=Y(0)+Z_1+2Z_3-1.5Z_5-\epsilon(0)+\epsilon(1)$, where $\log\{\epsilon(0)\}\sim\mathcal{N}(0,0.4^2)$ and $\log\{\epsilon(1)\}\sim\mathcal{N}(0,0.2^2)$.

Figure A.1 shows the distributions of the estimation error (i.e.; the estimator minus the true parameter value) based on 1000 repeated sampling. The same conclusion applies to this simulation setting. DSM is multiply robust and achieves the best performance among all types of estimators.

A.7 Figures

Figure A.2 compares the distributions of extreme and non-extreme propensity scores in the simulation.

Figure A.3 shows that the outcome distributions are highly skewed in the data from the job training program.

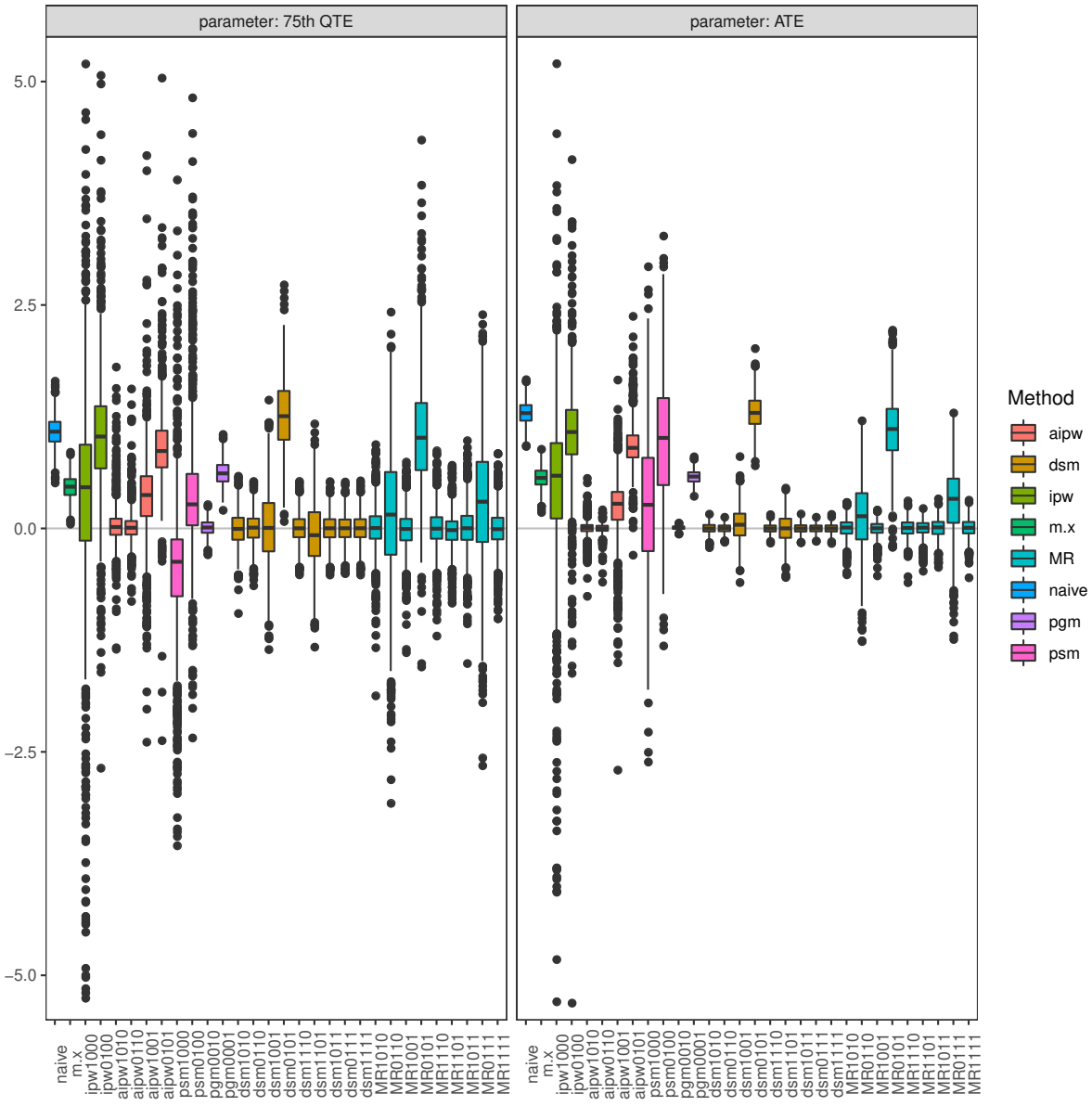


Figure A.1: Simulation results of various weighting and matching estimators when the error distribution is log-normal. There are two panels of results: the left for the 75th QTE and the right for the ATE. Each boxplot shows the distribution of the estimator subtracting the true parameter value based on 1000 Monte Carlo simulated datasets.

Figure A.2: Distribution of extreme and non-extreme propensity scores in the simulation

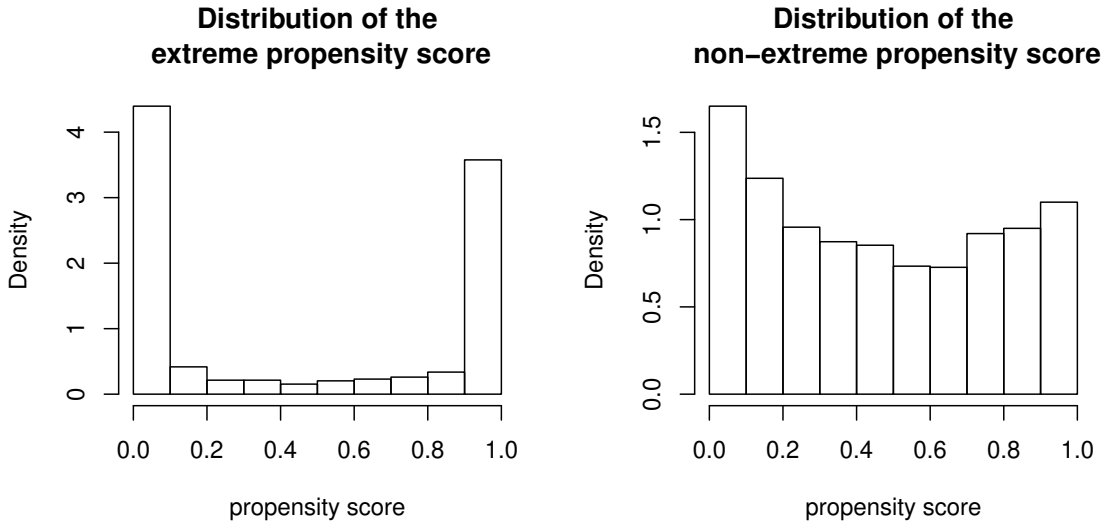
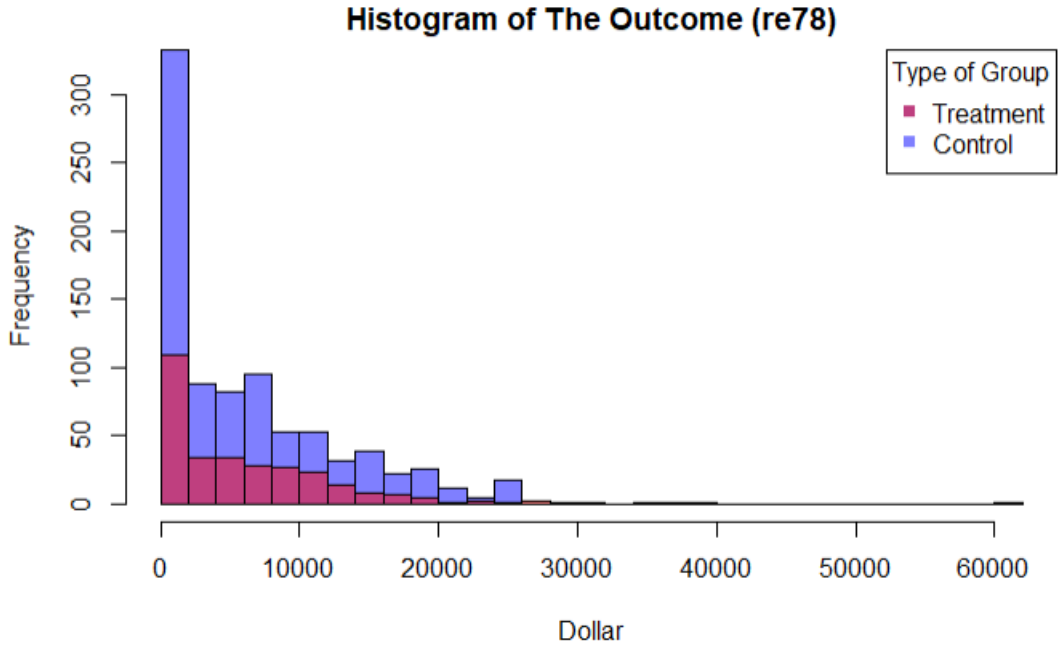


Figure A.3: Histogram of outcome variable (re78)



APPENDIX

B

APPENDIX FOR "PRACTICAL RECOMMENDATIONS ON DOUBLE SCORE MATCHING FOR ESTIMATING CAUSAL EFFECTS"

The appendix contains additional proofs and results for the main paper. Section B.1 derive some asymptotic results for the three matching estimators. Section B.2, B.3, B.4 provide the proof of Theorem 11, Corollary 1, and Corollary 2, respectively. Section B.5 explains key variables in the REFLECTIONS dataset. Section B.6 specifies the detailed configurations for the data generation process in the simulations. Section B.9 shows the complete results for variable selection in DSM, as a complement for Section 3.3.2. Section B.10 presents additional results to illustrate the weakness of DSM when matching without a caliper and without replacement.

B.1 Theory: Asymptotic results for matching estimators on the ATT

In this section, we will derive some asymptotic results for the three matching estimators introduced in Section 3.2. For simplicity, we only consider matching with replacement and without a caliper. Simulations in the following sections will show that this is the optimal setting for matching estimators. Although variable selection strategies such as LASSO are included in the simulations, we do not consider them at this point because they introduce substantial difficulties in theoretical analysis. Nevertheless, the asymptotic results in this section can still provide some insights for choosing the optimal matching estimators.

We first introduce additional notation that is useful in our theoretical analysis. We posit a working model $e(X;\alpha)$ for the propensity score $e(X)$ and a working model $\Psi\left(X;\beta\right)$ for the prognostic score $\Psi(X)$. The double score $S(X;\theta)=\left(e(X;\alpha),\Psi\left(X;\beta\right)\right)$ is the combination of these two scores. We denote $\theta^*=\left(\alpha^*,\beta^*\right)$ as the true parameter, i.e. $S=S(X;\theta^*)=\left(e(X;\alpha^*),\Psi\left(X;\beta^*\right)\right)=(e(X),\Psi(X))$. For simplicity of exposition, for a generic variable V, denote

$$\mu_a(V)=E\{Y(a)|V\},\quad \sigma_a^2(V)=\mathbb{V}\{Y(a)|V\},\quad e(V)=E(A|V),$$

where $\mu_a(V)$ is the mean function for potential outcome $Y(a)$, $\sigma_a^2(V)$ is the variance function, and $e(V)$ is the propensity score given the variable V.

In this paper, we only establish the asymptotic results for matching estimators when θ^* is known, following Abadie and ImbensAbadie and Imbens (2016). Although these can be extended for the estimated parametersYang and Zhang (2020), the resulting expressions are too cumbersome to make comparison with variances under the estimated $\hat{\theta}$. To formalize our analysis, we first posit some necessary assumptions required by the asymptotic properties of matching estimators.

Assumption 1 *There exist constants c_1 and c_2 such that $0<c_1\le e(X)\le c_2<1$ almost surely.*

Assumption 2 $\{Y(0),Y(1)\}\perp\!\!\!\perp A\,|X$.

Assumption 3 *For $a=0,1$, (1) the matching variable V has a compact and convex support $\mathscr{V}$, with a continuous density bounded and bounded away from zero: there exist constants C_{1L} and C_{1U} such that $C_{1L}\le f_1(V)/f_0(V)\le C_{1U}$, for any $V\in\mathscr{V}$; (2) $\mu_a(V)$ and $\sigma_a^2(V)$ satisfy Lipschitz continuity conditions: there exists a constant C_2 such that $\left|\mu_a(V_i)-\mu_a\left(V_j\right)\right|<C_2\left\|V_i-V_j\right\|$ for any V_i and V_j. and similarly for $\sigma_a^2(V)$; and (3) there exists $\delta>0$ such that*

$E\left\{\left|Y(a)^{2+\delta}\right||V\right\}$ *is uniformly bounded for any* $V \in \mathcal{V}$*; (4)* $n^{-1/2}\sum_{i=1}^{n}\left\{A_i-(1-A_i)M^{-1}K_{V,i}\right\}$ $\left\{\hat{\mu}_0(V_i)-\mu_0(V_i)\right\}=o_P(1)$.

Assumption 4 *The matching variable* V *deconfounds the control arm outcome from the treatment assignment:* $Y(0) \perp\!\!\!\perp A\,|V$

Assumption 1 is the standard positivity assumption, and Assumption 2 implies that there is no unmeasured confounder. Assumption 3 as a regularity and smoothness condition is considered by Abadie and ImbensAbadie and Imbens (2016) for the PSM estimator and Yang and ZhangYang and Zhang (2020) for the DSM estimator. The last statement implies that the nonparametric estimate for the bias correction term is consistent. Assumption 4 shows that the matching variable V is a balancing score, which is satisfied by the propensity score $e(X)$, the prognostic score $\Psi(X)$, and the double score $S(X)$. Importantly, only one model is needed to be correctly specified in the double score to deconfound the potential outcome from the treatment assignment, contributing to the double robustness of DSM.

In the following theorem, we establish the asymptotic result for the matching estimator $\hat{\tau}_{ATT}$.

Theorem 11 *Under Assumptions 1-4, the matching estimator on the ATT based on matching variable* V *is asymptotically normal:* $\sqrt{n}(\hat{\tau}_{\mathrm{ATT}}-\tau_{\mathrm{ATT}}) \to \mathcal{N}\left(0, V_{\tau_{\mathrm{ATT}}}\right)$*, in distribution, as* $n \to \infty$*, where*

$$V_{\tau_{\mathrm{ATT}}}=\frac{1}{p^2}E\left[e(V)\left\{\mu_1(V)-\mu_0(V)-\tau_{\mathrm{ATT}}\right\}^2\right]+\frac{1}{p^2}E\left\{e(V)\sigma_1^2(V)\right\}$$
$$+\frac{1}{p^2}E\left[\sigma_0^2(V)\left\{\frac{e^2(V)}{1-e(V)}+\frac{1}{M}e(V)+\frac{1}{2M}\frac{e^2(V)}{1-e(V)}\right\}\right]$$

and $p=E\{e(X)\}=E\{e(V)\}$ *is the proportion of treatment population. Specifically,* V *can be the propensity score* $e(X)$*, the prognostic score* $\Psi(X)$*, and the double score* $S(X)$*. Importantly, the double score* $S(X)$ *only requires one of the two scores to be correctly specified, which implies the double robustness of DSM.*

Here M is the number of controlled subjects matched to each treated subject. In our following simulations, M is always chosen as 1. The proof of this theorem is provided in the supplementary material. By replacing V by $e(X)$, $\Psi(X)$, or $S(X)$, we can derive the asymptotic variances for the PSM, PGM, or DSM estimator, respectively. Therefore, it is natural to compare these three variance terms. Unfortunately, there is no deterministic ordering

among the variances of the three matching estimators for the ATT. Instead, we consider the estimation of the average control arm outcome on the treated $\mu_{0,trt} = E\{Y(0)|A=1\}$. It is worth discussing since $\hat{\tau}_{ATT} = n_1^{-1}\sum_{u=1}^{n} A_i Y_i - \hat{\mu}_{0,trt}$, and the first term is the same for all the matching estimators. The following corollary establishes its asymptotic distribution.

Corollary 1 *Under Assumptions 1-4, the matching estimator on the average control arm outcome on the treated based on matching variable V is asymptotically normal: $\sqrt{n}(\hat{\mu}_{0,trt} - \mu_{0,trt}) \to N\left(0, V_{\mu_{0,trt}}\right)$, where*

$$V_{\mu_{0,trt}} = \frac{1}{p^2}E\left[e(V)\left\{\mu_0(V)-\mu_{0,trt}\right\}^2\right] + \frac{1}{p^2}E\left[\sigma_0^2(V)\cdot\left\{\frac{e^2(V)}{1-e(V)} + \frac{1}{M}e(V) + \frac{1}{2M}\frac{e^2(V)}{1-e(V)}\right\}\right]$$

In the following theorem, we compare DSM with the other two matching estimators under correct model specifications.

