ABSTRACT

ZHAO, HONGHE. Advances in Matching Methods for Causal Inference with Multiple Treatments. (Under the direction of Shu Yang and Emily Hector).

Matching on the generalized propensity score (GPS) is an effective and widely used approach for estimation of the pairwise treatment effects in observational studies that involve more than two treatment levels. In this dissertation, we develop innovative methods that enhance the efficiency and robustness of the GPS matching estimator, while also providing justification for its applicability to estimate treatment effects on time-to-event outcome.

In observational studies, the GPS is typically unknown. In the first chapter, we propose the outcome-adjusted balance measure to perform model selection for the GPS. The primary goal of the balance measure is to identify the GPS model specification such that the resulting ATE estimator is consistent and efficient. Following recent empirical and theoretical evidence, we establish that the optimal GPS model should only include covariates related to the outcomes. Given a collection of candidate GPS models, the outcome-adjusted balance measure imputes all baseline covariates by matching on each candidate model, and selects the model that minimizes a weighted sum of absolute mean differences between the imputed and original values of the covariates. The weights are defined to leverage the covariate-outcome relationship, so that GPS models without optimal variable selection are penalized. Under appropriate assumptions, we show that the outcome-adjusted balance measure consistently selects the optimal GPS model, so that the resulting GPS matching estimator is asymptotically normal and efficient.

While weighting methods are popular for comparing the effects of multi-level treatment in observational studies, their performance can be unstable in the presence of extreme values of the generalized propensity score. Matching methods are more resistant to GPS outliers but bear the risk of GPS model misspecification. To reduce the dependence of GPS estimator on the GPS model, in the second chapter, we propose a double score matching (DSM) estimator of the pairwise average treatment effects based on the GPS and the generalized prognostic score (GPGS). The de-biased DSM estimator not only maintains the advantage of matching methods but also alleviates the model dependence problem due to its double robustness: it consistently estimates the true pairwise ATE if either the GPS or the GPGS is correctly specified.

There is a growing interest in using observational studies to estimate the effects of

treatments on survival or time-to-event outcomes. However, few standard approaches can adequately accommodate multiple treatment levels, which are common in observational comparative effectiveness research. In the third chapter, we propose GPS matching estimators for the pairwise marginal hazard ratios among multiple treatment groups. First, we use GPS matching to impute the missing potential outcome processes. Then, pairwise hazard ratio estimates are obtained by fitting a marginal Cox proportional hazard model on the imputed dataset. We establish the asymptotic distributions of the GPS estimators based on known and estimated GPS. We evaluate our approach in a simulation study and and a case study where we analyze the IQVIA electronic medical records data.

Advances in Matching Methods for Causal Inference with Multiple Treatments

by
Honghe Zhao

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:

Wenbin Lu	Brian Reich
Shu Yang Co-chair of Advisory Committee	Emily Hector Co-chair of Advisory Committee

BIOGRAPHY

Honghe Zhao was born in Weihai, China and moved to the US in 2011. He attended The Oakridge School in Arlington, Texas for high school from 2011 to 2014. He earned his Bachelor's degree in Mathematics from The University of Texas at Austin in 2018 and began his PhD program in Statistics at North Carolina State University the same year. He will complete his PhD study in August 2023.

ACKNOWLEDGEMENTS

I am deeply grateful to Dr. Shu Yang for her invaluable guidance and expertise in forming research questions and for elevating my understanding of matching methods and research to a higher level. I am also deeply grateful to Dr. Emily Hector for her encouragement and insightful feedback that broadened the scope of my understanding in various topics in statistics. Without them this dissertation would not have been possible.

I would like to thank Dr. Brian Reich and Dr. Wenbin Lu for taking time out of their busy schedules to invest in me and offer valuable feedback.

I would like to thank my mom and the rest of my family for their unconditional support. Without them, I would not be where I am now, and none of these works would have been possible. I would also like to thank Nicholas Larsen for being a great friend and providing a listening ear during challenging times.

I would like to thank Yanning for her unwavering support, care, patience, and for her companionship.

TABLE OF CONTENTS

List of Tables . . . . . **vi**

List of Figures . . . . . **viii**

Chapter 1 INTRODUCTION **1**
1.1 Background 1
1.1.1 Causal inference with multiple treatments 2
1.1.2 Matching estimators 5
1.2 Challenges and Contributions 7

Chapter 2 Outcome-Adjusted Balance Measure for Generalized Propensity Score Model Selection **10**
2.1 Introduction 10
2.2 Background 13
2.2.1 Setup 13
2.2.2 Matching on the generalized propensity score 14
2.3 Model Selection 15
2.3.1 Imputing covariates via GPS matching 15
2.3.2 Types of covariates 17
2.3.3 Outcome-adjusted balance measure 18
2.4 Theoretical Properties 20
2.5 A Simulation Study 23
2.6 Applications 28
2.7 Discussion 30

Chapter 3 Double Score Matching in Observational Studies with Multilevel Treatments **32**
3.1 Introduction 32
3.2 Background 34
3.2.1 Setup, notation, and assumptions 34
3.2.2 Matching on the generalized propensity score 36
3.3 New Double Score Matching Estimator 38
3.4 Main Results 41
3.4.1 Asymptotic results 41
3.4.2 Resampling-based variance estimation 44
3.5 A Simulation Study 46
3.6 An Application 49
3.7 Discussion 53

Chapter 4 Propensity Score Matching for Estimation of Pairwise Marginal Hazard Ratios **55**

4.1 Notation, Model, and Assumptions . . . 57
4.1.1 Data structure . . . 57
4.1.2 Causal proportional hazard model . . . 58
4.1.3 Identification assumptions . . . 59
4.2 Methodology . . . 60
4.2.1 Matching . . . 60
4.2.2 Estimating equations . . . 61
4.2.3 Asymptotic results with known GPS . . . 63
4.2.4 Asymptotic results with estimated GPS . . . 64
4.2.5 Estimation of asymptotic variance . . . 65
4.3 Simulations . . . 66
4.4 Real Data Analysis . . . 71
4.5 Discussion and Future Studies . . . 72

Chapter 5 Discussion and Future Research . . . 74

References . . . 77

APPENDICES . . . 86
Appendix A Supplementary Material for Chapter 2 . . . 87
A.1 Notation and Assumptions . . . 87
A.2 Proof of Lemma 2 . . . 89
A.3 Proof of Theorem 1 . . . 91
A.4 Definition of $c(w)$ & Variance Estimation . . . 97
A.5 Steps to Perform Model Selection . . . 99
A.6 Existing Balance Measures . . . 99
A.7 Remaining Scenarios of the Simulation Study . . . 100
A.8 Nursing Home Data Codebook . . . 104
A.9 Application: Tutoring Data Analysis . . . 106
A.10 Simulation Result Details . . . 108
Appendix B Supplementary Material for Chapter 3 . . . 111
B.1 Proof of Theorem 3 . . . 112
B.2 Le Cam's Third Lemma . . . 114
B.3 Proof of Theorem 4 . . . 115
B.4 Additional Simulation Results . . . 118
Appendix C Supplementary Material for Chapter 4 . . . 126
C.1 Regularity Conditions and Lemmas . . . 126
C.2 Proof of Asymptotic Unbiasedness of the Partial Score Function . . . 128
C.3 Proof of Theorem 5 . . . 131
C.4 Proof of Theorem 6 . . . 134
C.5 Additional Simulations . . . 139
C.5.1 An algorithm that simulates the potential survival times . . . 139
C.5.2 Additional results when the causal effects are present . . . 140

LIST OF TABLES

Table 2.1 Coverage rates of asymptotic 95% confidence intervals using the GPSM point estimator and Abadie and Imbens (2016) variance estimator with GPS models selected by competing balance measures; simulation study 27

Table 2.2 The proposed $\texttt{OABM}_{\texttt{BCor}}$ evaluated at each treatment level, for each posited model; bold number indicates the lowest $\texttt{OABM}_{\texttt{BCor}}$ value for that treatment; nursing home data 30