Corollary 2 *When models are correctly specified, PGM is always more efficient than DSM for estimation of $\mu_{0,trt} = E\{Y(0)|A=1\}$. DSM is more efficient than PSM if and only if*

$$E\left(\left[\mu_0\{S(X)\}-\mu_0\{e(X)\}\right]^2\left\{\frac{e^2(X)}{1-e(X)} + \frac{1}{M}e(X) + \frac{1}{2M}\frac{e^2(X)}{1-e(X)} - 1\right\}\right) \geq 0$$

The first part of Corollary 2 coincides with Leacy and StaurtLeacy and Stuart (2014) that PGM has smaller variance than DSM when both models are correctly specified. This underscores the importance of incorporating variable selection into DSM to increase its efficiency, as shown in the simulations.

B.2 Proof of Theorem 1

Firstly, it is straightforward that

$$\sqrt{n}\{\hat{\tau}_{ATT}(\theta^*) - \tau_{ATT}\} = \frac{n}{n_1}\cdot\frac{n_1}{\sqrt{n}}\{\hat{\tau}_{ATT}(\theta^*) - \tau_{ATT}\}.$$

Also, because

$$\frac{n}{n_1} \to \frac{1}{p}$$

in probability, we only need to derive the limiting distribution for the second term, which can be expanded into three terms:

$$\frac{n_1}{\sqrt{n}}\{\hat{\tau}_{ATT}(\theta^*)-\tau_{ATT}\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\hat{\mu}_1(V_i)-\hat{\mu}_0(V_i)-\tau_{ATT}\right\}$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{Y_i-\hat{\mu}_1(V_i)\right\}$$
$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{V,i}\left\{Y_i-\hat{\mu}_0(V_i)\right\}.$$

The nonparametric estimators are assumed to be consistent estimates of the true means, thus

$$\frac{n_1}{\sqrt{n}}\{\hat{\tau}_{ATT}(\theta^*)-\tau_{ATT}\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{Y_i-\mu_1(V_i)\right\}$$
$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{V,i}\left\{Y_i-\mu_0(V_i)\right\}+o_P(1).$$

Denote

$$T_{1n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}, \tag{B.1}$$
$$T_{2n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{Y_i-\mu_1(V_i)\right\}, \tag{B.2}$$
$$T_{3n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{V,i}\left\{Y_i-\mu_0(V_i)\right\}. \tag{B.3}$$

We first show these three terms all have zero expectations:

$$E(T_{1n})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left(E\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}|A_i\right]\right)$$
$$=\sqrt{n}pE\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}|A_i=1\right\}$$
$$=\sqrt{n}pE\left[E\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}|A_i=1,V_i\right\}\right]=0.$$

The last line is because

$$\tau_{ATT} = E\{Y(1) - Y(0)|A=1\} = E\left[E\left\{\mu_1(V) - \mu_0(V)|A=1, V\right\}\right].$$

For the second term,

$$\begin{aligned} E(T_{2n}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left(E\left[A_i\left\{Y_i - \mu_1(V_i)\right\}|A_i\right]\right) \\ &= \sqrt{n} p E\left\{Y_i - \mu_1(V_i)|A_i=1\right\} = 0. \end{aligned}$$

The last line is simply from the definition of $\mu_1(V_i)$:

$$E\left\{\mu_1(V_i)|A_i=1\right\} = E\left[E\left\{Y_i|A_i=1, V_i\right\}\right] = E\left\{Y_i|A_i=1\right\}.$$

For the third term,

$$\begin{aligned} E(T_{3n}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1-A_i) M^{-1} K_{V,i}\left\{Y_i - \mu_0(V_i)\right\} \\ &= \sqrt{n}(1-p) E\left[M^{-1} K_{V,i}\left\{Y_i - \mu_0(V_i)\right\}|A_i=0\right] \\ &= \sqrt{n}(1-p) E\left(E\left[M^{-1} K_{V,i}\left\{Y_i - \mu_0(V_i)\right\}|A_i=0, V_i\right]\right) \\ &= \sqrt{n}(1-p) E\left(M^{-1} K_{V,i} E\left[\left\{Y_i - \mu_0(V_i)\right\}|A_i=0, V_i\right]\right) = 0. \end{aligned}$$

As a result, we prove that the asymptotic bias of $n^{1/2}\{\hat{\tau}_{ATT}(\theta^*) - \tau_{ATT}\}$ is zero. We show that the covariances of T_{1n}, T_{2n}, T_{3n} are zero:

$$
\begin{aligned}
\operatorname{cov}(T_{1n}, T_{2n}) =& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}, A_j\left\{Y_j-\mu_1\left(V_j\right)\right\}\right] \\
=& \frac{1}{n}\sum_{i=1}^{n}\operatorname{cov}\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}, A_i\left\{Y_i-\mu_1(V_i)\right\}\right] \\
=& \frac{1}{n}\sum_{i=1}^{n}E\left(\operatorname{cov}\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}, A_i\left\{Y_i-\mu_1(V_i)\right\}|A_i\right]\right) \\
&+\frac{1}{n}\sum_{i=1}^{n}\operatorname{cov}\left(E\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}|A_i\right], E\left[A_i\left\{Y_i-\mu_1(V_i)\right\}|A_i\right]\right) \\
=& pE\left[\operatorname{cov}\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}, Y_i-\mu_1(V_i)|A_i=1\right\}\right]+0 \\
=& pE\left[\operatorname{cov}\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}, Y_i-\mu_1(V_i)|A_i=1, V_i\right\}\right] \\
&+p\operatorname{cov}\left[E\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}|A_i=1, V_i\right\}, E\left\{Y_i-\mu_1(V_i)|A_i=1, V_i\right\}\right] \\
=& 0.
\end{aligned}
$$

The other two covariances are automatically zero by construction. Thus, the asymptotic variance is the summation of the three variance terms:

$$
\begin{aligned}
\mathbb{V}(T_{1n}) &= E\left(\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}\right]^2\right) \\
&= E\left[A_i\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}^2|V_i\right] \\
&= E\left[E(A_i|V_i)\left\{\mu_1(V_i)-\mu_0(V_i)-\tau_{ATT}\right\}^2\right] \\
&= E\left[e(V)\left\{\mu_1(V)-\mu_0(V)-\tau_{ATT}\right\}^2\right],
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{V}(T_{2n}) &= E\left(\left[A_i\left\{Y_i-\mu_1(V_i)\right\}\right]^2\right) \\
&= E\left[A_i\left\{Y_i-\mu_1(V_i)\right\}^2|V_i\right] \\
&= E\left[E(A_i|V_i)\left\{Y_i-\mu_1(V_i)\right\}^2\right] \\
&= E\left[e(V_i)\left\{Y_i-\mu_1(V_i)\right\}^2\right] \\
&= E\left\{e(V)\sigma_1^2(V)\right\}.
\end{aligned}
$$

Following Abdie and ImbensAbadie and Imbens (2016), the third term has the limiting

variance

$$\mathbb{V}(T_{3n}) \to E\left[\sigma_0^2(V)\cdot\left\{\frac{e^2(V)}{1-e(V)}+\frac{1}{M}e(V)+\frac{1}{2M}\frac{e^2(V)}{1-e(V)}\right\}\right].$$

Combining the three terms and apply the Slutsky's Theorem, this finishes the proof of Theorem 1.

B.3 Proof of Corollary 1

The matching estimator for the average control arm outcome for the treated can be written as

$$\begin{aligned}\hat{\mu}_{0,trt} &= n_1^{-1}\sum_{i=1}^{n} A_i\left\{M^{-1}\sum_{j\in J_{V,i}} Y_j + \hat{\mu}_0(V_i) - M^{-1}\sum_{j\in J_{V,i}}\hat{\mu}_0\big(V_j\big)\right\}\\ &= n_1^{-1}\sum_{i=1}^{n} A_i\hat{\mu}_0(V_i) + (1-A_i)M^{-1}K_{V,i}\left\{Y_i-\hat{\mu}_0(V_i)\right\}.\end{aligned}$$

Thus,

$$\sqrt{n}\left\{\hat{\mu}_{0,trt}(\theta^*)-\mu_{0,trt}\right\} = \frac{n}{n_1}\cdot\frac{n_1}{\sqrt{n}}\left\{\hat{\mu}_{0,trt}(\theta^*)-\mu_{0,trt}\right\},$$

and we only need to derive the asymptotic distribution of the second term:

$$\begin{aligned}\frac{n_1}{\sqrt{n}}\left\{\hat{\mu}_{0,trt}(\theta^*)-\mu_{0,trt}\right\} =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_i\left\{\mu_0(V_i)-\mu_{0,trt}\right\}\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{V,i}\left\{Y_i-\mu_0(V_i)\right\}+o_P(1).\end{aligned}$$

Denote

$$\begin{aligned}T'_{1n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_i\left\{\mu_0(V_i)-\mu_{0,trt}\right\},\\ T'_{3n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{V,i}\left\{Y_i-\mu_0(V_i)\right\} = T_{3n}.\end{aligned}$$

Similarly, the covariance of T'_{1n} and T'_{3n} can be shown as zero. The variance of T'_{1n} is

$$\begin{aligned}
\mathbb{V}\left(T'_{1n}\right) &= E\left(\left[A_i\left\{\mu_0(V_i)-\mu_{0,trt}\right\}\right]^2\right) \\
&= E\left[A_i\left\{\mu_0(V_i)-\mu_{0,trt}\right\}^2|V_i\right] \\
&= E\left[E\left(A_i|V_i\right)\left\{\mu_0(V_i)-\mu_{0,trt}\right\}^2\right] \\
&= E\left[e\left(V\right)\left\{\mu_0(V)-\mu_{0,trt}\right\}^2\right].
\end{aligned}$$

The variance of T'_{3n} is same as the variance of T_{3n}:

$$\mathbb{V}\left(T'_{3n}\right) = E\left[\sigma_0^2(V)\cdot\left\{\frac{e^2(V)}{1-e\left(V\right)}+\frac{1}{M}e\left(V\right)+\frac{1}{2M}\frac{e^2(V)}{1-e\left(V\right)}\right\}\right].$$

Combining the two terms and apply the Slutsky's Theorem, we finish the proof of Corollary 1.

B.4 Proof of Corollary 2

By Corollary 1, we obtain the asymptotic variance for the the three matching estimators:

$$\begin{aligned}
V_{\mu_{0,trt},PSM} &= \frac{1}{p^2}E\left(e\left(X\right)\left[\mu_0\left\{e\left(X\right)\right\}-\mu_{0,trt}\right]^2\right) \\
&\quad+\frac{1}{p^2}E\left[\sigma_0^2\left\{e\left(X\right)\right\}\cdot\left\{\frac{e^2(X)}{1-e\left(X\right)}+\frac{1}{M}e\left(X\right)+\frac{1}{2M}\frac{e^2(X)}{1-e\left(X\right)}\right\}\right], \\
V_{\mu_{0,trt},PGM} &= \frac{1}{p^2}E\left[e\left(\Psi\right)\left\{\mu_0(\Psi)-\mu_{0,trt}\right\}^2\right] \\
&\quad+\frac{1}{p^2}E\left[\sigma_0^2(\Psi)\cdot\left\{\frac{e^2(\Psi)}{1-e\left(S\right)}+\frac{1}{M}e\left(\Psi\right)+\frac{1}{2M}\frac{e^2(\Psi)}{1-e\left(\Psi\right)}\right\}\right], \\
V_{\mu_{0,trt},DSM} &= \frac{1}{p^2}E\left[e\left(S\right)\left\{\mu_0(S)-\mu_{0,trt}\right\}^2\right] \\
&\quad+\frac{1}{p^2}E\left[\sigma_0^2(S)\cdot\left\{\frac{e^2(S)}{1-e\left(S\right)}+\frac{1}{M}e\left(S\right)+\frac{1}{2M}\frac{e^2(S)}{1-e\left(S\right)}\right\}\right],
\end{aligned}$$

where for simplicity we denote $\Psi = \Psi(X)$ as the prognostic score and $S = S(X) = (e\left(X\right),\Psi(X))$ as the double score.

By the definition of the prognostic score,

$$\mu_0(S) = E\{Y(0)|S\} = E\{Y(0)|e(X),\Psi(X)\} = E\{Y(0)|X\} = \mu_0(X)$$
$$\mu_0(\Psi) = E\{Y(0)|\Psi(X)\} = E\{Y(0)|X\} = \mu_0(X).$$

Thus,

$$\mu_0(S) = \mu_0(\Psi) = \mu_0(X). \tag{B.4}$$

Similarly,

$$\sigma_0(S) = \sigma_0(\Psi) = \sigma_0(X). \tag{B.5}$$

We first compare the variances of PGM and DSM. By (B.4) and (B.5), we can show that their first terms are equivalent:

$$\begin{aligned} E\left[e(S)\left\{\mu_0(S)-\mu_{0,trt}\right\}^2\right] &= E\left[e(S)\left\{\mu_0(\Psi)-\mu_{0,trt}\right\}^2\right] \\ &= E\left(E\left[e(S)\left\{\mu_0(\Psi)-\mu_{0,trt}\right\}^2|\Psi\right]\right) \\ &= E\left[\left\{\mu_0(\Psi)-\mu_{0,trt}\right\}^2 E\{e(S)|\Psi\}\right] \\ &= E\left[e(\Psi)\left\{\mu_0(\Psi)-\mu_{0,trt}\right\}^2\right]. \end{aligned}$$

The second term in the asymptotic variance of DSM can be written as

$$\begin{aligned} & E\left[\sigma_0^2(S)\left\{\frac{1}{M}e(S)+\left(1+\frac{1}{2M}\right)\frac{e^2(S)}{1-e(S)}\right\}\right] \\ =& E\left(E\left[\sigma_0^2(\Psi)\left\{\frac{1}{M}e(S)+\left(1+\frac{1}{2M}\right)\frac{e^2(S)}{1-e(S)}\right\}\right]|\Psi\right) \\ =& E\left(\sigma_0^2(\Psi)E\left[\left\{\frac{1}{M}e(S)+\left(1+\frac{1}{2M}\right)\frac{e^2(S)}{1-e(S)}\right\}\right]|\Psi\right) \\ =& E\left(\sigma_0^2(\Psi)\left[\frac{1}{M}e(\Psi)+\left(1+\frac{1}{2M}\right)E\left\{\frac{e^2(S)}{1-e(S)}|\Psi\right\}\right]\right). \end{aligned}$$

By Jensen's inequality, we have

$$E\left\{\frac{e^2(S)}{1-e(S)}|\Psi\right\} \geq \frac{e^2(S|\Psi)}{1-e(S|\Psi)} = \frac{e^2(\Psi)}{1-e(\Psi)},$$

where the last equality is simply implied by the construction of the double score:

$$e(S|\Psi) = E\{e(S)|\Psi\} = E[E\{A|S\}|\Psi] = E(A|\Psi) = e(\Psi).$$

As a result, it follows that $V_{\mu_{0,trt},DSM} \geq V_{\mu_{0,trt},PGM}$.

To compare the variances of PSM and DSM, we decompose

$$Y(0) = \mu_0\{e(X)\} + \varepsilon_{0,S|e(X)} + \varepsilon_0, \tag{B.6}$$

where $\varepsilon_{0,S|e(X)}$ and ε_0 have mean zero and satisfy that $\mu_0\{e(X)\}, \varepsilon_{0,S|e(X)}, \varepsilon_0$ are mutually independent. Moreover, $\mu_0(S) = \mu_0\{e(X)\} + \varepsilon_{0,S|e(X)}$. With this decomposition, $\sigma_0^2(S) = E\left(\varepsilon_0^2\right)$ and $\sigma_0^2\{e(X)\} = E\left\{\varepsilon_{0,S|e(X)}^2 | e(X)\right\} + E\left(\varepsilon_0^2\right)$.

We first expand the first term in $V_{\mu_{0,trt},DSM}$:

$$\begin{aligned}
E\left[e(S)\left\{\mu_0(S_i) - \mu_{0,trt}\right\}^2\right] &= E\left[e(S)\left\{\mu_0\{e(X)\} + \varepsilon_{0,S|e(X)} - \mu_{0,trt}\right\}^2\right] && \text{(B.7)}\\
&= E\left(e(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right) && \text{(B.8)}\\
&\quad + 2E\left(e(S)\varepsilon_{0,S|e(X)}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right) && \text{(B.9)}\\
&\quad + E\left\{\varepsilon_{0,S|e(X)}^2\right\}. && \text{(B.10)}
\end{aligned}$$

(B.8) can be shown to be equivalent as the first term in $V_{\mu_{0,trt},PSM}$:

$$\begin{aligned}
E\left(e(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right) &= E\left\{E\left(e(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2 | e(X)\right)\right\}\\
&= E\left(E\{e(S)|e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right)\\
&= E\left(e(X)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right).
\end{aligned}$$

(B.9) is in fact a zero term:

$$\begin{aligned}
E\left(e(S)\varepsilon_{0,S|e(X)}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right) &= E\left(e(S)\left[\mu_0(S) - \mu_0\{e(X)\}\right]\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&= E\left(e(S)\mu_0(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&\quad - E\left(e(S)\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&= E\left\{E\left(e(S)\mu_0(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]|e(X)\right)\right\}\\
&\quad - E\left\{E\left(e(S)\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]|e(X)\right)\right\}\\
&= E\left(E\left[E\{AY(0)|S\}|e(X)\right]\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&\quad - E\left(E\{e(S)|e(X)\}\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&= E\left(E\{AY(0)|e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&\quad - E\left(e(X)\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&= 0.
\end{aligned}$$

Thus, the difference between the first terms of $V_{\mu_{0,trt},DSM}$ and $V_{\mu_{0,trt},PSM}$is simply $E\{\varepsilon^2_{0,S|e(X)}\}$. Then, by the decomposition from (B.6), the second term in $V_{\mu_{0,trt},PSM}$ can be written as:

$$
\begin{aligned}
&E\left[\sigma_0^2\{e(X)\}\cdot\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right]\\
=&E\left(\left[E\left\{\varepsilon^2_{0,S|e(X)}|e(X)\right\}+E\left(\varepsilon_0^2\right)\right]\cdot\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right)\\
=&E\left[\sigma_0^2(S)\cdot\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right]\\
&+E\left[E\left\{\varepsilon^2_{0,S|e(X)}|e(X)\right\}\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right].
\end{aligned}
$$

As a result, the difference between $V_{\mu_{0,trt},DSM}$ and $V_{\mu_{0,trt},PSM}$ is

$$
\begin{aligned}
V_{\mu_{0,trt},psm}-V_{\mu_{0,trt},dsm}&=\frac{1}{p^2}E\left[E\left\{\varepsilon^2_{0,S|e(X)}|e(X)\right\}\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}-1\right\}\right]\\
&=\frac{1}{p^2}E\left[\varepsilon^2_{0,S|e(X)}\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}-1\right\}\right]\\
&=\frac{1}{p^2}E(\left[\mu_0(S)-\mu_0\{e(X)\}\right]^2\\
&\left\{\frac{e^2(X)}{1-e(X)}+\frac{1}{M}e(X)+\frac{1}{2M}\frac{e^2(X)}{1-e(X)}-1\right\}).
\end{aligned}
$$

Unfortunately, this difference can be either positive or negative. Thus, we don't have a deterministic conclusion on the comparison between PSM and DSM. However, simulations show that DSM can be more efficient than PSM in most situations.

B.5 Key variables in the REFLECTIONS dataset

Table B.1 as an excerpt from Faries et al.'s bookFaries et al. (2020) summarizes all the continuous variables from the REFLECTIONS dataset that were used in the simulations. Detailed explanation for each variable can be found in the Peng et al. analysisPeng et al. (2015).

Table B.1: List of the continuous variables in the REFLECTIONS dataset.

Variable Name	Variable Label
Age	Age in years
SymDur	BMI at Baseline
BPIPain_B	BPI Pain score at Baseline
BPIInterf_B	BPI Interference score at Baseline
BMI_B	BMI at Baseline
DxDur	Time (in years) since initial Dx
PHQ8_B	PHQ8 total score
MFIpf_B	MFI Physical Fatigue at Baseline
TrtDur	Time (in years) since initial Trtmnt
PhysicalSymp_B	PHQ 15 total score at Baseline
GAD7_B	GAD7 total score at Baseline
CPFQ_B	CPFQ Total Score at Baseline
FIQ_B	FIQ Total Score at Baseline
SDS_B	SDS total score at Baseline
ISIX_B	ISIX total score at Baseline
CPFQ_B	CPFQ Total score at Baseline

B.6 Detailed configurations for the data generation

B.6.1 Model specification for the illustrating example

The covariates $X_1 - X_{16}$ are independent standard normal variables.

The true propensity score model is

$$\text{logit}\{P(A=1 \mid \mathbf{X})\} = -1 + 1.5X_1 - 0.5X_2 + 1.75X_3 + 0.5X_4 + 1.5X_5 - 1.25X_6 + 1.75X_7.$$

The true outcome models are

$$\begin{aligned} Y(0) &= 1 - 2.5X_1 - 1.5X_2 - X_3 + 3X_4 + 1.5X_8 - 3.5X_9 + 2.5X_{10} + \epsilon \\ Y(1) &= Y(0) + 3, \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0,1)$. This error term is also contained in the following section.

B.6.2 Model specification for the REFLECTIONS simulations

The covariates are selected from the simulated RELFECTIONS datasets, which is generated from the real REFLECTIONS datasets using the Iman-Conover methodFaries et al. (2020). The true propensity score and outcome models vary across different scenarios.

Linear model and constant effect

The true propensity score is

$$\begin{aligned}\text{logit}\{P(A=1\mid\mathbf{X})\} = (&-2+Age-0.5SymDur+1.5BPIPain\_B+0.5BPIInterf\_B\\ &+3BMI\_B-2.5DxDur+3.5PHQ8\_B)/2.\end{aligned}$$

The true outcome models are

$$\begin{aligned}Y(0) =& 1-2Age-SymDur-0.5BPIPain\_B+3BPIInterf\_B\\ &+1.5MFIpf\_B-3.5TrtDur+2.5PhysicalSymp\_B+\epsilon\\ Y(1) =& Y(0)+3,\end{aligned}$$

Nonlinear model and constant effect

The true propensity score is

$$\begin{aligned}\text{logit}\{P(A=1\mid\mathbf{X})\} = (&-6+3Age-0.5SymDur+1.5BPIPain\_B+0.5BPIInterf\_B\\ &+3BMI\_B-2.5DxDur+3.5PHQ8\_B+Age\cdot BPIPain\_B\\ &+1.5SymDur\cdot BPIInterf\_B+0.5BPIPain\_B\cdot BMI\_B\\ &+1.5BPIInterf\_B\cdot DxDur+2.5BMI\_B\cdot PHQ8\_B\\ &+1.5Age\cdot DxDur+2SymDur\cdot BPIPain\_B+0.5BPIPain\_B\\ &\cdot BPIInterf\_B+BPIInterf\_B\cdot BMI\_B+1.5BMI\_B\cdot DxDur\\ &-2SymDur\cdot SymDur+1.5BPIInterf\_B\cdot BPIInterf\_B\\ &+1.5DxDur\cdot DxDur)/8.\end{aligned}$$

The true outcome models are

$$\begin{aligned}
Y(0) = & 1 - 2Age - SymDur - 0.5BPIPain_B + 3BPIInterf_B + 1.5MFIpf_B \\
& - 3.5TrtDur + 2.5PhysicalSymp_B + SymDur \cdot BPIInterf_B \\
& + 2.5BPIPain_B \cdot PHQ8_B + BPIInterf_B \cdot TrtDur \\
& + 1.5MFIpf_B \cdot PhysicalSymp_B + 2Age \cdot TrtDur \\
& + 0.5SymDur \cdot BPIPain_B + BPIPain_B \cdot BPIInterf_B \\
& + 1.5BPIInterf_B \cdot MFIpf_B + 0.5MFIpf_B \cdot TrtDur \\
& - SymDur \cdot SymDur + 2.5BPIInterf_B \cdot BPIInterf_B \\
& + 1.5PhysicalSymp_B \cdot PhysicalSymp_B + \epsilon \\
Y(1) = & Y(0) + 3,
\end{aligned}$$

Linear model and heterogeneous effect

The true propensity score is

$$\begin{aligned}
\text{logit}\{P(A=1 \mid \mathbf{X})\} = & (-3 + 6Age - 0.5SymDur + 1.5BPIPain_B + 0.5BPIInterf_B \\
& + 5BMI_B - 2.5DxDur + 3.5PHQ8_B)/2.
\end{aligned}$$

The true outcome models are

$$\begin{aligned}
Y(0) = & 1 - 2Age - SymDur - 0.5BPIPain_B + 3BPIInterf_B \\
& + 1.5MFIpf_B - 3.5TrtDur + 2.5PhysicalSymp_B + \epsilon \\
Y(1) = & Y(0) + 3 + 5Age + 4BMI_B - 3MFIpf_B - 2.5GAD7_B.
\end{aligned}$$

Nonlinear model and heterogeneous effect

The true propensity score is

$$\begin{aligned}\text{logit}\{P(A=1\mid \mathbf{X})\} = (&-6+3Age-0.5SymDur+1.5BPIPain\_B+0.5BPIInterf\_B\\&+3BMI\_B-2.5DxDur+3.5PHQ8\_B+Age\cdot BPIPain\_B\\&+1.5SymDur\cdot BPIInterf\_B+0.5BPIPain\_B\cdot BMI\_B\\&+1.5BPIInterf\_B\cdot DxDur+2.5BMI\_B\cdot PHQ8\_B\\&+1.5Age\cdot DxDur+2SymDur\cdot BPIPain\_B\\&+0.5BPIPain\_B\cdot BPIInterf\_B+BPIInterf\_B\cdot BMI\_B\\&+1.5BMI\_B\cdot DxDur-2SymDur\cdot SymDur\\&+1.5BPIInterf\_B\cdot BPIInterf\_B+1.5DxDur\cdot DxDur)/8.\end{aligned}$$

The true outcome models are

$$\begin{aligned}Y(0) = 1&-2Age-SymDur-0.5BPIPain\_B+3BPIInterf\_B+1.5MFIpf\_B\\&-3.5TrtDur+2.5PhysicalSymp\_B+SymDur\cdot BPIInterf\_B\\&+2.5BPIPain\_B\cdot PHQ8\_B+BPIInterf\_B\cdot TrtDur\\&+1.5MFIpf\_B\cdot PhysicalSymp\_B+2Age\cdot TrtDur\\&+0.5SymDur\cdot BPIPain\_B+BPIPain\_B\cdot BPIInterf\_B\\&+1.5BPIInterf\_B\cdot MFIpf\_B+0.5MFIpf_B\cdot TrtDur\\&-SymDur\cdot SymDur+2.5BPIInterf\_B\cdot BPIInterf\_B\\&+1.5PhysicalSymp\_B\cdot PhysicalSymp\_B+\epsilon\\Y(1) = Y(0)&+3+5Age+4BMI\_B-2MFIpf\_B-1.5GAD7\_B\\&+2Age\cdot Age+BMI\_B\cdot BMI\_B-0.5MFIpf\_B\cdot MFIpf\_B\\&+Age\cdot BMI\_B-MFIpf\_B\cdot GAD7\_B+1.5Age\cdot GAD7\_B\\&-2BMI\_B\cdot MFIpf\_B,\end{aligned}$$

B.7 Performance of the matching estimators when the propensity score and the prognostic score are highly correlated

In Section 3.3.1, we stated that the prognostic score may not have better a overlap than the propensity score when the two scores are highly correlated. Thus, the difference between PSM and PGM can be negligible. In this section, we construct an extreme example that the prognostic score is exactly the same as the linear predictor of the propensity score using the model specification in Section B.6.1. In this case, the overlap of the true prognostic score should be the same as the overlap of the true propensity score in the logit scale. We also check the performance of PSM, PGM, and DSM based on the estimated scores. As illustrated in Figure B.1, PSM performed almost equivalently as PGM. DSM had a smaller bias than the other two estimators, but the improvement is not significant. Even though the prognostic score cannot increase the efficiency in this setting, the performance of DSM won't be worse than the two single score estimators.

B.8 Additional results for variable selection in PSM and PGM

To verify that the bias in the simulation studies in Section 3.3 is finite sample bias, we increase the number of replications from 100 to 500. The sample bias decreased significantly as illustrated in Figure B.2 and B.3.

B.9 Complete results for variable selection in double score matching

In Section 3.4.3, we only showed the results for matching with a caliper and with replacement when the generating model was linear and the treatment effect was heterogeneous. To complete our statements, Figure B.4 - B.7 present the results for all the scenarios, and our conclusion for DSM holds under different situations.

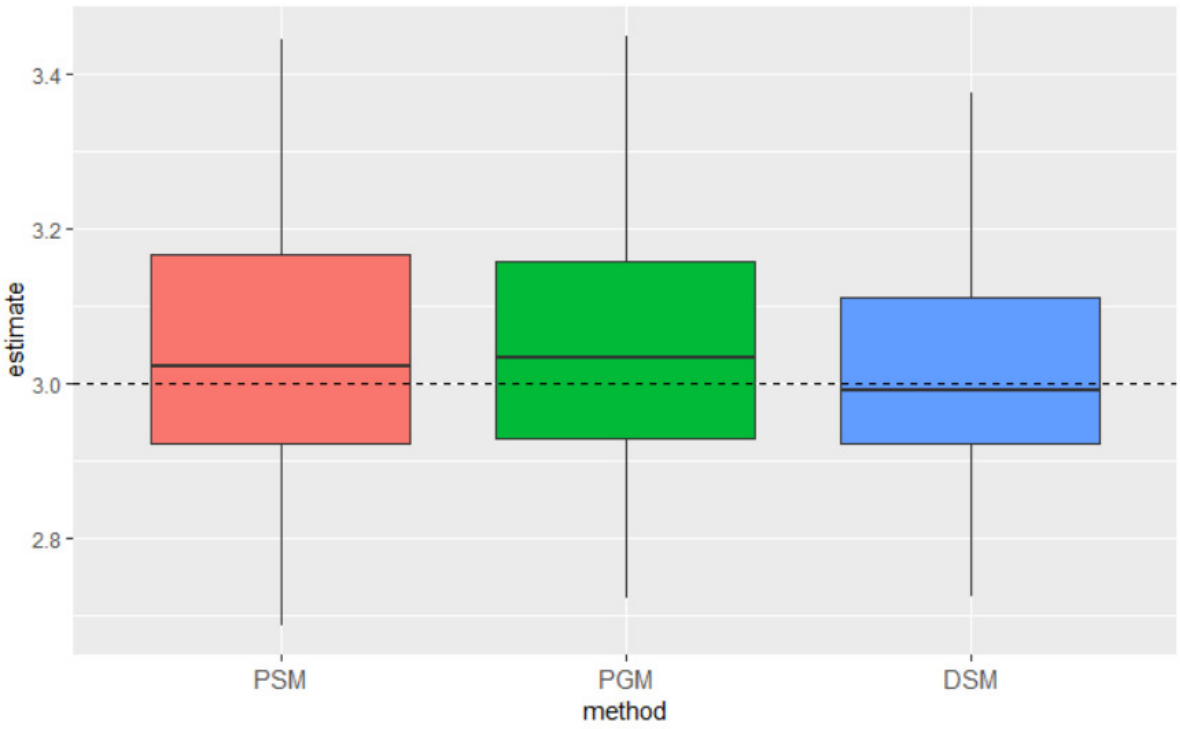


Figure B.1: Comparison of PSM, PGM, and DSM when the propensity score is equivalent as the prognostic score. Matching is done with caliper and with replacement.

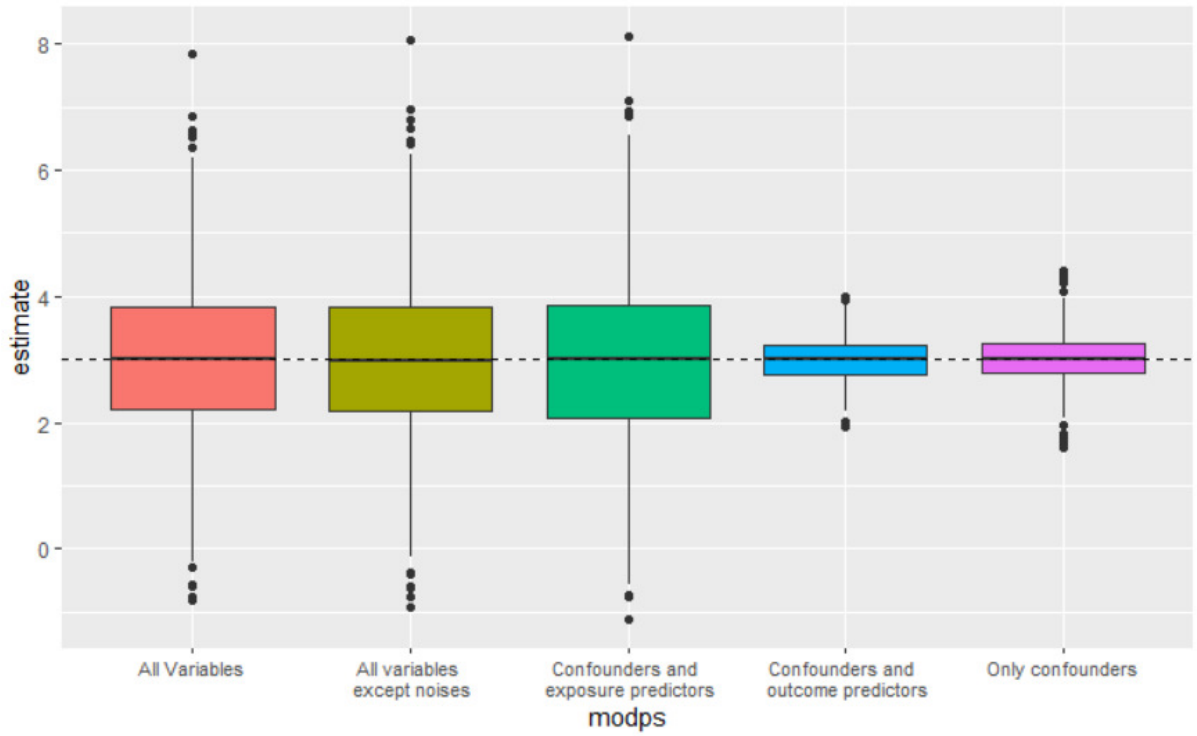


Figure B.2: Performance of PSM estimator under different variable selection strategies in the illustrating example, number of replications $K = 500$.