Table 2.3 GPSM estimates of the pairwise ATEs between nursing home 1 and all other nursing homes; GPSM-$\texttt{OABM}_{\texttt{BCor}}$ uses $\texttt{OABM}_{\texttt{BCor}}$ to select GPS model among the three posited models; GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$ uses a multinomial logit model for the GPS specification that involves linear terms of IVs and confounders; standard errors are displayed in parenthesis; * indicates significance at the 0.05 level; nursing home data 30

Table 3.1 Matching scheme to impute missing potential outcomes when $T = 3$; for estimating $\mathbb{E}\{Y(w)\}$ at each level, the average of the imputed column is equal to the average of the weighted column. 38

Table 3.2 Simulation results based on 1000 Monte Carlo simulated datasets for the coverage properties for the proposed double score matching estimators of the average treatment effects under four scenarios for the generalized propensity score (GPS) and generalized prognostic score (GPGS) models: empirical coverage rate and (empirical coverage rate $\pm$ 2$\times$Monte Carlo standard error) 50

Table 3.3 Results of fitting a linear regression of all first-order terms of the covariates for the generalized prognostic score of each tutoring service. 50

Table 3.4 Results of fitting two different GPS models for the three tutoring levels. 51

Table 3.5 Estimated ATEs of tutoring services and 95% Wald confidence intervals 53

Table 4.1 GPS matching for imputation of the missing potential outcomes when the number of treatment levels is 3 and the number of matches is 2; the matching index set $\mathcal{J}_2\{1, e_1(\boldsymbol{X}_n)\}$, for instance, denotes the set containing the indices of 2 subjects who received treatment level $\omega = 1$ and whose generalized propensity scores evaluated at 1 are closest to $e_1(\boldsymbol{X}_n)$, the GPS of n-th unit evaluated at $\omega = 1$; the full imputed dataset can be equivalently viewed as a weighted dataset, where $k_i(W_i)$ is the number of times individual i is used as a match. 61

Table 4.2 IQVIA EMR data analysis results. Pem: Pemetrexed; CP: Carboplatin + paclitaxel; Erl: Erlotinib; Doc: Docetaxel; Gem: Gemcitabine. "Pem vs CP", for instance, is the hazard ratio CP/Pem. Thus, a value smaller than one implies that receiving Carboplatin + paclitaxel results in better survival than Pemetrexed. 95% confidence intervals are calculated using the same methods as described in the simulation study section. 73

Table A.1 Coverage rates of asymptotic 95% confidence intervals under scenarios with weak IVs 104

Table A.2 The proposed $\texttt{OABM}_{\texttt{BCor}}$ evaluated at each treatment level, for each posited model; bold number indicates the lowest $\texttt{OABM}_{\texttt{BCor}}$ value for that treatment; tutoring data 107

Table A.3 GPSM estimates of the pairwise ATEs among three types of tutoring services; GPSM-$\texttt{OABM}_{\texttt{BCor}}$ uses $\texttt{OABM}_{\texttt{BCor}}$ to select GPS model among the three posited models; GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$ uses a multinomial logit model for the GPS specification that involves linear terms of IVs and confounders; standard errors are displayed in parenthesis; * indicates significance at the 0.05 level; tutoring data 108

Table A.4 Simulation results of the MSE of GPSM estimator using competing balance measures for GPS model selection when the outcome model is nonlinear in covariates 109

Table A.5 Simulation results of the MSE of GPSM estimator using competing balance measures for GPS model selection when the outcome model is nonlinear in covariates 110

Table B.1 Simulation results based on 1000 Monte Carlo simulated datasets for the coverage rates (CR) and standard errors (SE) for the proposed double score matching estimators of the average treatment effects under four scenarios with (P1) for the generalized propensity score (GPS) and prognostic score (GPGS) models. 122

Table B.2 Simulation results based on 1000 Monte Carlo simulated datasets for the coverage rates (CR) and standard errors (SE) for the proposed double score matching estimators of the average treatment effects under four scenarios with (P2) for the generalized propensity score (GPS) and prognostic score (GPGS) models. 125

LIST OF FIGURES

Figure 2.1 A DAG that illustrates the relationships among four types of covariates, treatment assignment, and outcome in a simple scenario. 17

Figure 2.2 Box plots of GPSM estimates. Numeric MSEs for $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$, $\mathtt{OABM}_{\mathtt{BCor}}$ and $\mathtt{OABM}_{\mathtt{OLS}}$ are explicitly shown above their corresponding box plots. True pairwise treatment effects are denoted by the horizontal dotted lines. 25

Figure 2.3 Proportion of model selection by six balance measures. 26

Figure 3.1 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios under (P1) for the generalized propensity score (GPS) and generalized prognostic score (GPGS). 48

Figure 3.2 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios under (P2) for the generalized propensity score (GPS) and generalized prognostic score (GPGS). 49

Figure 3.3 Densities of fitted generalized propensity scores based on the two models. 51

Figure 3.4 Standardized difference in means based on the two models. 52

Figure 4.1 Simulation results when the GPS model is *correctly* specified. . . . 69

Figure 4.2 Simulation results when the GPS model is *incorrectly* specified. . 70

Figure A.1 Box plots of 1000 generalized propensity score matching estimates under scenarios with weak IVs. The 6 benchmark GPS models are greyed out on the x-axis. Numeric MSEs for $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$, $\mathtt{OABM}_{\mathtt{OLS}}$, and $\mathtt{OABM}_{\mathtt{BCor}}$ are explicitly shown above their corresponding box plots. True pairwise treatment effects are denoted by the horizontal dotted lines. 101

Figure A.2 Proportion of model selection over 1000 simulations under scenarios with weak IVs. The optimal benchmark model $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$ is colored in red. 103

Figure B.1 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effect $E\{Y(2) - Y(1)\}$ under four scenarios under (P1) for the generalized propensity score (GPS) and generalized prognostic score (GPGS) for different sample sizes $N_w = 100, 200, 300$ for $w = 1, 2, 3$. 119

Figure B.2 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P1) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = 100, N_2 = 200$ and $N_3 = 300$. 120

Figure B.3 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P1) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = N_2 = N_3 = 300$, when $\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$. 121
Figure B.4 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P2) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = 100, N_2 = 200$ and $N_3 = 300$. 123
Figure B.5 Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P2) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = N_2 = N_3 = 300$, when $\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$. 124

Figure C.1 Simulation results when the GPS model is *correctly* specified and treatment effect is present. 141
Figure C.2 Simulation results when the GPS model is *correctly* specified and treatment effect is present. 142

CHAPTER

1

INTRODUCTION

1.1 Background

Randomized controlled trials (RCTs) are widely considered the gold standard for estimating treatment effects. This is because by design, treatment randomization ensures that treatment groups are comparable, reducing the risk of selection bias. However, practical limitations such as resource and ethical constraints often make it difficult to conduct RCTs (Sanson-Fisher et al. 2007). In such cases, the rich information contained in observational data could, under appropriate assumptions, still be leveraged to draw causal conclusions. In recent decades, a significant amount of statistical research has been dedicated to developing causal inference methods for observational data.

In observational studies, the estimation of causal effects is fraught with confounding bias that arises from systematic differences between treatment groups (Joffe and Rosenbaum 1999). Under the assumptions of unconfoundedness and overlap, methods that correct for this bias can be broadly classified into two categories. The outcome regression approach models the multiple regression of the outcome on the treatment and measured potential confounders. However, this approach relies on correctly specifying the outcome regression

model. An alternative approach is based on modeling the propensity score, defined as the probability of receiving treatment given the confounders. The treatment effect can then be estimated by matching (Dehejia and Wahba 2002), weighting (Lunceford and Davidian 2004), stratification (Rosenbaum and Rubin 1983), or regression (Vansteelandt and Daniel 2014) on the estimated propensity scores. Propensity score methods are motivated by Rosenbaum and Rubin (1983), who showed that propensity scores have a balancing property, such that the conditional distribution of the potential confounders given the balancing scores are the same for treated and control groups. While propensity score methods provide some protection against misspecification of the outcome model, the propensity score model is still required to be correctly specified.