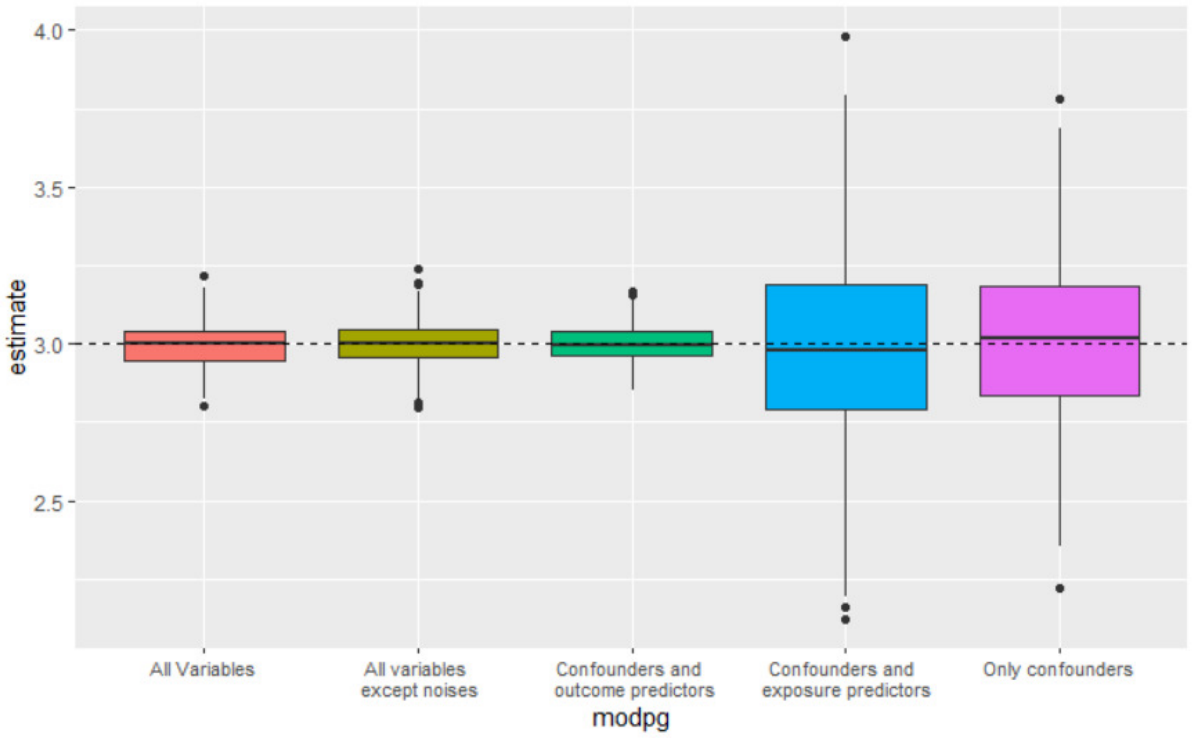


Figure B.3: Performance of PGM estimator under different variable selection strategies in the illustrating example, number of replications $K = 500$.

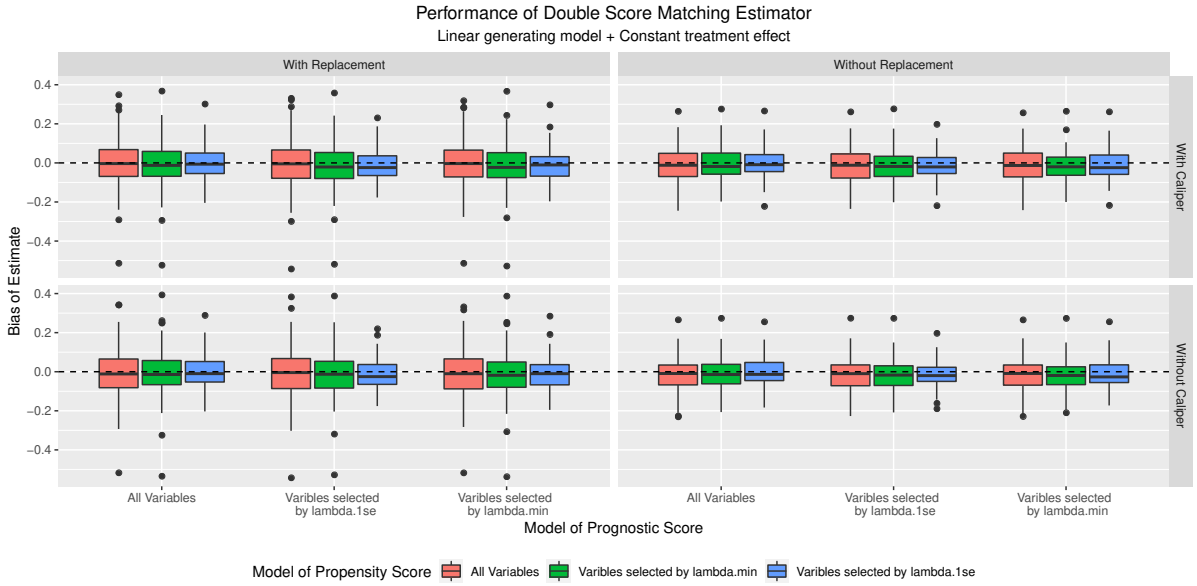


Figure B.4: Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was linear and the treatment effect was constant.

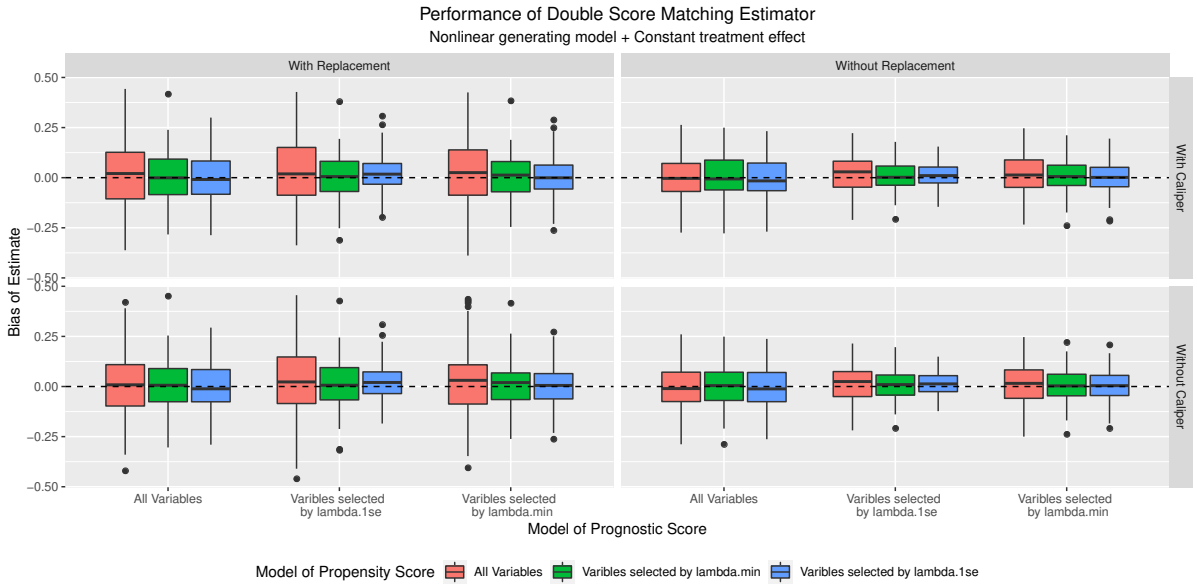


Figure B.5: Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was nonlinear and the treatment effect was constant.

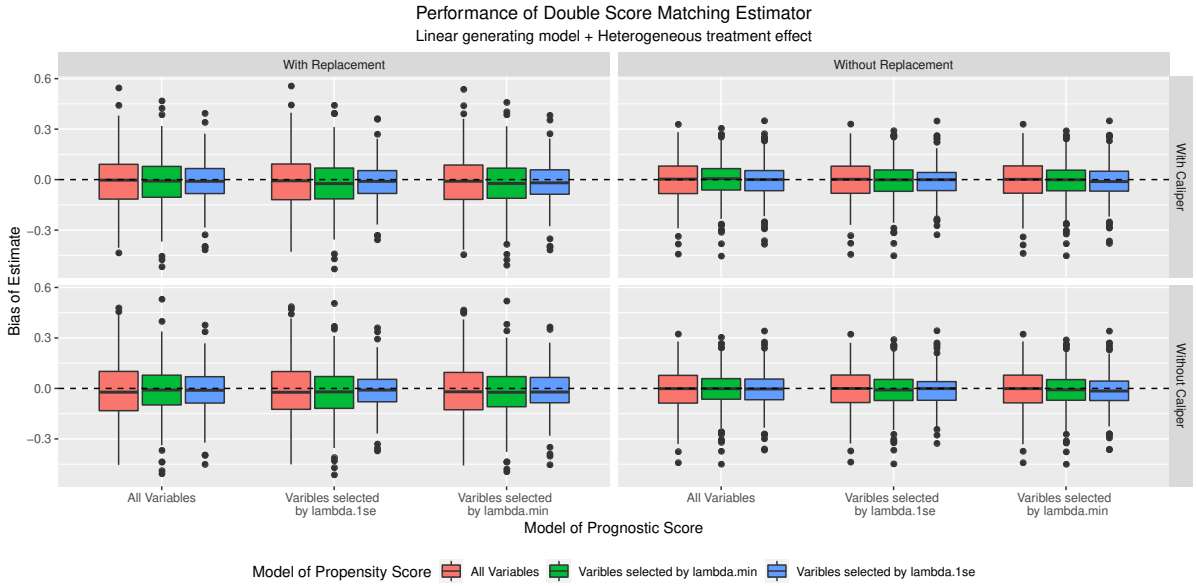


Figure B.6: Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was linear and the treatment effect was heterogeneous.

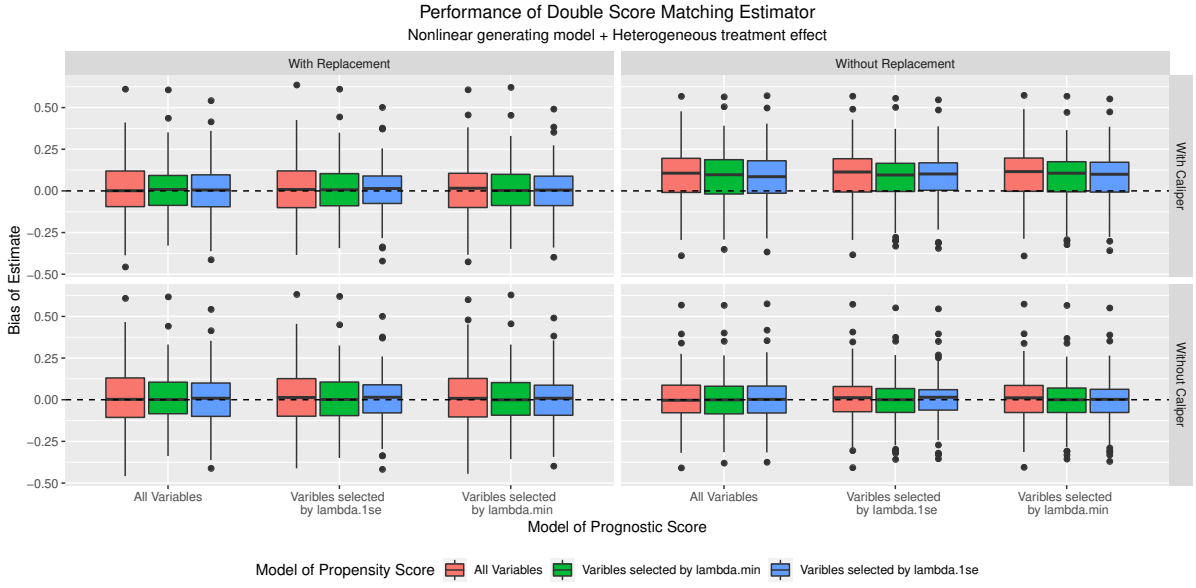


Figure B.7: Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was nonlinear and the treatment effect was heterogeneous.

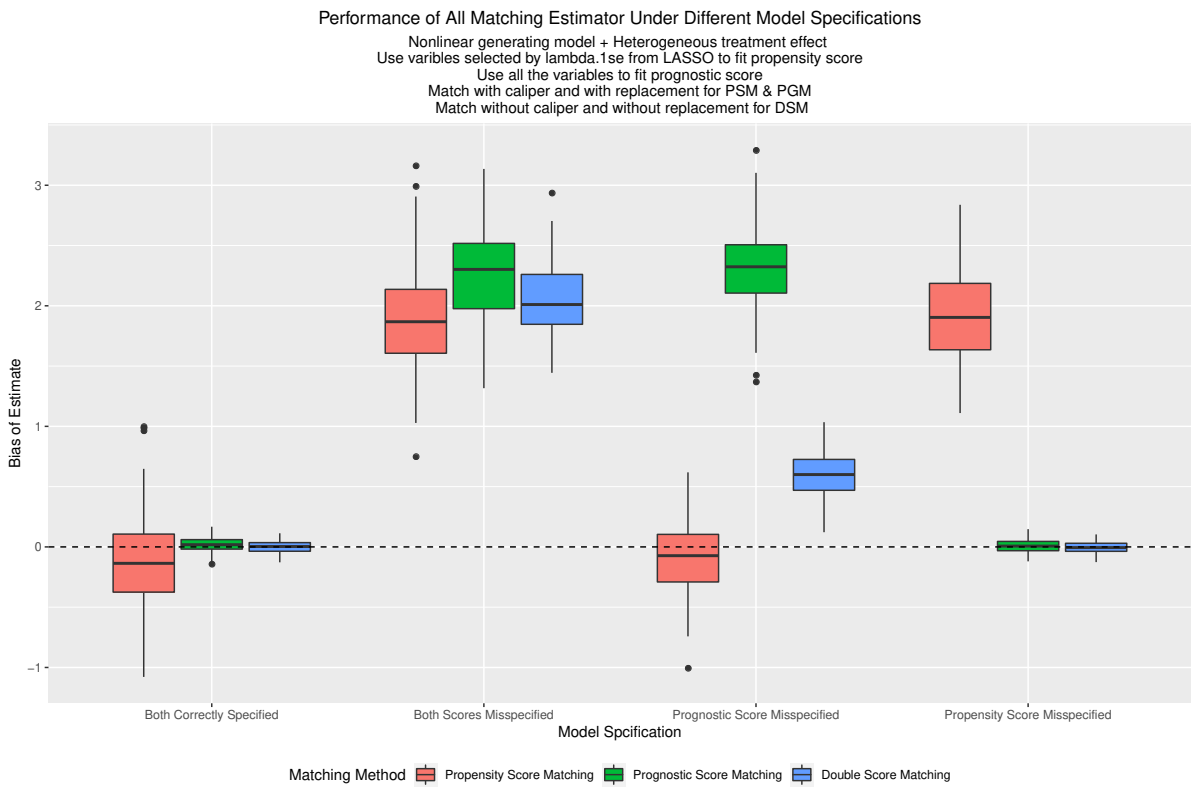


Figure B.8: Comparison of PSM, PGM, and DSM when both models might be misspecified. DSM was applied without a caliper and without replacement.

B.10 Additional results for double score matching without a caliper and without replacement

In Section 3.4.3, we recommended matching with replacement for DSM, while the caliper did not make a difference. However, as illustrated by Figure 3.14, matching without a caliper and without replacement achieved the best performance as well. The reason that we excluded this configuration is that the double robustness property of DSM was weakened under this choice of caliper and replacement, as illustrated in Figure B.8. When the propensity score was correctly specified but the prognostic score was misspecified, significant bias appeared in DSM. However, this phenomenon did not happen when we match with replacement. As a result, we do not recommend using DSM without a caliper and without replacement, even though it has satisfying performance when models are correctly specified.