The parametric methods mentioned above primarily concern settings where treatment assignment is binary and with continuous outcomes. In recent years, there has been a growing interest in extending these approaches to evaluate treatment effect when there are more than two categorical treatment levels (Lopez and Gutman 2017), which is also known in the literature as "multiple treatments" or "multi-level treatments." The theoretical work of Imbens (2000) and Imai and Van Dyk (2004) introduced the generalized propensity score (GPS), which extends the propensity score framework for binary treatments to multiple treatments. Subsequently, several propensity score methods, such as matching (Yang et al. 2016), weighting (Imbens 2000), and subclassification (Yang et al. 2016), have been reformulated to accommodate multiple treatments. They have valuable applications in new priority areas of health services research, particularly comparative effectiveness research, which seeks to compare the effects of multiple therapies.

The focus of this dissertation is on developing new tools for improving the GPS matching estimator. In the following sections, we provide background on commonly used approaches for making causal inference under the multiple treatment setting, as well as a brief introduction to matching estimators.

1.1.1 Causal inference with multiple treatments

In this section, we provide a brief overview of the approaches to causal inference in observational studies with multiple treatments, which work under the assumption of unconfoundedness and overlap. We primarily focus on the regression and propensity score based methods, as some of them will appear as comparisons to the proposed methods in this dissertation.

A seemingly viable approach to evaluate treatment effects between multiple treatment

levels is to conduct a series of binary comparisons (SBCs) (Lechner 2001). This approach works by grouping subjects into separate sub-populations, each with two treatments, and then applying approaches designed for binary treatment. Although often used in the empirical literature, the main issue is that each pairwise treatment effect from SBC generalizes only to subjects eligible for that specific pair of treatments, as opposed to those eligible for all treatments (Yang et al. 2016). Lopez and Gutman (2017) illustrated in a few examples that using SBCs could lead to non-transitive causal estimates, increased bias, and ambiguity about which treatment is optimal.

The traditional outcome regression approach, also known as regression adjustment, directly models the relationship between the outcome and pretreatment covariates by treatment groups (Linden et al. 2016; Imbens and Rubin 2015). Under the unconfoundedness assumption, the pairwise ATEs can be identified by positing a parametric regression model for the conditional mean of outcome given the covariates under each treatment group. Estimates based on these regression models can be used to impute the missing potential outcomes. Consistency of the outcome regression estimator heavily depends on the correct specification of these regression models. If no prior knowledge is available, such a specification may be difficult, and may become increasingly difficult if the number of covariates is large. In addition, when the covariate distributions between treatment groups are far apart, this approach relies on extrapolation (Imbens and Rubin 2015).

Propensity score weighting, also known as inverse probability of treatment weighting, or inverse probability weighting (IPW), attempts to obtain an unbiased estimator for treatment effects similar to how the inverse of the selection probability adjusts for sampling bias (Horvitz and Thompson 1952). In a common weighting scheme, units in one group are weighted by by their estimated inverse probability of being in that group. The framework of propensity score weighting makes for a natural extension to the multiple treatment setting (Imbens 2000). The large sample distributions of the weighting-based estimators can be characterized using the theory of M-estimation, which yields estimated standard errors that accounts for the uncertainty associated with the first step estimation of the propensity score. However, a major challenge with IPW is that subjects with low generalized propensity scores that are close to zero may lead to extreme weights, resulting in high instability in estimation (Kang and Schafer 2007). This issue gets more pronounced as the number of treatments increases (Yang et al. 2016).

An important extension of the IPW is augmented inverse probability weighting (AIPW), where the IPW estimator is augmented using predictions from an outcome regression model. To implement AIPW method in a multiple treatment setting, one can first obtain

the estimated GPS, possibly from a multinomial logistic regression model, and then the predicted outcomes for each treatment group from outcome models that describe the conditional expectation of the outcome variable given measured covariates and treatment status. The resulting estimator is known as having a double robustness (DR) property such that the estimator remains consistent as long as either the propensity score model or the outcome model is correctly specified. AIPW estimator is asymptotically efficient within a broad class of estimators that includes the IPW estimator Robins et al. (1994). Lunceford and Davidian (2004) reviewed the theoretical properties of IPW, AIPW, and several other propensity score weighting estimators in the context of two treatments and continuous outcome.

Propensity score matching (PSM) is the most widely used approach in the binary treatment setting (Pearl 2010). There are a variety of different matching algorithms. One particular version of the PSM, where matching is done with replacement and with a fixed number of matches per unit, targets inference for the overall population. PSM imputes the missing potential outcomes by finding the closest subject in the control group for each individual in the treatment group and vice versa. A main advantage of PSM is that, by matching subjects based on a scalar propensity score instead of a vector of covariates, confounding can be minimized. The theoretical properties of this version of the PSM estimator were established by a series of papers by Abadie and Imbens (2006, 2011, 2012, 2016). Despite its popularity in the binary treatment case, extending the PSM framework to the multiple treatment setting was not straightforward. This is because when there are more than two treatment levels, there is no scalar function of the covariates such that strong unconfoundedness holds. As a result, it was widely believed that the main advantage of PSM does not seem to carry over to multiple treatment levels (Imbens 2000; Lechner 2001; Rassen et al. 2013). However, by assuming weak unconfoundedness, Yang et al. (2016) showed that PSM can be extended to the multi-level treatment setting, without giving up the dimension reduction property of the propensity score. We provide an overview of the matching estimators in the next section.

Besides the simple parametric modelling approaches, there is also an extensive literature on using machine learning methods to capture potential nonlinearities and higher-order interactions. Some of these approaches have been compared under the binary outcome and multiple treatment settings (Hu et al. 2020). The relative gain by using such flexible methods depends on the sample size, the number of predictors, and the true structure of the underlying propensity score or outcome models. A non-parametric modeling approach is Bayesian additive regression trees (BART), which allows for very flexible functional form

between the covariates and outcome (Hill 2011). Another doubly robust approach, targeted maximum likelihood estimation (TMLE) (Schuler and Rose 2017), combines outcome and propensity score estimation using the Super Learner (Van der Laan et al. 2007), and a targeting step to optimize the parameter of interest with respect to bias/variance.

1.1.2 Matching estimators

A variety of different matching algorithms have been proposed in the literature, such as nearest neighbor matching, kernal matching, and full matching, etc. For a comprehensive review, see Stuart (2010). A simple and widely used nearest neighbor matching procedure attempts to match each unit to a small number of units with similar characteristics in the opposite treatment arm. In this section, we provide a brief overview of this matching procedure and refer to it as "matching estimators" in the remainder of this dissertation.

Although exact matching is in many ways the ideal, the primary difficulty is that it becomes increasingly difficult to obtain perfect matches when the covariates are high dimensional. To overcome the curse of dimensionality, Rosenbaum and Rubin (1983) showed that, under the unconfoundedness assumption, adjusting for the propensity score only is sufficient to remove all confounding, which motivates the use of propensity score matching estimators. Using the propensity score for matching reduces the dimensionality of the procedure by summarizing the information contained in the covariates in a single variable.

Propensity score matching enjoys several attractive features. First, it creates a clear separation between the design and the analysis (Ho et al. 2007). As the design stage does not involve any outcome data information, a careful design can improve the objectiveness of the outcome data analysis (Rubin 2007, 2008). Additionally, matching has been shown to be robust to propensity score model misspecification and to the presence of extreme values of the estimated propensity score (Waernbaum 2012; Greifer and Stuart 2021). Furthermore, matching can be viewed as hot deck imputation in the missing data literature (Little and Rubin 2014). As a result, it offers a natural way to construct valid estimators for general parameters of the overall population, such as quantiles (Ford 1983).

The asymptotic properties of the matching estimators of average treatment effects have been thoroughly investigated. Abadie and Imbens (2006) showed that matching estimators are in general not $N^{1/2}$-consistent and asymptotically normal unless the matching variables contain at most one continuously distributed variable. Abadie and Imbens (2011) proposed a bias correction that renders matching estimators $N^{1/2}$-consistent and asymptotically

normal irrespective of the number of covariates. This technique combines some of the advantages and disadvantages of both matching and regression estimators. Given the insight that matching estimators have a martingale representation (Abadie and Imbens 2012), Abadie and Imbens (2016) derived large sample approximations to the distribution of propensity score matching estimators while taking into account the first step estimation of the propensity score. They showed that matching on the estimated propensity score is more efficient than matching on the true propensity score in large samples. To make the matching methods accessible to practitioners, Imbens (2015) used three examples to demonstrate practical implementations from the theoretical literature, and provided detailed recommendations on how the procedures should be performed.

Existing inference methods for the matching estimators have been based on the asymptotic standard error or the resampling-bootstrap procedure. Inspired by the large sample distribution of the PSM estimator, Abadie and Imbens (2016) proposed an estimator of the asymptotic variance, which accounts for estimation of the propensity score. Abadie and Imbens (2008) provided an example which shows that the standard naive bootstrap fails to provide an asymptotically valid standard error and quantiles for a matching estimator. They recommended using the asymptotic standard error derived in Abadie and Imbens (2006) or subsampling (Politis and Romano1994) for inference. Otsu and Rai (2017) proposed a weighted bootstrap approach for matching directly on the covariates, which shows comparable performance to the asymptotic variance estimator by Abadie and Imbens (2006) in simulations. However, their consistency result does not extend to propensity score matching because conditioning on both treatments and covariates precludes taking into account the effect of the estimation of propensity scores. Adusumilli (2018) proposed a bootstrap procedure by resampling the potential errors as a pair, while also re-assigning new values for the treatments using the estimated propensity score.

While PSM is the most popular matching estimator, matching procedures based on other balancing scores have also been developed. The prognostic score, which summarizes covariate correlations with the outcomes, is also a balancing score. Prognostic score matching (PGM) is proposed by Hansen (2008) as a alternative to PSM. Unlike PSM that seeks to balance covariate distributions between treatment levels, PGM attempts to directly balance the potential outcome distributions. Wyss et al. (2015) used simulations and an empirical example to illustrate the superior efficiency of PGM compared to PSM when the propensity scored distributions are separated. Hansen (2006) was first to suggest the idea of matching jointly on the prognostic score and propensity score, or double score

matching (DSM). Leacy and Stuart (2014) later demonstrated the double robustness of this procedure in a simulation study. Antonelli et al. (2018) formally established the double robustness and convergence of the DSM estimator. Yang and Zhang (2023) derived the large sample distributions for the average and quantile treatment effects. The best practices of DSM estimator were investigated in Zhang et al. (2021).

1.2 Challenges and Contributions

Like other propensity score-based methods, the GPS matching estimator relies on correctly specifying the GPS model. However, existing model selection and variable selection techniques are suboptimal for GPS model selection, as their model evaluation criteria are based on the prediction accuracy of the treatment assignment, rather than the estimation performance of the matching estimator. Similarly, existing balance measures such as absolute mean difference, Mahalanobis distance, or Kolmogorov-Smirnov distance often fail to prioritize balance of the prognostically important covariates. This misalignment in variable/model selection goal may lead to substantial efficiency loss of the propensity score methods, including the GPS matching estimator.

In Chapter 2, we propose a balance measure that serves as a tool for model selection of the generalized propensity score. Empirical evidence suggests that the GPS matching estimator is most efficient when the GPS model includes covariates that are strong outcome predictors in the GPS model, while excluding covariates that are uncorrelated with the outcome. Thus, we propose a weighted balance measure that takes into account the outcome-covariate relationship, such that minimization of this measure yields the optimal GPS model. The weighting scheme encourages inclusion of outcome predictive covariates in the GPS model, while penalizing inclusion of covariates that are not correlated with the outcome. The weights could be parametrically or non-parametrically estimated depending on the knowledge of the outcome model. We demonstrate that our proposed balance measure achieves model selection consistency. That is, given a set of candidate models, it is guaranteed to select the GPS model that includes the optimal set of covariates in large samples. We show in a simulation study that, compared to other existing balance measures, using the proposed method for GPS model selection leads to the GPS estimator with the smallest mean squared error. The proposed method is used to analyze the MDA nursing home data in order to compare the quality of several nursing homes in Massachusetts.

An alternative approach that offers an additional layer of protection against misspec-

ification of the GPS is through the property of double robustness. An estimator with the double robustness property will consistently estimate the treatment effect as long as either the outcome model or the propensity score model is correctly specified. Augmented inverse probability estimator is perhaps the most well known estimator that is doubly robust for estimation of the ATE (Bang and Robins 2005). However, as the number of treatment level increases, the generalized propensity scores are more likely to be close to zero or one. This negatively affects the stability of the AIPW estimates (Kang and Schafer 2007). Matching methods, on the other hand, are more robust to practical violation of overlap assumption as it avoids inverting the GPS. While doubly robust matching estimators, also known as double score matching, have been proposed in the literature, its large sample distributions and performance evaluation have been limited to the binary treatment setting (Antonelli et al. 2018; Yang and Zhang 2023).

To address this limitation, in Chapter 3, we extend the framework of double score matching estimator of the average treatment effect to the multiple treatment setting. We first extend the notion of the prognostic score to multiple treatment levels by introducing the generalized prognostic score. We estimate the mean potential outcome for each treatment level separately. The key insight is that, to remove confounding while maintaining double robustness for estimating the mean potential outcome at each treatment level, matching on the generalized propensity score and generalize prognostic score for that treatment level is sufficient. However, the matching variable is two dimensional and the conditional bias does not vanish at a fast enough rate, preventing the estimator from being $N^{1/2}$-consistent. To circumvent this, we propose a bias-corrected double score matching estimator, where we subtract the double score matching estimator by a non-parametrically estimated conditional bias term. In addition, we propose a resampling-based variance procedure that accounts for variability induced by estimation of the two scores as well as matching. In a simulation study, we demonstrate the double robustness of the the debiased double score matching and show that it is preferable to AIPW when covariate distributions between treatment groups are strongly separated. The proposed approach is applied to the analysis of the tutoring data, which compares the effectivenesess of different tutoring services.

Most propensity score methods focus on estimation of the average treatment effect when outcome is continuous. For estimation of treatment effect on time-to-event outcome, standard techniques often combine propensity score methods with survival models. One of the most popular techniques for analyzing survival data is to use the Cox proportional hazard model (Cox 1972). Matching on the propensity score has shown advantage to

weighting, and has been empirically evaluated to be effective for analyzing survival data when used in combination with the Cox model (Austin 2013). Tang et al. (2019) extended the theoretical framework of using PSM for estimating the marginal hazard ratio. However, when there are multiple treatment arms, a theoretical framework for estimating pairwise marginal hazard ratios after propensity score matching is lacking.

In Chapter 4, we extend the PSM framework to estimate the pairwise marginal hazard ratios among all treatment levels. We begin by using generalized propensity score matching to impute the missing potential outcome processes. The resulting estimator is then obtained by solving a weighted Cox partial score equation, with weights corresponding to the number of times each unit is used as a match for other treatment levels. We show that this weighted Cox partial score function is asymptotically unbiased for zero. We derive the large sample distributions of the matching estimator based on the true and estimated generalized propensity scores. By comparing the asymptotic variances, we observe that matching on the estimated generalized propensity score leads to a more efficient treatment effect estimator than matching on the true GPS, which is parallel to the conclusion for estimation of the ATE. Following Abadie and Imbens (2016), we propose an asymptotic variance estimator that accounts for estimation of the generalized propensity score. We demonstrate in a simulation study that the proposed estimators of marginal hazard ratios are effective at minimizing bias, robust to GPS model misspecification and less variable than GPS weighting. We apply the GPS matching estimator for analyzing the IQVIA lung cancer electronic records data.