APPENDIX

C

APPENDIX FOR "ON THE INFLUENCE FUNCTION OF THE DOUBLY ROBUST REGRESSION ESTIMATOR"

C.1 Analysis on (4.8): the propensity score parameter

We focus on deriving the adjustment term from the non-parametric estimate (4.9) in the influence function based on Newey's paper (Newey 1994). To apply the theorems, we need to write out the M-estimating equations for the doubly robust regression estimator. For simplicity, denote $v_i = \hat{e}(X_i)$ as the estimated propensity score and $Y_{a,i} = Y_i(a)$ as the potential outcome for treatment level a. By Taylor expansion:

$$\begin{aligned} & \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[\mu_1\{X_i, \hat{e}(X_i)\} - \mu_1\{X_i, e^*(X_i)\} \right] \\ = \; & \mathbb{E}\left[\frac{\partial}{\partial \alpha} \mu_1 \{X, e(X;\alpha)\} \mid_{\alpha=\alpha^*} \right] n^{1/2} (\hat{\alpha} - \alpha^*) \end{aligned}$$

The Bahadur representation of $\hat{\alpha}$ can be easily derived from the standard M-estimation theory.

$$\hat{\alpha}-\alpha^*=\frac{1}{n}\sum_{i=1}^{n}M_e^{-1}\psi_e(X_iA_i;\alpha^*)+o_p(1),$$

where $M_e=\mathbb{E}\left\{-\frac{\partial}{\partial\alpha}\psi_e(X,A;\alpha)\,|_{\alpha=\alpha^*}\right\}$. Thus, we only need to derive the derivative of $\mathbb{E}[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}\,|_{\alpha=\alpha^*}]$ and evaluate at α^*, where $\alpha^*=\text{plim}\hat{\alpha}$. By (4.7), it can be written as

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}\,|_{\alpha=\alpha^*}\right]=\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=a\}\right]|_{\alpha=\alpha^*}\\
&+\mathbb{E}\left[\frac{\partial\, r_{X,1}\{e(X;\alpha)\}^{\mathrm{T}}}{\partial\alpha}\beta_1(\alpha^*)\right]|_{\alpha=\alpha^*}\\
&+\mathbb{E}\left[r_{X,1}\{e(X;\alpha^*)\}^{\mathrm{T}}\frac{\partial\beta_1(\alpha)}{\partial\alpha}\right]|_{\alpha=\alpha^*}\\
=\;&\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=1\}\,|_{\alpha=\alpha^*}\right] &\text{(C.1)}\\
&-\mathbb{E}\left[\frac{\partial\,\mathbb{E}\{X\mid e(X;\alpha),A=1\}^{\mathrm{T}}}{\partial\alpha}\right]|_{\alpha=\alpha^*}\beta_1(\alpha^*) &\text{(C.2)}\\
&+\mathbb{E}\left[r_{X,1}\{e(X;\alpha^*)\}^{\mathrm{T}}\right]\frac{\partial\beta_1(\alpha)}{\partial\alpha}\,|_{\alpha=\alpha^*}, &\text{(C.3)}
\end{aligned}$$

where $\beta_1(\alpha)$ is the plim of $\hat{\beta}_1(\alpha)$ as $n\to\infty$, which is a function of α, not X or Y. For simplicity, denote $r^*_{X,1}=r_{X,1}\{e(X;\alpha^*)\}=r_{X,1}\{e^*(X)\}$ and $r^*_{Y,1}=r_{Y,1}\{e(X;\alpha^*)\}=r_{Y,1}\{e^*(X)\}$. Then

$$\begin{aligned}
\beta_1(\alpha) &= \mathbb{E}\big(r_{X,1}r_{X,1}^{\mathrm{T}}\mid A=1\big)^{-1}\mathbb{E}\big(r_{X,1}r_{Y,1}\mid A=1\big)\\
\beta_1(\alpha^*) &= \mathbb{E}\big(r^*_{X,1}r^{*\mathrm{T}}_{X,1}\mid A=1\big)^{-1}\mathbb{E}\big(r^*_{X,1}r^*_{Y,1}\mid A=1\big).
\end{aligned}$$

C.1.1 Analysis on (C.1) and (C.2): derivatives of the conditional expectations

To derive the derivative in the first and second terms, we need to apply the orthogonality conditions from the conditional expectations. Note that

$$\mathbb{E}(h_1\{e(X;\alpha)\}A[Y-\mathbb{E}\{Y \mid e(X;\alpha), A=1\}]) = 0 \tag{C.4}$$

$$\mathbb{E}(h_2\{e(X;\alpha)\}A[X-\mathbb{E}\{X \mid e(X;\alpha), A=1\}]) = 0. \tag{C.5}$$

Take $h_1\{e(X;\alpha)\} = \pi\{e(X;\alpha)\}^{-1}$, where $\pi\{e(X;\alpha)\} = P\{A=1 \mid e(X;\alpha)\}$, we get

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

Apply the law of total expectation theorem given X on the left hand side, the previous equation changes into

$$\mathbb{E}\left[\frac{e(X)\mathbb{E}(Y \mid X, A=1)}{\pi\{e(X;\alpha)\}}\right] = \mathbb{E}\left[\frac{A\mathbb{E}\{Y \mid e(X;\alpha), A=1\}}{\pi\{e(X;\alpha)\}}\right]. \tag{C.6}$$

Take derivative with respect to α on both sides and evaluate at α^*. For the left hand side, we have

$$\mathbb{E}\left[-\frac{e(X)\mathbb{E}(Y \mid X, A=1)}{\pi\{e(X;\alpha)\}^2} \cdot \frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\right]|_{\alpha=\alpha^*}.$$

For the right hand side, we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{\pi\{e(X;\alpha)\}A\frac{\partial}{\partial\alpha}\mathbb{E}\{Y \mid e(X;\alpha), A=1\} - A\mathbb{E}\{Y \mid e(X;\alpha), A=1\}\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}}{\pi\{e(X;\alpha)\}^2}\right]|_{\alpha=\alpha^*} \\
= \;& \mathbb{E}\left[\frac{A}{\pi\{e(X;\alpha^*)\}} \cdot \frac{\partial}{\partial\alpha}\mathbb{E}\{Y \mid e(X;\alpha), A=1\}|_{\alpha=\alpha^*}\right] \\
& -\mathbb{E}\left[\frac{e(X)\mathbb{E}\{Y \mid e(X;\alpha^*), A=1\}}{\pi\{e(X;\alpha^*)\}^2} \cdot \frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}|_{\alpha=\alpha^*}\right].
\end{aligned}$$

Following Hahn and Ridder (2013), we further decompose the first derivative into two terms. Note that α affects $\mathbb{E}\{Y \mid e(X;\alpha), A=1\}$ through the value of the propensity score and the term that is conditioned on. To clarify these two sources of variation, we write $\mathbb{E}\{Y \mid e(X;\alpha), A=1\}$ as $\mathbb{E}\{Y \mid e(X;\alpha_2) = e(X;\alpha_1), A=1\}$ and then apply the chain rule:

$$
\begin{aligned}
& \mathbb{E}\left[\frac{A}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=1\}\mid_{\alpha=\alpha^*}\right] \\
= \; & \mathbb{E}\left[\frac{A}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial}{\partial\alpha_1}\mathbb{E}\{Y\mid e(X;\alpha^*)=e(X;\alpha_1),A=1\}\mid_{\alpha_1=\alpha^*}\right] \\
& +\mathbb{E}\left[\frac{A}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial}{\partial\alpha_2}\mathbb{E}\{Y\mid e(X;\alpha_2)=e(X;\alpha^*),A=1\}\mid_{\alpha_2=\alpha^*}\right] \\
= \; & \mathbb{E}\left[\frac{e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\mathbb{E}\{Y\mid v,A=1\}}{\partial v}\mid_{v=v^*}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\mid_{\alpha=\alpha^*}\right] \\
& +\mathbb{E}\left[\frac{\partial}{\partial\alpha_2}\mathbb{E}\{Y\mid e(X;\alpha_2)=e(X;\alpha^*),A=1\}\mid_{\alpha_2=\alpha^*}\right],
\end{aligned}
$$

where $v=e(X;\alpha)$ and $v^*=e(X;\alpha^*)$. The first term is derived by conditioning on X, while the second term is based on taking conditional expectation given $e(X;\alpha^*)$ since $\partial\mathbb{E}\{Y\mid e(X;\alpha_2)=e(X;\alpha^*),A=1\}/\partial\alpha_2\mid_{\alpha_2=\alpha^*}$is a function of $e(X;\alpha^*)$. Plug the previous results back to (C.6):

$$
\begin{aligned}
& \mathbb{E}\left[\frac{\partial}{\partial\alpha_2}\mathbb{E}\{Y\mid e(X;\alpha_2)=e(X;\alpha^*),A=1\}\mid_{\alpha_2=\alpha^*}\right] \\
= \; & \mathbb{E}\left[\frac{e(X)\mathbb{E}\{Y\mid e(X;\alpha^*),A=1\}}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right] \\
& -\mathbb{E}\left[\frac{e(X)\mathbb{E}(Y\mid X,A=1)}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right] \\
& -\mathbb{E}\left[\frac{e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\mathbb{E}\{Y\mid v,A=1\}}{\partial v}\mid_{v=v^*}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\mid_{\alpha=\alpha^*}\right].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& \mathbb{E}\left[\frac{\partial}{\partial \alpha}\mathbb{E}\{Y \mid e(X;\alpha), A=1\}\,|_{\alpha=\alpha^*}\right] \\
= \;& \mathbb{E}\left[\frac{\partial}{\partial \alpha_1}\mathbb{E}\{Y \mid e(X;\alpha^*)=e(X;\alpha_1), A=1\}\,|_{\alpha_1=\alpha^*}\right] \\
& +\mathbb{E}\left[\frac{\partial}{\partial \alpha_2}\mathbb{E}\{Y \mid e(X;\alpha_2)=e(X;\alpha^*), A=1\}\,|_{\alpha_2=\alpha^*}\right] \\
= \;& \mathbb{E}\left(\frac{e(X)[\mathbb{E}\{Y \mid e(X;\alpha^*), A=1\}-\mathbb{E}(Y \mid X, A=1)]}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial \alpha}\pi\{e(X;\alpha)\}\,|_{\alpha=\alpha^*}\right) \\
& -\mathbb{E}\left[\frac{\pi\{e(X;\alpha^*)\}-e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\,\mathbb{E}\{Y \mid v, A=1\}}{\partial v}\,|_{v=v^*}\cdot\frac{\partial\, e(X;\alpha)}{\partial \alpha}\,|_{\alpha=\alpha^*}\right].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \mathbb{E}\left[\frac{\partial}{\partial \alpha}\mathbb{E}\{X \mid e(X;\alpha), A=1\}\,|_{\alpha=\alpha^*}\right] \\
= \;& \mathbb{E}\left(\frac{e(X)[\mathbb{E}\{X \mid e(X;\alpha^*), A=1\}-X]}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial \alpha}\pi\{e(X;\alpha)\}\,|_{\alpha=\alpha^*}\right) \\
& -\mathbb{E}\left[\frac{\pi\{e(X;\alpha^*)\}-e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\,\mathbb{E}\{X \mid v, A=1\}}{\partial v}\,|_{v=v^*}\cdot\frac{\partial\, e(X;\alpha)}{\partial \alpha}\,|_{\alpha=\alpha^*}\right].
\end{aligned}
$$

C.1.2 Analysis on (C.3): derivative of $\beta_1(\alpha)$

$\beta_1(\alpha)$ is the limit of the ordinary least squares estimator $\hat{\beta}_1(\alpha)$ and thus

$$
\beta_1(\alpha)=\mathbb{E}\big(r_{X,1}r_{X,1}^{\mathrm{T}} \mid A=1\big)^{-1}\mathbb{E}\big(r_{X,1}r_{Y,1} \mid A=1\big).
$$

Or equivalently,

$$
\mathbb{E}\big(r_{X,1}r_{X,1}^{\mathrm{T}} \mid A=1\big)\cdot\beta_1(\alpha)=\mathbb{E}\big(r_{X,1}r_{Y,1} \mid A=1\big).
$$

Take derivative on both sides and apply the product rule for matrices,

$$
\mathbb{E}\big(r_{X,1}r_{X,1}^{\mathrm{T}} \mid A=1\big)\cdot\frac{\partial\,\beta_1(\alpha)}{\partial \alpha}+\sum_{j=1}^{p}\frac{\partial}{\partial \alpha_j}\mathbb{E}\big(r_{X,1}r_{X,1}^{\mathrm{T}} \mid A=1\big)\cdot B_j=\frac{\partial}{\partial \alpha}\mathbb{E}\big(r_{X,1}r_{Y,1} \mid A=1\big),
$$

where $B_j = \left(\underbrace{\mathbf{0},...,\mathbf{0}}_{j-1}, \beta_1(\alpha), \mathbf{0},...,\mathbf{0} \right)$ is a $p \times p$ matrix with the j-th column as $\beta_1(\alpha)$ and other columns as zero vectors. Thus,

$$\frac{\partial \beta_1(\alpha)}{\partial \alpha} = \mathbb{E}\left(r_{X,1} r_{X,1}^{\mathrm{T}} \mid A=1\right)^{-1} \left\{ \frac{\partial}{\partial \alpha} \mathbb{E}\left(r_{X,1} r_{Y,1} \mid A=1\right) - \sum_{j=1}^{p} \frac{\partial}{\partial \alpha_j} \mathbb{E}\left(r_{X,1} r_{X,1}^{\mathrm{T}} \mid A=1\right) \cdot B_j \right\}.$$

We then derive the derivatives using the techniques in Section C.1.1. By the product rule,

$$\begin{aligned} \frac{\partial}{\partial \alpha} \mathbb{E}\left(r_{X,1} r_{Y,1} \mid A=1\right) &= \mathbb{E}\left(\frac{\partial r_{X,1}}{\partial \alpha} r_{Y,1} \mid A=1 \right) + \mathbb{E}\left(r_{X,1} \frac{\partial r_{Y,1}}{\partial \alpha} \mid A=1 \right) \\ &= -\mathbb{E}\left(\frac{\partial}{\partial \alpha} \mathbb{E}\{X \mid e(X;\alpha), A=1\} \cdot r_{Y,1} \mid A=1 \right) \\ &\quad -\mathbb{E}\left(r_{X,1} \cdot \frac{\partial}{\partial \alpha} \mathbb{E}\{Y \mid e(X;\alpha), A=1\} \mid A=1 \right). \end{aligned}$$

We only need to derive the first term because the two terms have similar structures. By decomposing the effect of α into two sources,

$$\begin{aligned} & \mathbb{E}\left(\frac{\partial}{\partial \alpha} \mathbb{E}\{X \mid e(X;\alpha), A=1\} \cdot r_{Y,1} \mid A=1 \right) \\ = \; & \mathbb{E}\left(\frac{\partial}{\partial \alpha_1} \mathbb{E}\{X \mid e(X;\alpha) = e(X;\alpha_1), A=1\} \mid_{\alpha_1=\alpha} \cdot r_{Y,1} \mid A=1 \right) \\ & + \mathbb{E}\left(\frac{\partial}{\partial \alpha_2} \mathbb{E}\{X \mid e(X;\alpha_2) = e(X;\alpha), A=1\} \mid_{\alpha_2=\alpha} \cdot r_{Y,1} \mid A=1 \right) \\ = \; & \mathbb{E}\left(\frac{\partial \mathbb{E}\{X \mid v, A=1\}}{\partial v} \mid_{v=e(X;\alpha)} \cdot \frac{\partial e(X;\alpha)}{\partial \alpha} \cdot r_{Y,1} \mid A=1 \right) \\ & + \mathbb{E}\left(\frac{\partial}{\partial \alpha_2} \mathbb{E}\{X \mid e(X;\alpha_2) = e(X;\alpha), A=1\} \mid_{\alpha_2=\alpha} \cdot \mathbb{E}\left\{ r_{Y,1} \mid e(X;\alpha), A=1 \right\} \mid A=1 \right) \\ = \; & \mathbb{E}\left(\frac{\partial \mathbb{E}\{X \mid v, A=1\}}{\partial v} \mid_{v=e(X;\alpha)} \cdot \frac{\partial e(X;\alpha)}{\partial \alpha} \cdot r_{Y,1} \mid A=1 \right). \end{aligned}$$