CHAPTER

2

OUTCOME-ADJUSTED BALANCE MEASURE FOR GENERALIZED PROPENSITY SCORE MODEL SELECTION

2.1 Introduction

Estimating the causal effects from observational data with more than two treatment levels has become an increasingly important goal in socioeconomic and biomedical research. Examples of observational studies with multiple treatment levels are ubiquitous. In public policy, the National Antidrug Media Campaign was implemented nationwide in the United States, with teens receiving varying degrees of exposure to the media campaign (Zanutto et al. 2005). In medicine, the REFLECTIONS was a 12-month prospective study that involved three fibromyalgia medication cohorts from 58 outpatient sites in the United States and Porto Rico (Yang et al. 2016). In health care, the Minimum Data Set is part

of a federally mandated process for clinical assessment of multiple Medicare or Medicaid certified nursing homes (Scotina and Gutman 2019). To deal with confounders, methods that weight (McCaffrey et al. 2013; Li and Li 2019), stratify (Zanutto et al. 2005; Yang et al. 2016), and match (Yang et al. 2016; Scotina and Gutman 2019) the sample based on the generalized propensity score (GPS) (Imbens 2000) have been proposed. While the popular AIPW estimator enjoys nice properties including semiparametric efficiency and double robustness, inversely weighting by extreme values of the GPS can still jeopardize its performance. In practice, it is common for the GPS to take on extreme values, and the instability issue with weighting methods can become more pronounced as the number of treatments increases. The matching estimator is an attractive alternative due to its resistance to those extreme probabilities (Frölich 2004). The idea of matching is also intuitively appealing as it seeks to replicate the ideal randomized experiment (Rubin 2006; Stuart 2010). GPS matching has been widely implemented in practice (Bennett et al. 2020; Scotina et al. 2020; Brown et al. 2020). However, as with all GPS based methods, GPS matching requires carefully modeling the generalized propensity score, which is typically unknown in practice. Given multiple postulated models, little work has been done to develop GPS model selection strategies, especially for improving the efficiency of the matching estimator.

When treatment is binary, considerable progress has been made towards developing modeling strategies for the propensity score (PS) (Rosenbaum and Rubin 1983). The essential purpose a PS model specification serves is in assisting the estimation and inference of the ATE(s) rather than in explaining or predicting treatment assignment. Standard diagnostics for model prediction performance should be avoided as they often fail to suggest PS models that provide unbiased and efficient estimation of the ATE (Westreich et al. 2011). This is supported by a large thread of empirical evidence, which surprisingly suggests that including instrumental variables (covariates that are part of the true PS model) in the PS model specification inflates variance of the resulting ATE estimators (Brookhart et al. 2006; Austin et al. 2007; Myers et al. 2011; Pearl 2011; Yang et al. 2020). Additionally, including precision variables (covariates only related to outcomes but not treatment) in the PS leads to improved efficiency for ATE estimation (Brookhart et al. 2006; Patrick et al. 2011). Building upon the work of Lunceford and Davidian (2004) and Hahn (2004), Tang et al. (2022) theoretically justify that both excluding instrumental variables and including precision variables in the PS will help improve asymptotic efficiency in the Horvitz-Thompson estimator, the ratio estimator, and the doubly robust estimator. Rotnitzky and Smucler (2020) provide a graphical criterion for identifying the optimal

covariate adjustment set for non-parametric efficient estimation of the ATE.

In the binary treatment case, when baseline covariates are high-dimensional, many regularized regression methods have been developed to perform variable selection for the PS model. Shortreed and Ertefaie (2017) propose the outcome-adaptive LASSO with a penalty function that takes into account association between covariates and outcome, and association between covariates and treatment. Ju et al. (2019b) propose a collaborative-controlled LASSO that uses the C-TMLE algorithm based on LASSO to minimize a bias-variance tradeoff in the estimated treatment effect. Tang et al. (2022) improve upon the outcome-adaptive LASSO by incorporating the ball covariance (Pan et al. 2020), which makes the method free of dependence on the outcome model specification. These methods are useful not only for screening out redundant, highly correlated variables but also for arriving at a propensity score model that is consistent and efficient in ATE estimation. However, all of these methods technically must presuppose a parametric logit model for the PS, which could be restrictive given the primary objective is no longer to explain or predict treatment assignment.

A reasonable diagnostic for the PS model is to assess the resulting covariate balance in the matched sample or balance within each stratum after stratifying on the quantiles of the PS. Austin et al. (2007) investigate the issue of variable selection by comparing the ability of different PS model specifications in balancing baseline covariates. Austin (2009), Belitser et al. (2011), and Ali et al. (2014) carry out simulations that compare the ability of different balance measures (standardized mean difference, KS distance, etc.) in assessing whether a PS model is adequate in reducing finite sample bias. However, most of these works (except Belitser and others) only focus on standard measures of balance, which assign equal weights to all covariates and hence do not make a distinction among types of covariates. Caruana et al. (2015) propose a weighted standardized mean difference for PS variable selection, where the weights are coefficients from regressing the observed outcome on the covariates.

We propose a recipe for GPS model selection for estimating pairwise ATEs in observational studies with multiple (≥ 2) treatment levels. Motivated by the aforementioned literature, the optimal GPS model is the one that includes only covariates that are predictors of the potential outcomes. Given a set of postulated GPS models, if the set happens to contain the optimal GPS model, then the proposed outcome-adjusted balance measure is able to consistently select the optimal model in large samples. The balance measure evaluates the discrepancies between two estimators of the average covariates, i.e., the GPSM estimator and the sample average of covariates, and imposes weights

that penalize models with variable selection different from the optimal GPS model. The weights incorporate the association of covariates and outcome, which can be estimated either parametrically or nonparametrically. We show the selection consistency, i.e., the optimal model can be selected with probability one asymptotically, and that the resulting GPSM estimator of the ATEs based on this model selection criterion is not only consistent and asymptotically normal. Most importantly, this estimator will be efficient based on the empirical evidence.

2.2 Background

2.2.1 Setup

Following the potential outcomes framework, we consider an observational study with T unordered treatment levels. Let $W_i \in \mathbb{W} = \{1, \ldots, T\}$ denote the treatment assignment. For each unit, there are T potential outcomes, denoted by $Y_i(w)$, for $w \in \mathbb{W}$. Implicit in this notation is the stable-unit-treatment-value assumption, which requires no interference between units and no different versions of potential outcome for each treatment level. The observed outcome is the potential outcome under the treatment received: $Y_i = Y_i(W_i)$. We also observe d baseline covariates $\boldsymbol{X}_i = \left(X_i^{(1)}, ..., X_i^{(d)}\right)^{\mathrm{T}}$. We use the symbol $\boldsymbol{X}_i^{\mathcal{S}}$ to denote a nonempty subset of the baseline covariates so that $\boldsymbol{X}_i^{\mathcal{S}} \subseteq \boldsymbol{X}_i$. Define the indicator variable $D_i(w) \in \{0, 1\}$ as $D_i(w) = 1$ if $W = w$, and $D_i(w) = 0$ otherwise. We assume that the full data $\{\boldsymbol{X}_i, W_i, Y_i(1), ..., Y_i(T)\}$, $i = 1, ..., N$, are independent and identically distributed (i.i.d) so that the observed data $\{\boldsymbol{X}_i, W_i, Y_i\}$, $i = 1, ..., N$ are also independent and identically distributed. Suppose each covariate has been normalized to have mean zero and variance one.

Imbens (2000) introduces the generalized propensity score as an extension of the propensity score to multiple treatments.

Definition 1 (Generalized propensity score) *The generalized propensity score is the conditional probability of receiving each treatment level:* $p(w \mid \boldsymbol{x}) = pr(W = w \mid \boldsymbol{X} = \boldsymbol{x})$.

Assumption 1 (Overlap) *For all values of* $\boldsymbol{x}$*, the probability of receiving any level of the treatment is positive:* $p(w \mid \boldsymbol{x}) > 0$ *for all* $w, \boldsymbol{x}$.

Assumption 2 (Weak unconfoundedness) $D(w) \perp\!\!\!\perp Y(w) \mid \boldsymbol{X}$ *for all* $w \in \mathbb{W}$.

Assumption 1 rules out deterministic treatment assignment mechanisms and allows all units to have positive probabilities of receiving any treatment level (Yang and Ding 2018). When Assumption 1 is violated, it implies that there is a sub-population for which no information on some potential outcomes is available. Assumption 2 holds if all baseline covariates that are associated with both treatment assignment and the outcome are captured. Therefore, in order to make Assumption 2 hold, practitioners often collect a rich set of pre-treatment variables, rendering variable selection a critical matter to consider.

As a result of Assumptions 1 and 2, weak unconfoundedness is preserved if we condition on the generalized propensity score:

$$D(w) \perp\!\!\!\perp Y(w) \mid p(w \mid \boldsymbol{X}) \tag{2.1}$$

One key insight from Yang et al. (2016) is that the conditional independence (2.1) is sufficient for the identification of the average potential outcomes for a single treatment level, namely $E\{Y(w)\}$ for all w, in that $E\{Y(w)\} = E[E\{Y \mid W = w, p(w \mid \boldsymbol{X})\}]$. This in turn makes the identification of the pairwise average causal effects $E\{Y(w') - Y(w)\}$ feasible.

2.2.2 Matching on the generalized propensity score

We briefly review the generalized propensity score matching (GPSM) estimator of the pairwise average treatment effects. Different from matching algorithms that construct matches only for individuals in a particular treatment group without replacement, here matching is done with replacement to impute the $T-1$ missing potential outcomes for every unit in the sample. Define the generalized propensity score matching function as $m(w, p) = \underset{j:W_j=w}{\operatorname{argmin}} ||p(w \mid \boldsymbol{X}_j) - p||$.

Since the propensity scores are continuous, ties are unlikely. In case there are ties, we can randomly break them, and subsequent arguments still hold. Given the generalized propensity score matching function, we impute $Y_i(w)$ as $\widehat{Y}_i(w) = Y_{m\{w,p(w|\boldsymbol{X}_i)\}}$.

That is, for each treatment w, we impute the potential outcome of unit i by the observed outcome from a unit in treatment w that has generalized propensity score $p(w \mid \boldsymbol{X})$ most similar to that of unit i. Implicit in this rule is that when $W_i = w$, the imputation of $Y_i(w)$ is just Y_i. We formulate the GPSM estimator of $E\{Y(w)\}$ as $\widehat{E}\{Y(w)\} = N^{-1} \sum_{i=1}^{N} \widehat{Y}_i(w)$.

The final GPSM estimator of the pairwise average treatment effect is

$$\widehat{\tau}_{\text{gpsm}}(w, w') = \widehat{E}\left\{Y(w')\right\} - \widehat{E}\left\{Y(w)\right\}. \tag{2.2}$$

Yang et al. (2016) show that including all confounders (covariates that are associated with both treatment assignment and potential outcome) in the GPS is sufficient for $\widehat{\tau}_{\text{gpsm}}(w, w')$ to be asymptotically normal and consistent for the true pairwise ATE, i.e. $E\left\{Y(w') - Y(w)\right\}$. In practice, to ensure the key unconfoundedness assumption holds, a rich set of pre-treatment covariates are collected and used to estimate the GPS. However, such an approach completely ignores the consequence for efficiency loss, especially if one includes instrumental variables, or fails to include precision variables in the GPS specification. Moreover, severe misspecification of the functional form of the optimal GPS (such as using the wrong link function, or failure to include higher-order terms) could easily lead to biased estimation of the ATEs.

2.3 Model Selection

2.3.1 Imputing covariates via GPS matching

The GPSM estimator has been used to estimate the ATEs, but we propose to use it for estimating the mean of $\boldsymbol{X}$. This might appear bizarre and unnecessary, since $E(\boldsymbol{X})$ can already be directly estimated by the sample average. The key insight is that the GPSM estimator of $E(\boldsymbol{X})$ becomes useful for gauging the quality of a GPS model specification. The following lemmas provide the stepping stones for constructing our proposed balance measure.

Lemma 1 *Given an observed covariate $X \in \boldsymbol{X}$, there exists a subset of the baseline covariates $\boldsymbol{X}^{\mathcal{S}} \subseteq \boldsymbol{X}$, such that for all $w \in \mathbb{W}$,*

$$D(w) \;\perp\!\!\!\perp\; X \;\mid\; p(w \mid \boldsymbol{X}^{\mathcal{S}}). \tag{2.3}$$

Note that (2.3) is parallel to (2.1). Because of (2.1), the imputed potential outcome for each treatment level is representative of the true potential outcome distribution over the whole sample for that treatment level. Similarly, because of (2.3), the imputed covariate for treatment level w, namely, $\widehat{X}_i(w) = X_{m\{w, p(w|\boldsymbol{X}_i^{\mathcal{S}})\}}$ is representative of the covariate distribution over the whole sample for that treatment level. We define the

GPSM estimator of $E(X)$ as the sample average of imputed covariate for treatment level w, namely, $\widehat{X}_{\text{gpsm}}(w) = N^{-1}\sum_{i=1}^{N}\widehat{X}_i(w)$. Denote the sample average of covariate X as $\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$. We have the following key result.

Lemma 2 *Assume the regularity conditions (Assumption 12 in Supplementary Material) for applying the martingale central limit theorem hold. Given an observed covariate X, choose $\boldsymbol{X}^{\mathcal{S}}$ so that (2.3) holds. Use the generalized propensity score $p(w \mid \boldsymbol{X}^{\mathcal{S}})$ as the matching variable. Then for all $w \in \mathbb{W}$, we have $\sqrt{N}\{\widehat{X}_{\text{gpsm}}(w) - \overline{X}\} \xrightarrow{d} \mathcal{N}(0, \sigma_X^2)$, where*

$$\sigma_X^2 = E\left[Var\left\{X \mid p\left(w \mid \boldsymbol{X}^{\mathcal{S}}\right)\right\}\left\{\frac{3}{2}\frac{1}{p\left(w \mid \boldsymbol{X}^{\mathcal{S}}\right)} - \frac{1}{2}p\left(w \mid \boldsymbol{X}^{\mathcal{S}}\right) - 1\right\}\right].$$

To appreciate this result, suppose we choose $\boldsymbol{X}^{\mathcal{S}} = \boldsymbol{X}$ to be the full set of observed covariates, then observing a small difference between the two estimators across all the covariates (i.e. $\sum_{j=1}^{d}|\widehat{X}_{\text{gpsm}}^{(j)}(w) - \overline{X}^{(j)}|$ is small) is an indication of the correct specification of $p(w \mid \boldsymbol{X}^{\mathcal{S}})$, and a large difference would indicate an incorrectly specified GPS model. Thus, in order to assess the quality of a proposed GPS specification $p(w \mid \boldsymbol{X}^{\mathcal{S}})$, it is meaningful to impute the covariates by matching on $p(w \mid \boldsymbol{X}^{\mathcal{S}})$, and compare how close their imputed distributions are to their observed sample distributions.

In the binary treatment setting, existing empirical literature (Austin 2009; Belitser et al. 2011; Ali et al. 2014; Caruana et al. 2015) uses different balance measures for PS model selection. The best PS model is chosen by minimizing a balance measure, which calculates the discrepancies between covariate distributions in the control and treatment arms after PS adjustments such as matching or stratification. In this article, we consider the following existing balance measures. The absolute mean difference (`AMD`) evaluates the sum of absolute differences of the covariate sample means between treatment arms, where the sum is taken over all baseline covariates. The Kolmogorov-Smirnov distance (`KSdist`) on the other hand calculates the sum of the discrepancies between the covariate CDFs and is therefore more sensitive to small differences in the shape of the distributions than `AMD`. The Mahalanobis distance (`Mdist`) computes a weighted sum of squared differences between the covariate sample means, which takes into account correlations among the covariates. The weighted balance measure (`WBM`) weights each covariate mean difference in `AMD` by the prognostic importance of the covariate. For comparison with our proposed balance measure, we extend their definitions to the multi-level treatment settings to measure the discrepancies between the imputed and the observed covariate distributions

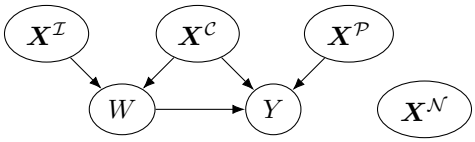


Figure 2.1: A DAG that illustrates the relationships among four types of covariates, treatment assignment, and outcome in a simple scenario.