The derivation follows because $\partial \mathbb{E}\{X \mid e(X;\alpha_2) = e(X;\alpha), A=1\} / \partial \alpha_2$ is a function of $e(X;\alpha)$ and $\mathbb{E}\left\{ r_{Y,1} \mid e(X;\alpha), A=1 \right\} = 0$ by its definition. Thus,

$$\begin{aligned}\frac{\partial}{\partial\alpha}\mathbb{E}\big(r_{X,1}r_{Y,1}\mid A=1\big) &= -\mathbb{E}\left(\frac{\partial\mathbb{E}\{X\mid v,A=1\}}{\partial v}\mid_{v=e(X;\alpha)}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\cdot r_{Y,1}\mid A=1\right)\\ &\quad -\mathbb{E}\left(r_{X,1}\cdot\frac{\partial\mathbb{E}\{Y\mid v,A=1\}}{\partial v}\mid_{v=e(X;\alpha)}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha}\mid A=1\right).\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial}{\partial\alpha_j}\mathbb{E}\big(r_{X,1}r_{X,1}^{\mathrm{T}}\mid A=1\big) &= -\mathbb{E}\left(\frac{\partial\mathbb{E}\{X\mid v,A=1\}}{\partial v}\mid_{v=e(X;\alpha)}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha_j}\cdot r_{X,1}^{\mathrm{T}}\mid A=1\right)\\ &\quad -\mathbb{E}\left(\frac{\partial\mathbb{E}\{X\mid v,A=1\}}{\partial v}\mid_{v=e(X;\alpha)}\cdot\frac{\partial e(X;\alpha)}{\partial\alpha_j}\cdot r_{X,1}^{\mathrm{T}}\mid A=1\right)^{\mathrm{T}}.\end{aligned}$$

In all, the adjustment term from the propensity score parameter is

$$\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right]M_e^{-1}\psi_e(A,X;\alpha^*),$$

where

$$\begin{aligned}\psi_e(A,X;\alpha) &= \frac{\partial e(X;\alpha)}{\partial\alpha}\cdot\frac{A-e(X;\alpha)}{e(X;\alpha)\{1-e(X;\alpha)\}},\\ M_e &= \mathbb{E}\left\{-\frac{\partial}{\partial\alpha}\psi_e(A,X;\alpha)\mid_{\alpha=\alpha^*}\right\},\\ \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right] &= \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y\mid e(X;\alpha),A=1\}\mid_{\alpha=\alpha^*}\right]\\ &\quad -\mathbb{E}\left[\frac{\partial\mathbb{E}\{X\mid e(X;\alpha),A=1\}^{\mathrm{T}}}{\partial\alpha}\right]\mid_{\alpha=\alpha^*}\beta_1(\alpha^*)\\ &\quad +\mathbb{E}\big[r_{X,1}\{e(X;\alpha^*)\}^{\mathrm{T}}\big]\frac{\partial\beta_1(\alpha)}{\partial\alpha}\mid_{\alpha=\alpha^*},\end{aligned}$$

and $\beta_1(\alpha^*)=\mathbb{E}\big(r_{X,1}^*r_{X,1}^{*\mathrm{T}}\mid A=1\big)^{-1}\mathbb{E}\big(r_{X,1}^*r_{Y,1}^*\mid A=1\big)$,

$$
\begin{aligned}
& \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{Y \mid e(X;\alpha), A=1\}\mid_{\alpha=\alpha^*}\right] \\
= \;& \mathbb{E}\left(\frac{e(X)[\mathbb{E}\{Y \mid e(X;\alpha^*), A=1\} - \mathbb{E}(Y \mid X, A=1)]}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right) \\
& -\mathbb{E}\left[\frac{\pi\{e(X;\alpha^*)\} - e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\,\mathbb{E}\{Y \mid \nu, A=1\}}{\partial\,\nu}\mid_{\nu=\nu^*}\cdot\frac{\partial\, e(X;\alpha)}{\partial\,\alpha}\mid_{\alpha=\alpha^*}\right] \\
& \mathbb{E}\left[\frac{\partial}{\partial\alpha}\mathbb{E}\{X \mid e(X;\alpha), A=1\}\mid_{\alpha=\alpha^*}\right] \\
= \;& \mathbb{E}\left(\frac{e(X)[\mathbb{E}\{X \mid e(X;\alpha^*), A=1\} - X]}{\pi\{e(X;\alpha^*)\}^2}\cdot\frac{\partial}{\partial\alpha}\pi\{e(X;\alpha)\}\mid_{\alpha=\alpha^*}\right) \\
& -\mathbb{E}\left[\frac{\pi\{e(X;\alpha^*)\} - e(X)}{\pi\{e(X;\alpha^*)\}}\cdot\frac{\partial\,\mathbb{E}\{X \mid \nu, A=1\}}{\partial\,\nu}\mid_{\nu=\nu^*}\cdot\frac{\partial\, e(X;\alpha)}{\partial\,\alpha}\mid_{\alpha=\alpha^*}\right],
\end{aligned}
$$

and the derivative for $\beta_1(\alpha)$ is provided in this section.

C.2 Analysis on (4.9): the non-parametric estimation

First, we would like to decompose (4.9) for further analysis. To simplify notations, we use $\hat{\beta}_1^*$ to denote $\hat{\beta}_1(\alpha^*)$ and β_1^* to denote its limit in probability.

$$
\begin{aligned}
& \hat{\mu}_1\{X_i, e^*(X_i)\} - \mu_1\{X_i, e^*(X_i)\} && \text{(C.7)} \\
= \;& \hat{\mathbb{E}}\{Y \mid e^*(X_i), A=1\} + \left[X_i - \hat{\mathbb{E}}\{X \mid e^*(X_i), A=1\}\right]^{\mathrm{T}}\hat{\beta}_1^* \\
& -\mathbb{E}\{Y \mid e^*(X_i), A=1\} - [X_i - \mathbb{E}\{X \mid e^*(X_i), A=1\}]^{\mathrm{T}}\beta_1^* \\
= \;& \hat{\mathbb{E}}\{Y \mid e^*(X_i), A=1\} - \mathbb{E}\{Y \mid e^*(X_i), A=1\} \\
& +\left[X_i - \hat{\mathbb{E}}\{X \mid e^*(X_i), A=1\}\right]^{\mathrm{T}}\hat{\beta}_1^* - [X_i - \mathbb{E}\{X \mid e^*(X_i), A=1\}]^{\mathrm{T}}\hat{\beta}_1^* \\
& +[X_i - \mathbb{E}\{X \mid e^*(X_i), A=1\}]^{\mathrm{T}}\hat{\beta}_1^* - [X_i - \mathbb{E}\{X \mid e^*(X_i), A=1\}]^{\mathrm{T}}\beta_1^* \\
= \;& \hat{\mathbb{E}}\{Y \mid e^*(X_i), A=1\} - \mathbb{E}\{Y \mid e^*(X_i), A=1\} \\
& -\left[\hat{\mathbb{E}}\{X \mid e^*(X_i), A=1\} - \mathbb{E}\{X \mid e^*(X_i), A=1\}\right]^{\mathrm{T}}\hat{\beta}_1^* \\
& +[X_i - \mathbb{E}\{X \mid e^*(X_i), A=1\}]^{\mathrm{T}}\left(\hat{\beta}_1^* - \beta_1^*\right) \\
\cong \;& \hat{\mathbb{E}}\{Y \mid e^*(X_i), A=1\} - \mathbb{E}\{Y \mid e^*(X_i), A=1\} && \text{(C.8)} \\
& -\left[\hat{\mathbb{E}}\{X \mid e^*(X_i), A=1\} - \mathbb{E}\{X \mid e^*(X_i), A=1\}\right]^{\mathrm{T}}\beta_1^* && \text{(C.9)} \\
& +[X_i - \mathbb{E}\{X \mid e^*(X_i), A=1\}]^{\mathrm{T}}\left(\hat{\beta}_1^* - \beta_1^*\right), && \text{(C.10)}
\end{aligned}
$$

where $\cong$ is because $\left[\hat{\mathbb{E}}\{X \mid e^*(X_i), A=1\} - \mathbb{E}\{X \mid e^*(X_i), A=1\}\right]^{\mathrm{T}}\left(\hat{\beta}_1^* - \beta_1^*\right)$ as a higher order term can be negligible. (C.8) stands for the adjustment term from the nonparametric estimator $\hat{\mathbb{E}}\{Y \mid e^*(X_i), A=1\}$, and (C.9) is the adjustment term from $\hat{\mathbb{E}}\{X \mid e^*(X_i), A=1\}$ multiplied by β_1^*. (C.10) accounts for the variability from estimating $\hat{\beta}_1^*$, which comes from a M-estimating equation with nonpametric terms involved. These can all be derived from the pathwise derivative based on Newey's paper (Newey 1994). Denote $\{F_\theta\}$ as a path of the distribution of $\{X, A, Y\}$, and it is equal to the true distribution when $\theta = 0$. To simplify notations, we further denote $\lambda_1(X, \theta) = \mathbb{E}_\theta\{Y \mid e^*(X), A=1\}$, $\lambda_2(X, \theta) = \mathbb{E}_\theta\{X \mid e^*(X), A=1\}$, $\lambda_1(X) = \mathbb{E}\{Y \mid e^*(X), A=1\}$, and $\lambda_2(X) = \mathbb{E}\{X \mid e^*(X), A=1\}$.

C.2.1 Analysis on (C.8) and (C.9): the conditional expectations

We formalize how we would apply Newey's techniques to handle nonparametric terms in the estimating equations. Denote

$$\hat{\mu}_Y(F_\theta) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}_\theta\{Y \mid e^*(X_i), A=1\}$$

and

$$\mu_Y(F_\theta) = \text{plim}\hat{\mu}_Y(F_\theta) = \mathbb{E}_\theta[\mathbb{E}_\theta\{Y \mid e^*(X), A=1\}] = \mathbb{E}_\theta\{\lambda_1(X, \theta)\}.$$

If we can find $d_1(X, A, Y)$ with $\mathbb{E}\{d_1(X, A, Y)\} = 0$ and $\mathbb{E}\{\|d_1(X, A, Y)\|^2\} < \infty$ such that

$$\frac{\partial \mu_Y(F_\theta)}{\partial \theta} = \mathbb{E}\{d_1(X, A, Y) S(X, A, Y)\},$$

where $S(X, A, Y)$ is the score function for the true distribution. Then by Theorem 2.1 in Newey's paper, $d_1(X, A, Y)$ is the corresponding influence function for $\hat{\mu}_Y$, i.e.

$$\hat{\mu}_Y - \mu_Y = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}\{Y \mid e^*(X_i), A=1\} - \mu_Y = \frac{1}{n}\sum_{i=1}^{n} d_1(X, A, Y) + o_p(1),$$

where $\hat{\mu}_Y = \hat{\mu}_Y(F_0)$ and $\mu_Y = \mu_Y(F_0)$ are the simplified notations under the true distribution. By the product rule,

$$\begin{aligned}\frac{\partial \mu_Y(F_\theta)}{\partial \theta} &= \frac{\partial \mathbb{E}_\theta\{\lambda_1(X,\theta)\}}{\partial \theta} \\ &= \mathbb{E}\{\lambda_1(X)S(X,A,Y)\} + \frac{\partial \mathbb{E}\{\lambda_1(X,\theta)\}}{\partial \theta} \\ &= \mathbb{E}\left[\left\{\lambda_1(X)-\mu_Y\right\}S(X,A,Y)\right] + \frac{\partial \mathbb{E}\{\lambda_1(X,\theta)\}}{\partial \theta}.\end{aligned}$$

The first term has already been a outer product form we need. Thus, if we can find a $\alpha_1(X,A,Y)$ such that $\mathbb{E}\{\alpha_1(X,A,Y)\}=0$ and

$$\frac{\partial \mathbb{E}\{\lambda_1(X,\theta)\}}{\partial \theta} = \mathbb{E}\{\alpha_1(X,A,Y)S(X,A,Y)\},$$

the influence function can be derived as $d_1(X,A,Y)=\left\{\lambda_1(X)-\mu_Y\right\}+\alpha_1(X,A,Y)$. The first term is included in (4.10), and the second term accounts for the adjustment of nonparametric estimation in (C.8).

Note that the orthogonality condition C.4 applies for aribitrary θ:

$$\mathbb{E}_\theta\left(h_1\{e^*(X)\}A[Y-\mathbb{E}_\theta\{Y \mid e^*(X), A=1\}]\right)=0.$$

Let $h_1\{e^*(X)\}=\pi\{e^*(X)\}^{-1}=\mathbb{E}\{A \mid e^*(X)\}^{-1}$, then the previous equation becomes

$$\mathbb{E}_\theta\left[\frac{AY}{\pi\{e^*(X)\}}\right]=\mathbb{E}_\theta\left[\frac{A\lambda_1(X,\theta)}{\pi\{e^*(X)\}}\right]$$

Take derivative on both sides,

$$\begin{aligned}\frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{AY}{\pi\{e^*(X)\}}\right] &= \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{A\lambda_1(X)}{\pi\{e^*(X)\}}\right] + \frac{\partial}{\partial \theta}\mathbb{E}\left[\frac{A\lambda_1(X,\theta)}{\pi\{e^*(X)\}}\right] \\ &= \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{A\lambda_1(X)}{\pi\{e^*(X)\}}\right] + \frac{\partial \mathbb{E}[\lambda_1(X,\theta)]}{\partial \theta},\end{aligned}$$

which is because $\lambda_1(X,\theta)$ is a function of $e^*(X)$. Then

$$\begin{aligned}\frac{\partial \mathbb{E}[\lambda_1(X,\theta)]}{\partial \theta} &= \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{AY}{\pi\{e^*(X)\}}\right]-\frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{A\lambda_1(X)}{\pi\{e^*(X)\}}\right]\\ &= \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{A\{Y-\lambda_1(X)\}}{\pi\{e^*(X)\}}\right]\\ &= \mathbb{E}\left[\frac{A\{Y-\lambda_1(X)\}}{\pi\{e^*(X)\}}\cdot S(X,A,Y)\right].\end{aligned}$$

Thus, the adjustment term for (C.8) is

$$\alpha_1(X,A,Y)=\frac{A\{Y-\lambda_1(X)\}}{\pi\{e^*(X)\}}.$$

Similarly, the adjustment term for (C.9) is

$$\alpha_2(X,A,Y)=\frac{A\{X-\lambda_2(X)\}^{\mathrm{T}}\beta_1^*}{\pi\{e^*(X)\}}.$$

C.2.2 Analysis on (C.10): the OLS estimator $\hat{\beta}_1^*$

The total contribution from (C.10) is

$$\begin{aligned}&\frac{1}{n^{1/2}}\sum_{i=1}^{n}[X_i-\mathbb{E}\{X\mid e^*(X_i),A=1\}]^{\mathrm{T}}\left(\hat{\beta}_1^*-\beta_1^*\right)\\ =\;& n^{1/2}\cdot\overline{r_{X,1,i}^*}^{\mathrm{T}}\cdot\left(\hat{\beta}_1^*-\beta_1^*\right)\\ \cong\;& n^{1/2}\cdot\mathbb{E}\left(r_{X,1}^*\right)^{\mathrm{T}}\cdot\left(\hat{\beta}_1^*-\beta_1^*\right),\end{aligned}$$

where $\cong$ follows by Taylor expansion. $\hat{\beta}_1^*$ is the solution of the following M-estimating equation:

$$\sum_{i=1}^{n}m_{\beta,1}\left(X_i,A_i,Y_i;\beta_1,\hat{h}\right)=\sum_{i=1}^{n}A_i\left(\hat{r}_{X,1,i}^*\hat{r}_{X,1,i}^{*\mathrm{T}}\beta_1-\hat{r}_{X,1,i}^*\hat{r}_{Y,1,i}^*\right)=0,$$

where $\hat{r}_{Y,1,i}^*=Y-\hat{\mathbb{E}}\{Y\mid e^*(X_i),A=1\}$, $\hat{r}_{X,1,i}^*=X-\hat{\mathbb{E}}\{X\mid e^*(X_i),A=1\}$, and $\hat{h}$ accounts for the nonparametric estimators of the conditional expectations. For a path $\{F_\theta\}$, let $h(\theta)=h(F_\theta)=(\mathbb{E}_\theta\{Y\mid e^*(X),A=1\},\mathbb{E}_\theta\{X\mid e^*(X),A=1\})$. We also denote $h_0=h(F_0)=(\mathbb{E}\{Y\mid e^*(X),A=1\},\mathbb{E}\{X\mid e^*(X),A=1\})$. Then by (3.10) in Newey (1994), the influence function of $\hat{\beta}_1^*$ is

$$\psi_{\beta,1}(X,A,Y) = -M_{\beta,1}^{-1}\left\{m_{\beta,1}\left(X,A,Y;\beta_1^*,h_0\right)+\alpha_{\beta,1}(X,A,Y)\right\},$$

where

$$M_{\beta,1} = \frac{\partial \mathbb{E}\left\{m_{\beta,1}\left(X,A,Y;\beta_1,h_0\right)\right\}}{\partial \beta_1}\,|_{\beta_1=\beta_1^*} = \mathbb{E}\left(A r_{X,1}^* r_{X,1}^{*\mathrm{T}}\right),$$

and $\alpha_{\beta,1}(X,A,Y)$ satisfies $\mathbb{E}\left\{\alpha_{\beta,1}(X,A,Y)\right\}=0$ and

$$\frac{\partial \mathbb{E}\left[m_{\beta,1}\left\{X,A,Y;\beta_1^*,h(\theta)\right\}\right]}{\partial \theta} = \mathbb{E}\left\{\alpha_{\beta,1}(X,A,Y)S(X,A,Y)\right\}.$$

We rewrite $m_{\beta,1}\left\{X,A,Y;\beta_1^*,h(\theta)\right\}$ based on the simplified notations for the conditional expectations:

$$\begin{aligned}
& m_{\beta,1}\left\{X,A,Y;\beta_1^*,h(\theta)\right\} \\
= \;& A\{X-\lambda_2(X,\theta)\}\{X-\lambda_2(X,\theta)\}^{\mathrm{T}}\beta_1^* - A\{X-\lambda_2(X,\theta)\}\{Y-\lambda_1(X,\theta)\} \\
= \;& A\left\{XX^{\mathrm{T}} - X\lambda_2(X,\theta)^{\mathrm{T}} - \lambda_2(X,\theta)X^{\mathrm{T}} + \lambda_2(X,\theta)\lambda_2(X,\theta)^{\mathrm{T}}\right\}\beta_1^* \\
& -A\{XY - X\lambda_1(X,\theta) - \lambda_2(X,\theta)Y + \lambda_1(X,\theta)\lambda_2(X,\theta)\}.
\end{aligned}$$

Thus,

$$\frac{\partial \mathbb{E}\left[m_{\beta,1}\left\{X,A,Y;\beta_1^*,h(\theta)\right\}\right]}{\partial \theta} \tag{C.11}$$

$$= \left[\frac{\partial \mathbb{E}\left\{A\lambda_2(X,\theta)\lambda_2(X,\theta)^{\mathrm{T}}\right\}}{\partial \theta} - \frac{\partial \mathbb{E}\left\{AX\lambda_2(X,\theta)^{\mathrm{T}}\right\}}{\partial \theta}\right]\beta_1^* \tag{C.12}$$

$$+\frac{\partial \mathbb{E}\{AX\lambda_1(X,\theta)\}}{\partial \theta} - \frac{\partial \mathbb{E}\{A\lambda_1(X,\theta)\lambda_2(X,\theta)\}}{\partial \theta} \tag{C.13}$$

$$+\frac{\partial \mathbb{E}\{A\lambda_2(X,\theta)Y\}}{\partial \theta} - \frac{\partial \mathbb{E}\left\{A\lambda_2(X,\theta)X^{\mathrm{T}}\beta_1^*\right\}}{\partial \theta}. \tag{C.14}$$

Recall that we have orthogonality condition for $\lambda_2(X,\theta)=\mathbb{E}_\theta\{X \mid e^*(X), A=1\}$:

$$\mathbb{E}_\theta\left(h_2\{e^*(X)\}A[X-\mathbb{E}_\theta\{X \mid e^*(X), A=1\}]\right)=0.$$

Let $h_2\{e^*(X)\}=\lambda_2(X,\theta)^{\mathrm{T}}$ and take transpose,

$$\mathbb{E}_\theta\left\{AX\lambda_2(X,\theta)^{\mathrm{T}}\right\} = \mathbb{E}_\theta\left\{A\lambda_2(X,\theta)\lambda_2(X,\theta)^{\mathrm{T}}\right\}.$$

Take derivative on both sides,

$$\begin{aligned}\frac{\partial\mathbb{E}_\theta\left\{AX\lambda_2(X,\theta)^{\mathrm{T}}\right\}}{\partial\theta} &= \frac{\partial\mathbb{E}_\theta\left\{A\lambda_2(X,\theta)\lambda_2(X,\theta)^{\mathrm{T}}\right\}}{\partial\theta}\\ \frac{\partial\mathbb{E}\left\{AX\lambda_2(X,\theta)^{\mathrm{T}}\right\}}{\partial\theta}+\frac{\partial\mathbb{E}_\theta\left\{AX\lambda_2(X)^{\mathrm{T}}\right\}}{\partial\theta} &= \frac{\partial\mathbb{E}\left\{A\lambda_2(X,\theta)\lambda_2(X,\theta)^{\mathrm{T}}\right\}}{\partial\theta}\\ &\quad+\frac{\partial\mathbb{E}_\theta\left\{A\lambda_2(X,\theta)\lambda_2(X)^{\mathrm{T}}\right\}}{\partial\theta}.\end{aligned}$$

Thus,

$$(C.12)=\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[A\{X-\lambda_2(X)\}\lambda_2(X)^{\mathrm{T}}\beta_1^*\right].$$

Let $h_2\{e^*(X)\}=\lambda_1(X,\theta)$ and take transpose,

$$\mathbb{E}_\theta\{AX\lambda_1(X,\theta)\} = \mathbb{E}_\theta\{A\lambda_1(X,\theta)\lambda_2(X,\theta)\}.$$

Take derivative on both sides,

$$\begin{aligned}\frac{\partial\mathbb{E}_\theta\{AX\lambda_1(X,\theta)\}}{\partial\theta} &= \frac{\partial\mathbb{E}_\theta\{A\lambda_1(X,\theta)\lambda_2(X,\theta)\}}{\partial\theta}\\ \frac{\partial\mathbb{E}\{AX\lambda_1(X,\theta)\}}{\partial\theta}+\frac{\partial\mathbb{E}_\theta\{AX\lambda_1(X)\}}{\partial\theta} &= \frac{\partial\mathbb{E}\{A\lambda_1(X,\theta)\lambda_2(X,\theta)\}}{\partial\theta}+\frac{\partial\mathbb{E}_\theta\{A\lambda_1(X)\lambda_2(X)\}}{\partial\theta}.\end{aligned}$$

Thus,

$$(C.13)=\frac{\partial}{\partial\theta}\mathbb{E}_\theta[-A\{X-\lambda_2(X)\}\lambda_1(X)].$$

Lastly,

$$
\begin{aligned}
(C.14) &= \frac{\partial}{\partial\theta}\mathbb{E}\left\{A\lambda_2(X,\theta)\left(Y-X^{\mathrm{T}}\beta_1^*\right)\right\} \\
&= \frac{\partial}{\partial\theta}\mathbb{E}\left[\mathbb{E}\left\{A\lambda_2(X,\theta)\left(Y-X^{\mathrm{T}}\beta_1^*\right)\mid X,A\right\}\right] \\
&= \frac{\partial}{\partial\theta}\mathbb{E}\left[e(X)\lambda_2(X,\theta)\left\{\mathbb{E}(Y\mid X,A=1)-X^{\mathrm{T}}\beta_1^*\right\}\right] \\
&= \frac{\partial}{\partial\theta}\mathbb{E}\left(\lambda_2(X,\theta)\mathbb{E}\left[e(X)\left\{\mathbb{E}(Y\mid X,A=1)-X^{\mathrm{T}}\beta_1^*\right\}\mid e^*(X)\right]\right).
\end{aligned}
$$

Denote $\varphi\{e^*(X)\}=\mathbb{E}\left[e(X)\left\{\mathbb{E}(Y\mid X,A=1)-X^{\mathrm{T}}\beta_1^*\right\}\mid e^*(X)\right]$, then

$$
\begin{aligned}
(C.14) &= \mathbb{E}\left(\mathbb{E}_\theta\left[X\varphi\{e^*(X)\}\mid e^*(X),A=1\right]\right) \\
&= \mathbb{E}\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[X\varphi\{e^*(X)\}\mid e^*(X),A=1\right]\right) \\
&= \mathbb{E}\left(\mathbb{E}\left[X\varphi\{e^*(X)\}S\{X\mid e^*(X),A=1\}\mid e^*(X),A=1\right]\right) \\
&= \mathbb{E}\left(\mathbb{E}\left[X\varphi\{e^*(X)\}S\{X\mid e^*(X),A=1\}\cdot\frac{A}{\pi\{e^*(X)\}}\mid e^*(X),A\right]\right) \\
&= \mathbb{E}\left[X\varphi\{e^*(X)\}S\{X\mid e^*(X),A\}\cdot\frac{A}{\pi\{e^*(X)\}}\right] \\
&= \mathbb{E}\left([X-\mathbb{E}\{X\mid e^*(X),A=1\}]\varphi\{e^*(X)\}S\{X\mid e^*(X),A\}\cdot\frac{A}{\pi\{e^*(X)\}}\right) \\
&= \mathbb{E}\left([X-\mathbb{E}\{X\mid e^*(X),A=1\}]\varphi\{e^*(X)\}\cdot\frac{A}{\pi\{e^*(X)\}}\cdot S(X,A,Y)\right).
\end{aligned}
$$

The last line is because

$$
\begin{aligned}
\mathbb{E}\left([X-\mathbb{E}\{X\mid e^*(X),A=1\}]\varphi\{e^*(X)\}S\{Y\mid X,e^*(X),A\}\cdot\frac{A}{\pi\{e^*(X)\}}\right) &= 0 \\
\mathbb{E}\left([X-\mathbb{E}\{X\mid e^*(X),A=1\}]\varphi\{e^*(X)\}S\{e^*(X),A\}\cdot\frac{A}{\pi\{e^*(X)\}}\right) &= 0,
\end{aligned}
$$

and

$$
S(X,A,Y)=S\{e^*(X),A\}+S\{X\mid e^*(X),A\}+S\{Y\mid X,e^*(X),A\}.
$$

Thus,

$$\begin{aligned}\alpha_{\beta,1}(X,A,Y) &= A\{X-\lambda_2(X)\}\lambda_2(X)^{\mathrm{T}}\beta_1^* \\ &\quad -A\{X-\lambda_2(X)\}\lambda_1(X) \\ &\quad +\{X-\lambda_2(X)\}\varphi\{e^*(X)\}\cdot\frac{A}{\pi\{e^*(X)\}}.\end{aligned}$$

Meanwhile,

$$\begin{aligned}m_{\beta,1}\left(X,A,Y;\beta_1^*,h_0\right) &= -AX\left(Y-X^{\mathrm{T}}\beta_1^*\right) \\ &\quad -A\{X-\lambda_2(X)\}\lambda_2(X)^{\mathrm{T}}\beta_1^* \\ &\quad +A\{X-\lambda_2(X)\}\lambda_1(X) \\ &\quad +A\lambda_2(X)\left(Y-X^{\mathrm{T}}\beta_1^*\right).\end{aligned}$$

Then the influence function of $\hat{\beta}_1^*$ is

$$\psi_{\beta,1}(X,A,Y)=\mathbb{E}\left(Ar_{X,1}^*r_{X,1}^{*\mathrm{T}}\right)^{-1}\left[A\{X-\lambda_2(X)\}\left(Y-X^{\mathrm{T}}\beta_1^*-\frac{A\varphi\{e^*(X)\}}{\pi\{e^*(X)\}}\right)\right].$$

To sum up, the contribution from the nonparametric estimation is

$$\frac{A\{Y-\lambda_1(X)\}}{\pi\{e^*(X)\}}-\frac{A\{X-\lambda_2(X)\}^{\mathrm{T}}\beta_1^*}{\pi\{e^*(X)\}}+\mathbb{E}\left(r_{X,1}^*\right)^{\mathrm{T}}\psi_{\beta,1}(X,A,Y).$$

On the other hand, the influence funtion from (C.10) is

$$\mu_1\{X_i,e^*(X_i)\}-\mu_1=\lambda_1(X)+\{X-\lambda_2(X)\}^{\mathrm{T}}\beta_1^*-\mu_1.$$

In all, the influence function of $\hat{\mu}_1$ is

$$\begin{aligned}&\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}|_{\alpha=\alpha^*}\right]M_e^{-1}\psi_e(A,X;\alpha^*) \\ +\quad &\frac{A\{Y-\lambda_1(X)\}}{\pi\{e^*(X)\}}+\frac{\pi\{e^*(X)\}-A}{\pi\{e^*(X)\}}\{X-\lambda_2(X)\}^{\mathrm{T}}\beta_1^* \\ +\quad &\mathbb{E}\left(r_{X,1}^*\right)^{\mathrm{T}}\psi_{\beta,1}(X,A,Y)+\lambda_1(X)-\mu_1.\end{aligned}$$

Thus, the influence function of the doubly robust regression estimator is

$$
\begin{aligned}
& \left(\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_1\{X,e(X;\alpha)\}|_{\alpha=\alpha^*}\right]-\mathbb{E}\left[\frac{\partial}{\partial\alpha}\mu_0\{X,e(X;\alpha)\}|_{\alpha=\alpha^*}\right]\right)M_e^{-1}\psi_e(A,X;\alpha^*) \\
+ \quad & \frac{A\{Y-\mathbb{E}\{Y\mid e^*(X),A=1\}\}}{\pi\{e^*(X)\}}+\frac{\pi\{e^*(X)\}-A}{\pi\{e^*(X)\}}\{X-\mathbb{E}\{X\mid e^*(X),A=1\}\}^{\mathrm{T}}\beta_1^* \\
- \quad & \frac{(1-A)\{Y-\mathbb{E}\{Y\mid e^*(X),A=0\}\}}{1-\pi\{e^*(X)\}}+\frac{\pi\{e^*(X)\}-A}{1-\pi\{e^*(X)\}}\{X-\mathbb{E}\{X\mid e^*(X),A=0\}\}^{\mathrm{T}}\beta_0^* \\
+ \quad & \mathbb{E}\left(r_{X,1}^*\right)^{\mathrm{T}}\psi_{\beta,1}(X,A,Y)-\mathbb{E}\left(r_{X,0}^*\right)^{\mathrm{T}}\psi_{\beta,0}(X,A,Y). \\
+ \quad & \lambda_1(X)-\lambda_0(X)-\left(\mu_1-\mu_0\right)
\end{aligned}
$$

C.3 Lemmas

Lemma 4 (Pathwise derivarive; Newey, 1994) *Denote $S(X,A,Y)$ as the score function for the true distribution. Suppose that (i) the set of scores for regular paths is linear; (ii) for any $\varepsilon>0$ and measureable $s(X,A,Y)$ with $\mathbb{E}\{s(X,A,Y)\}=0$ and $\mathbb{E}\left\{s(X,A,Y)^2\right\}<\infty$ there is a regular path with score $S(X,A,Y)$ satisfying $\mathbb{E}\{|s(X,A,Y)-S(X,A,Y)|\}<\varepsilon$; (iii) $\hat{\mu}$ is asymptotically linear and regular. Then there is $d(X,A,Y)$ such that*

$$
\frac{\partial\mu(F_\theta)}{\partial\theta}=\mathbb{E}\{d(X,A,Y)S(X,A,Y)\},
$$

and $d(X,A,Y)$ is the influence function of $\hat{\mu}$. We call $d(X,A,Y)$ the pathwise derivative of $\mu(F)$.

The proof of the lemma is included in Newey (1994).

Lemma 5 (Pathwise derivarive for M-estimators; Newey, 1994) *Suppose $\hat{\beta}$ is the solution of the following M-estimating equation:*

$$
\sum_{i=1}^{n} m\left(X_i,A_i,Y_i;\beta,\hat{h}\right)=0,
$$

where $\hat{h}$ denotes a function depending on the parameter and the data, e.g., the nonparametric estimators of the conditional expectations. Denote β^* as the probability limit of $\hat{\beta}$. For a path $\{F_\theta\}$, let $h(\theta)=h(F_\theta)$. We also denote $h_0=h(F_0)$. Then by (3.10) in Newey (1994), the influence function of $\hat{\beta}$ is

$$
\psi_\beta(X,A,Y)=-M_\beta^{-1}\left\{m_\beta\left(X,A,Y;\beta,h_0\right)+\alpha_\beta(X,A,Y)\right\},
$$

where

$$M_\beta = \frac{\partial \mathbb{E}\left\{m_\beta\left(X, A, Y; \beta, h_0\right)\right\}}{\partial \beta}\,|_{\beta=\beta^*},$$

and $\alpha_\beta(X, A, Y)$ *satisfies* $\mathbb{E}\left\{\alpha_\beta(X, A, Y)\right\} = 0$ *and*

$$\frac{\partial \mathbb{E}\left[m_\beta\left\{X, A, Y; \beta^*, h(\theta)\right\}\right]}{\partial \theta} = \mathbb{E}\left\{\alpha_\beta(X, A, Y) S(X, A, Y)\right\}.$$

The proof of the lemma is included in Newey (1994).