(see Supplementary Material for more detail).

2.3.2 Types of covariates

For efficiency considerations, it is necessary to properly distinguish the types of covariates, and to examine the consequence of Lemma 1 and Lemma 2 in light of this distinction. We consider a simple scenario where covariates can be categorized into four types, with causal relationships represented by the directed acyclic graph (Figure 2.1). We let $\boldsymbol{X}^{\mathcal{C}} = \{X^{(j)} \in pa(Y) \cap pa(W) : W \text{ and } X^{(j)} \text{ are d-connected given } pa(Y) \backslash \{X^{(j)}\}\}$ denote the confounders, where $pa(Y)$ denotes the parents of Y. We further define $\boldsymbol{X}^{\mathcal{P}} = pa(Y) \backslash \boldsymbol{X}^{\mathcal{C}}$ as precision variables, $\boldsymbol{X}^{\mathcal{I}} = pa(W) \backslash \boldsymbol{X}^{\mathcal{C}}$ as instrumental variables (IVs), and $\boldsymbol{X}^{\mathcal{N}} = \boldsymbol{X} \backslash \boldsymbol{X}^{\mathcal{C} \cup \mathcal{P} \cup \mathcal{N}}$ as noise variables. Refer to Pearl (2000) for a thorough discussion of the definitions of "d-connected" and "parents", as well as necessary assumptions involved in defining a DAG. The above categorization of the baseline covariates leads to new implications for Lemma 1 and Lemma 2, which we now discuss in the following two remarks.

Remark 1 *If $X \in \boldsymbol{X}^{\mathcal{I} \cup \mathcal{C}}$ is either an instrumental variable or a confounder, then taking $\boldsymbol{X}^{\mathcal{S}}$ to include X is required for (2.3) to hold. That is, the GPS model has to include at least X for X to be conditionally independent of $D(w)$. If $X \in \boldsymbol{X}^{\mathcal{P} \cup \mathcal{N}}$ is either a precision variable or a noise variable, then (2.3) holds regardless of the choice of $\boldsymbol{X}^{\mathcal{S}}$ because the path from X to W is always blocked.*

Remark 2 *(a) We highlight Lemma 2 as the key result. Because the sample average $\overline{X}$ is consistent for the population mean $E(X)$ independent of any GPS specification $p(w \mid \boldsymbol{X}^{\mathcal{S}})$, we would expect the difference between the matching estimator $\widehat{X}_{\text{gpsm}}(w)$ and the sample mean $\overline{X}$ to be small in large samples if (i) X and $D(w)$ are independent conditional on $p(w \mid \boldsymbol{X}^{\mathcal{S}})$ and (ii) $p(w \mid \boldsymbol{X}^{\mathcal{S}})$ is correctly specified. The same logic would not apply to the potential outcomes, since $Y(w)$ are missing for individuals who are not in treatment group w.*

(b) Lemma 2 relies on Lemma 1. Therefore if $X \in \boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}}$ *is either an IV or a confounder, then* $\widehat{X}_{\text{gpsm}}(w) - \overline{X} \xrightarrow{p} 0$ *as long as* X *is included in* $\boldsymbol{X}^{\mathcal{S}}$*. If* $X \in \boldsymbol{X}^{\mathcal{P}\cup\mathcal{N}}$ *is either a precision variable or a noise variable, then* $\widehat{X}_{\text{gpsm}}(w) - \overline{X} \xrightarrow{p} 0$ *regardless of whether* $\boldsymbol{X}^{\mathcal{S}}$ *includes* X*.*

So far no theoretical result has justified which combination of the variables to include in the GPS model for the GPSM estimator to be efficient. Thus, we rely on existing empirical and theoretical findings in the binary treatment case (Brookhart et al. 2006; Austin et al. 2007; Myers et al. 2011; Tang et al. 2022) to establish that the optimal GPS specification should also include only the confounders and precision variables (i.e., $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}})$).

Despite the popularity of `AMD`, `KSdist`, and `Mdist`, they place equal emphasis on balancing *all* baseline covariates, which is not ideal for identifying the optimal GPS model. To see this, a model selection criterion based on minimizing these balance measures will favor models that include $\boldsymbol{X}^{\mathcal{I}}$ and $\boldsymbol{X}^{\mathcal{C}}$ regardless of whether they include $\boldsymbol{X}^{\mathcal{P}}$ or $\boldsymbol{X}^{\mathcal{N}}$, since $\boldsymbol{X}^{\mathcal{P}}$ or $\boldsymbol{X}^{\mathcal{N}}$ will always be balanced due to Remark 2(b). As a result, the matching estimator based on a GPS model that minimizes the absolute mean difference would be consistent for the ATE but not necessarily efficient. While using `WBM` increases the chance of including prognostically important variables, it could still fail to exclude $\boldsymbol{X}^{\mathcal{I}}$ and $\boldsymbol{X}^{\mathcal{N}}$ in large samples due to the small weights they would receive. Nonetheless, it is possible to equip the absolute mean difference with outcome information to help identify the optimal GPS model.

2.3.3 Outcome-adjusted balance measure

To conform with evidence that advocates $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}})$ as the optimal GPS model, we make adjustments to the absolute mean difference balance measure by leveraging the outcome information. We let $\rho\left(X^{(j)}, Y \mid W = w\right)$ denote a generic metric of correlation between the observed outcome Y and jth observed covariate $X^{(j)}$ conditional on treatment level $W = w$. We let $\rho_N\left(X^{(j)}, Y \mid W = w\right)$ denote its empirical version. Define the outcome-adjusted balance measure for the GPS model $p(w \mid \boldsymbol{X}^{\mathcal{S}})$ at $w \in \mathbb{W}$ to be:

$$\begin{aligned}\texttt{OABM}_{\rho}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\} &= \boldsymbol{\zeta}_{\rho}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}^{\mathrm{T}} \left|\widehat{\boldsymbol{X}}_{\text{gpsm}}(w) - \overline{\boldsymbol{X}}\right| \\ &= \sum_{j=1}^{d} \zeta_{\rho}^{(j)}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\} \left|\widehat{X}^{(j)}_{\text{gpsm}}(w) - \overline{X^{(j)}}\right|,\end{aligned}$$

where $\boldsymbol{\zeta}_\rho\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\} = \left[\zeta_\rho^{(1)}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}, ..., \zeta_\rho^{(d)}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}\right]^{\mathrm{T}}$ is defined as follows:

$$\zeta_\rho^{(j)}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\} = \begin{cases} 1/\rho_N\left(X^{(j)}, Y \mid W = w\right) & \text{if } X^{(j)} \in \boldsymbol{X}^{\mathcal{S}} \\ \delta_w \rho_N\left(X^{(j)}, Y \mid W = w\right) & \text{if } X^{(j)} \notin \boldsymbol{X}^{\mathcal{S}}. \end{cases}$$

That is, if $X^{(j)}$ is included in $\boldsymbol{X}^{\mathcal{S}}$, then $\zeta_\rho^{(j)}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}$ takes the value $1/\rho_N(X^{(j)}, Y \mid W = w)$; if $X^{(j)}$ is excluded from $\boldsymbol{X}^{\mathcal{S}}$, then $\zeta_\rho^{(j)}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}$ assumes the value $\delta_w \rho_N(X^{(j)}, Y \mid W = w)$. We let δ_w be a positive tuning parameter that is proportional to $N^{1/3}$.