APPENDIX

D

APPENDIX FOR "SENSITIVITY ANALYSIS IN CAUSAL INFERENCE WITH INFORMATIVE SAMPLING"

D.1 Proof of Theorem 8

The data generating process can be characterized by $F(x) = \mathbb{P}(X \leq x)$, $\omega_s(x) = \mathbb{P}(S = s \mid X = x)$, $\pi_a(x) = \mathbb{P}(A = a \mid S = 1, X = x)$, and $F_a\big(y \mid x\big) = \mathbb{P}\big(Y \leq y \mid A = a, S = 1, X = x\big) = \mathbb{P}(Y \leq y \mid A = a\ , X = x)$. Consider the distribution $\widetilde{\mathbb{P}}_\theta$ defined by the following parametric submodel:

$$
\begin{aligned}
d\widetilde{F}_{\theta}(x) &= dF(x)\{1+\epsilon h(x)\} \\
d\widetilde{F}_{a,\theta}\left(y \mid x\right) &= dF_{a}\left(y \mid x\right)\left\{1+\eta_{a} k_{a}\left(y,x\right)\right\} \\
\widetilde{\pi}_{a,\theta}(x) &= \frac{\left[\pi_{1}(x)\,exp\left\{\delta\, l(x)\right\}\right]^{a}\pi_{0}(x)^{1-a}}{\pi_{1}(x)\,exp\left\{\delta\, l(x)\right\}+\pi_{0}(x)} \\
\widetilde{\omega}_{s,\theta}(x) &= \frac{\left[\omega_{1}(x)\,exp\left\{\rho\, k(x)\right\}\right]^{s}\omega_{0}(x)^{1-s}}{\omega_{1}(x)\,exp\left\{\rho\, k(x)\right\}+\omega_{0}(x)},
\end{aligned}
$$

where $\theta=\left(\epsilon,\eta_{0},\eta_{1},\delta,\rho\right)$, $\mathbb{E}\{h(X)\}=\mathbb{E}\left\{k_{a}\left(y,x\right)\mid X,A=a\right\}=0$, and $l(x)$ and $k(x)$ are arbitary functions of x. The corresponding score functions can be derived by calculating partial derivatives of the observed data log likelihood $d\widetilde{F}_{\theta}(z)=d\widetilde{F}_{\theta}\left(x,s,a,y\right)$ and evaluating it at $\theta=0$:

$$
\begin{aligned}
\frac{\partial\, dlog\widetilde{F}_{\theta}(z)}{\partial\epsilon}|_{\theta=0} &= h(x) \\
\frac{\partial\, dlog\widetilde{F}_{\theta}(z)}{\partial\eta_{a}}|_{\theta=0} &= SI(A=a)\,k_{a}\left(y,x\right) \\
\frac{\partial\, dlog\widetilde{F}_{\theta}(z)}{\partial\delta}|_{\theta=0} &= S\{A-\pi_{1}(x)\}\,l(x) \\
\frac{\partial\, dlog\widetilde{F}_{\theta}(z)}{\partial\rho}|_{\theta=0} &= \{S-\omega_{1}(x)\}\,k(x).
\end{aligned}
$$

The target parameter under $\widetilde{P}_{\theta}$ is

$$
\begin{aligned}
\psi_a(\widetilde{P}) &= \int_x \int_y y\, d\widetilde{F}_{a,\theta}(y \mid x)\widetilde{\pi}_{a,\theta}(x)\widetilde{\omega}_{1,\theta}(x)\, d\widetilde{F}_\theta(x) \\
&+ \int_x \frac{\int_y y\, exp\{\gamma_a s_a(y)\}\, d\widetilde{F}_{a,\theta}(y \mid x)}{\int_y exp\{\gamma_a s_a(y)\}\, d\widetilde{F}_{a,\theta}(y \mid x)} \widetilde{\pi}_{1-a,\theta}(x)\widetilde{\omega}_{1,\theta}(x)\, d\widetilde{F}_\theta(x) \\
&+ \int_x y\, d\widetilde{F}_{a,\theta}(y \mid x)\widetilde{\pi}_{a,\theta}(x) \frac{exp(-\gamma_s a)}{\eta(x)} \widetilde{\omega}_{0,\theta}(x)\, d\widetilde{F}_\theta(x) \\
&+ \int_x \frac{\int_y y\, exp\{\gamma_a s_a(y)\}\, d\widetilde{F}_{a,\theta}(y \mid x)}{\int_y exp\{\gamma_a s_a(y)\}\, d\widetilde{F}_{a,\theta}(y \mid x)} \widetilde{\pi}_{1-a,\theta}(x) \\
&\frac{exp\{-\gamma_s(1-a)\}}{\eta(\gamma_s, X)} \widetilde{\omega}_{0,\theta}(x)\, d\widetilde{F}_\theta(x).
\end{aligned}
$$

Its partial derivatives are

$$
\begin{aligned}
\frac{\partial \psi_a(\widetilde{P})}{\partial \epsilon}|_{\theta=0} &= \int_x h(x)\left(\mu_a(y;x)\pi_a(x) + \frac{\mu_a[y\, exp\{\gamma_a s_a(y)\};x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]}\pi_{1-a}(x)\right)\omega_1(x) \\
&+ h(x)(\mu_a(y;x)\pi_a(x)exp(-\gamma_s a) + \frac{\mu_a[y\, exp\{\gamma_a s_a(y)\};x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]}\pi_{1-a}(x) \\
&exp\{-\gamma_s(1-a)\})\frac{\omega_0(x)}{\eta(\gamma_s, x)} dF(x)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \psi_a(\widetilde{P})}{\partial \eta_a}|_{\theta=0} &= \int_x [\mu_a\{y\, k_a(y,x);x\}\pi_a(x)\left\{\omega_1(x) + \frac{exp(-\gamma_s a)}{\eta(\gamma_s, x)}\omega_0(x)\right\} \\
&+ (\frac{\mu_a[y\, exp\{\gamma_a s_a(y)\}\, k_a(y,x);x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]} \\
&- \frac{\mu_a[exp\{\gamma_a s_a(y)\}\, k_a(y,x);x]\mu_a[y\, exp\{\gamma_a s_a(y)\};x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]^2}) \\
&\pi_{1-a}(x)\left\{\omega_1(x) + \frac{exp(-\gamma_s(1-a))}{\eta(\gamma_s, x)}\omega_0(x)\right\}] dF(x)
\end{aligned}
$$

$$\frac{\partial \psi_a(\widetilde{P})}{\partial \eta_{1-a}}|_{\theta=0} = 0$$

$$\begin{aligned}\frac{\partial \psi_a(\widetilde{P})}{\partial \delta}|_{\theta=0} &= \int_x (-1)^{a+1}\left(\mu_a(y;x)-\frac{\mu_a[y\,exp\{\gamma_a s_a(y)\};x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]}\right)\\ &\quad\left\{\omega_1(x)+\frac{exp(-\gamma_s)\omega_0(x)}{\eta(\gamma_s,x)^2}\right\}dF(x)\end{aligned}$$

$$\begin{aligned}\frac{\partial \psi_a(\widetilde{P})}{\partial \rho}|_{\theta=0} &= \int_x (\mu_a(y;x)\pi_a(x)+\frac{\mu_a[y\,exp\{\gamma_a s_a(y)\};x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]}\pi_{1-a}(x)\\ &\quad -\mu_a(y;x)\pi_a(x)\frac{exp(-\gamma_s a)}{\eta(\gamma_s,x)}\\ &\quad -\frac{\mu_a[y\,exp\{\gamma_a s_a(y)\};x]}{\mu_a[exp\{\gamma_a s_a(y)\};x]}\pi_{1-a}(x)\frac{exp\{-\gamma_s(1-a)\}}{\eta(\gamma_s,x)})\\ &\quad \omega_0(x)\omega_1(x)k(x)dF(x)\end{aligned}$$

Any mean zero observed data random variable can be expressed as:

$$d(O)=a(X)+\{S-\omega_1(X)\}b(X)+S\{A-\pi_1(X)\}c(X)+S^{[}a=0]1\sum I(A=a)d_a(Y;X),$$

where $\mathbb{E}\{a(X)\}=\mathbb{E}\{d_a(Y;X)\mid A=a,S=1,X\}=0$ and $b(X),c(X)$ are arbitrary functions of X. To find the efficient influence function, we need to find appropriate $a(X),b(X),c(X)$, $d_a(Y;X)$ such that

$$
\begin{aligned}
\frac{\partial \psi_a(\widetilde{P})}{\partial \epsilon}|_{\theta=0} &= \mathbb{E}\{a(X)h(X)\} \\
\frac{\partial \psi_a(\widetilde{P})}{\partial \eta_a}|_{\theta=0} &= \mathbb{E}\{SI(A=a)k_a(Y,X)d_a(Y,X)\} \\
\frac{\partial \psi_a(\widetilde{P})}{\partial \delta}|_{\theta=0} &= \mathbb{E}\big[S\{A-\pi_1(X)\}^2 l(X)c(X)\big] \\
\frac{\partial \psi_a(\widetilde{P})}{\partial \rho}|_{\theta=0} &= \mathbb{E}\big[\{S-\omega_1(x)\}^2 k(X)b(X)\big].
\end{aligned}
$$

By the law of iterated expectation, the right hand side of the last three equations can be rewritten as:

$$
\begin{aligned}
\mathbb{E}\{SI(A=a)k_a(Y,X)d_a(Y,X)\} &= \mathbb{E}[\mathbb{E}\{k_a(Y,X)d_a(Y,X)\,|\,A=a,S=1,X\} \\
&\quad \omega_1(X)\pi_a(X)] \\
\frac{\partial \psi_a(\widetilde{P})}{\partial \delta}|_{\theta=0} &= \mathbb{E}\{\omega_1(X)\pi_0(X)\pi_1(X)l(X)c(X)\} \\
\frac{\partial \psi_a(\widetilde{P})}{\partial \rho}|_{\theta=0} &= \mathbb{E}\{\omega_0(X)\omega_1(X)k(X)b(X)\}.
\end{aligned}
$$

Compared with the partial derivatives, it can be shown that

$$
\begin{aligned}
a(X) &= \left(\mu_a(Y;X)\pi_a(X)+\frac{\mu_a\big[Yexp\{\gamma_a s_a(Y)\};X\big]}{\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]}\pi_{1-a}(X)\right)\omega_1(X) \\
&\quad +(\mu_a(Y;X)\pi_a(X)exp\big(-\gamma_s a\big)+\frac{\mu_a\big[Yexp\{\gamma_a s_a(Y)\};X\big]}{\mu_a\big[exp\{\gamma_a s_a(Y)\};X\big]} \\
&\quad \pi_{1-a}(X)exp\{-\gamma_s(1-a)\})\frac{\omega_0(X)}{\eta(\gamma_s,X)}
\end{aligned}
$$

$$
\begin{aligned}
b(X) &= \mu_a(Y;X)\pi_a(X)+\frac{\mu_a\left[Yexp\left\{\gamma_a s_a(Y)\right\};X\right]}{\mu_a\left[exp\left\{\gamma_a s_a(Y)\right\};X\right]}\pi_{1-a}(X) \\
&\quad -\mu_a(Y;X)\pi_a(X)\frac{exp\left(-\gamma_s a\right)}{\eta\left(\gamma_s,X\right)} \\
&\quad -\frac{\mu_a\left[Yexp\left\{\gamma_a s_a(Y)\right\};X\right]}{\mu_a\left[exp\left\{\gamma_a s_a(Y)\right\};X\right]}\pi_{1-a}(X)\frac{exp\left\{-\gamma_s(1-a)\right\}}{\eta\left(\gamma_s,X\right)}
\end{aligned}
$$

$$
c(x) = (-1)^{a+1}\left(\mu_a(Y;X)-\frac{\mu_a\left[Yexp\left\{\gamma_a s_a(Y)\right\};X\right]}{\mu_a\left[exp\left\{\gamma_a s_a(Y)\right\};X\right]}\right)\left\{1+\frac{exp\left(-\gamma_s\right)}{\eta\left(\gamma_s,X\right)^2}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right\}
$$

$$
\begin{aligned}
d_a(Y;X) &= \left\{Y-\mu_a\{Y;X\}\right\}\left\{1+\frac{exp\left(-\gamma_s a\right)}{\eta\left(\gamma_s,X\right)}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right\} \\
&\quad +\left(\frac{Yexp\left\{\gamma_a s_a(Y)\right\}}{\mu_a\left[exp\left\{\gamma_a s_a(Y)\right\};X\right]}-\frac{exp\left\{\gamma_a s_a(Y)\right\}\mu_a\left[Yexp\left\{\gamma_a s_a(Y)\right\};X\right]}{\mu_a\left[exp\left\{\gamma_a s_a(Y)\right\};X\right]^2}\right) \\
&\quad \frac{\pi_{1-a}(X)}{\pi_a(X)}\left\{1+\frac{exp\left(-\gamma_s(1-a)\right)}{\eta\left(\gamma_s,X\right)}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right\}
\end{aligned}
$$

$$
d_{1-a}(Y;X)=0.
$$

Thus, the efficient influence function has been derived, and it can be further simplfied as follows:

$$
\begin{aligned}
\phi_a(P)(O) &= SI(A=a)(\{Y-\mu_a(Y;X)\}\left\{1+\frac{exp(-\gamma_s a)}{\eta(\gamma_s,X)}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right\} \\
&+\left(\frac{Yexp\{\gamma_a s_a(Y)\}}{\mu_a[exp\{\gamma_a s_a(Y)\};X]}-\frac{exp\{\gamma_a s_a(Y)\}\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]^2}\right) \\
&\cdot\frac{\pi_{1-a}(X)}{\pi_a(X)}\cdot\left[1+\frac{exp\{-\gamma_s(1-a)\}}{\eta(\gamma_s,X)}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right]) \\
&+S\left(\mu_a(Y;X)\pi_a(X)+\frac{\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]}\pi_{1-a}(X)\right)-\psi_a(P) \\
&+S(A-\pi_1(X))(-1)^{a+1}\left(\mu_a(Y;X)-\frac{\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]}\right) \\
&\left\{1+\frac{exp(-\gamma_s)}{\eta(\gamma_s,X)^2}\cdot\frac{\omega_0(X)}{\omega_1(X)}\right\} \\
&+(1-S)\{\mu_a(Y;X)\pi_a(X)\cdot\frac{exp(-\gamma_s a)}{\eta(\gamma_s,X)}+\frac{\mu_a[Yexp\{\gamma_a s_a(Y)\};X]}{\mu_a[exp\{\gamma_a s_a(Y)\};X]} \\
&\pi_{1-a}(X)\cdot\frac{exp\{-\gamma_s(1-a)\}}{\eta(\gamma_s,X)}\}.
\end{aligned}
$$

D.2 Proof of Theorem 9

By assumption (9),

$$
\begin{aligned}
m(X,A) &= \alpha(X)+\gamma_s A \\
&= \alpha(X)+\gamma_s e(X)+\gamma_s\{A-e(X)\} \\
&= m^*(X)+\widetilde{\gamma}_s\widetilde{R}_A(X).
\end{aligned}
$$

Then, since $\widetilde{R}_A(X)$ is orthogonal to $m^*(X)$ and has unit variance,

$$
\text{var}\{m(X,A)\} = \text{var}\left\{m^*(X)+\widetilde{\gamma}_s\widetilde{R}_A(X)\right\}=\text{var}\{m^*(X)\}+\widetilde{\gamma}_s^2.
$$

The variance of the latent variable Z_{s_1} is then

$$\text{var}\left(Z_{s_1}\right) = \text{var}\{m(X,A)+\epsilon_1\} = \text{var}\{m^*(X)\}+\widetilde{\gamma}_s^2+\frac{\pi^2}{3}.$$

Thus, the partial variance explained by X and A is

$$\begin{aligned}\rho_{X,A}^2 &= \frac{\text{var}\{m(X,A)\}}{\text{var}\left(Z_{s_1}\right)} \\ &= \frac{\text{var}\{m^*(X)\}+\widetilde{\gamma}_s^2}{\text{var}\{m^*(X)\}+\widetilde{\gamma}_s^2+\frac{\pi^2}{3}}.\end{aligned}$$

Similarly, the partial variance explained by X is

$$\rho_X^2 = \frac{\text{var}\{m(X)\}}{\text{var}\{m(X)\}+\pi^2/3}.$$

Finally, the partial variance explained by A given X is

$$\begin{aligned}\rho_{A|X}^2 &= \frac{\rho_{X,A}^2-\rho_X^2}{1-\rho_X^2} \\ &= 1-\frac{\text{var}\{m(X)\}+\pi^2/3}{\text{var}\{m^*(X)\}+\pi^2/3+\sigma_A^2\gamma_s^2}.\end{aligned}$$

Solve for γ_s, we get

$$\left|\gamma_s\right| = \frac{1}{\sigma_A}\sqrt{\frac{\text{var}\{m(X)\}+\pi^2/3}{1-\rho_{A|X}^2}-\text{var}\{m^*(X)\}-\pi^2/3}.$$