The purpose of the weighting design $\boldsymbol{\zeta}_\rho\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}$ is to penalize posited models that differ in variable selection from the optimal GPS model, which should only include $\boldsymbol{X}^{\mathcal{P}}$ and $\boldsymbol{X}^{\mathcal{C}}$. Consider a strong predictor $X^{(j)}$ of the outcome Y at treatment level w, which would be reflected by a large $\rho_N\left(X^{(j)}, Y \mid W = w\right)$. If $X^{(j)}$ is excluded from a candidate GPS model and therefore unbalanced, we impose a large penalty $\delta_w \rho_N\left(X^{(j)}, Y \mid W = w\right)$. On the contrary, if we learn that $\rho_N\left(X^{(j)}, Y \mid W = w\right)$ is small, this will indicate that $X^{(j)}$ is weakly correlated or uncorrelated with the outcome. In that case, if $X^{(j)}$ is included in a candidate GPS model and therefore balanced, we penalize such a model by imposing a large weight $1/\rho_N(X^{(j)}, Y \mid W = w)$.

The choice of the correlation metric ρ should depend on whether one is confident in correctly specifying the parametric outcome model. For instance, if one has strong knowledge that the potential outcome $Y(w)$ is linear in the covariates with partial regression coefficients $(\theta_w^{(1)}, ..., \theta_w^{(d)})$, one should let $\rho_N\left(X^{(j)}, Y \mid W = w\right) = |\widehat{\theta}_w^{(j)}|$ and select the GPS model that minimizes $\texttt{OABM}_{\texttt{OLS}}$. On the contrary, if one knows little about the relationship between $Y(w)$ and $\boldsymbol{X}$, it is recommended to use the ball correlation (Pan et al. 2020) as the correlation metric given that it is free of dependence on the modeling assumptions, and one should choose the GPS model that minimizes $\texttt{OABM}_{\texttt{BCor}}$. Other model-free correlation metrics, such as the distance correlation (Székely et al. 2007), are also feasible alternatives for ρ.

The role of the tuning parameter δ_w is to ensure adequate finite sample performance. If the outcome model is correctly specified, choosing $\delta_w = N^{1/3}$ would be sufficient. If this is not the case, and one uses the ball correlation as the metric of correlation, then for every treatment level w, we first let $BCor_N^* = \{\max_{j:X^{(j)} \in \boldsymbol{X}^{\mathcal{H}_0}} BCor_N(X^{(j)}, Y \mid W = w) + \min_{j:X^{(j)} \in \boldsymbol{X}^{\mathcal{H}_1}} BCor_N(X^{(j)}, Y \mid W = w)\}/2$, and then choose δ_w by minimizing $\left|1/BCor_N^* - \delta_w BCor_N^*\right|$. Here, $\boldsymbol{X}^{\mathcal{H}_0}$ consists of baseline covariates for which the ball

covariance test between each covariate $X^{(j)}$ and Y conditional on $W = w$ fails to reject the null hypothesis, and $\boldsymbol{X}^{\mathcal{H}_1}$ is defined as the baseline covariates for which the ball covariance test between each covariate $X^{(j)}$ and Y conditional on $W = w$ rejects the null hypothesis. Therefore, $\max_{j:X^{(j)}\in\boldsymbol{X}^{\mathcal{H}_0}} BCor_N(X^{(j)}, Y \mid W = w)$ corresponds to the largest empirical ball correlation value among variables tested to be unrelated to Y at treatment level w, whereas $\min_{j:X^{(j)}\in\boldsymbol{X}^{\mathcal{H}_1}} BCor_N(X^{(j)}, Y \mid W = w)$ refers to the smallest empirical ball correlation value among variables tested to be related to Y at treatment level w, and $BCor_N^*$ is therefore a threshold ball correlation value that separates the covariates that are truly related to the outcome from those that are not.

The outcome-adjusted balance measure is computed separately for each treatment level w. Therefore it is possible that for a different treatment arm, relevant baseline covariates might be different, and a different GPS model might minimize $\texttt{OABM}_\rho$ corresponding to that arm. In such cases, the estimated GPS across all the treatments might not sum to one for some individuals. We allow this to occur in finite samples because the goal is not to estimate the GPS as accurately as possible, but rather to leverage its role of being a balancing score and efficiently estimate the mean potential outcomes. Also, in large samples, because of the model selection consistency result in Theorem 1, such a phenomenon will occur with small probabilities. The selected GPS model for treatment level w can then be used to construct a point estimate for the mean potential outcome $E\{Y(w)\}$, and then for the pairwise treatment effects. For variance estimation, we recommend using the Abadie and Imbens (2016) variance estimator, with slight adjustments made to allow the possibility that different GPS models could be selected by $\texttt{OABM}_\rho$ for different treatment levels. The variance estimator formula and summarizing steps to perform GPS model selection are contained in the Supplementary Material.

2.4 Theoretical Properties

In this section, we present asymptotic results for the outcome-adjusted balance measure as well as for the resulting GPSM estimator. Define the conditional mean and variance of $Y(w)$ given $p(w \mid \boldsymbol{X}^{\mathcal{S}})$ as $\bar{\mu}\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\} = E\left\{Y(w) \mid p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}$ and $\bar{\sigma}^2\{w, p(w \mid \boldsymbol{X}^{\mathcal{S}})\} = Var\left\{Y(w) \mid p(w \mid \boldsymbol{X}^{\mathcal{S}})\right\}$. Define the mean potential outcome as $\mu(w) = E\{Y(w)\}$. Let $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}})$ denote the optimal generalized propensity score model.

In addition to the assumptions made in Section 2.2, model selection consistency of $\texttt{OABM}_\rho$ relies on correct specification of the parametric outcome model if ρ is chosen to be

parametric.

Assumption 3 (Correct specification of parametric outcome model) *For all $w \in \mathbb{W}$, the relationship between the potential outcome $Y_i(w)$ and covariates $\boldsymbol{X}_i^{\mathcal{C}\cup\mathcal{P}}$ can be characterized by a known parametric density/mass function $f_w(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\phi})$, $i = 1, ..., N$, where the parameters of interest $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ are associated with the effect of $\boldsymbol{X}_i^{\mathcal{C}\cup\mathcal{P}}$, and $\boldsymbol{\phi}$ are the nuisance parameters.*

Theorem 1 (Model selection consistency) *Let the collection of posited generalized propensity score models be $\mathscr{P}$ and suppose the optimal model $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}) \in \mathscr{P}$ is one of the posited models. Suppose Assumptions 1 and 2 hold. (i) Suppose also that Assumption 3 is satisfied. Let $\rho_N(X^{(j)}, Y \mid W = w)$ be the absolute value of a CAN (consistent and asymptotic normal) estimator of the coefficient of $X^{(j)}$ in the parametric outcome model of $Y(w)$.*

Then for all w, $\lim_{n\to\infty} \mathbb{P}[\texttt{OABM}_{\texttt{CAN}}\{w, p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}})\} \leq \texttt{OABM}_{\texttt{CAN}}(w, p_k)$ for $p_k \in \mathscr{P}] = 1$.

(ii) Suppose that $\rho_N(X^{(j)}, Y \mid W = w) = o(1)$ for all $X^{(j)} \in \boldsymbol{X}^{\mathcal{I}}$. Let $\rho_N(X^{(j)}, Y \mid W = w)$ be the empirical ball correlation between $X^{(j)}$ on Y conditional on $W = w$. Then for all w, we have $\lim_{n\to\infty} \mathbb{P}[\texttt{OABM}_{\texttt{BCor}}\{w, p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}})\} \leq \texttt{OABM}_{\texttt{BCor}}(w, p_k)$ for $p_k \in \mathscr{P}] = 1$.

Remark 3 *When ρ_N is a nonparametric correlation metric such as the ball correlation or the distance correlation, the above consistency result holds approximately because W is a collider of $\boldsymbol{X}^{\mathcal{I}}$ and $\boldsymbol{X}^{\mathcal{C}}$ (see Figure 2.1). A consequence of this is that in general $\rho(X^{(j)}, Y \mid W = w) \neq 0$ for $X^{(j)} \in \boldsymbol{X}^{\mathcal{I}}$. In such cases, the reliability of $\texttt{OABM}_{\texttt{BCor}}$ depends on how small such correlation $\rho(X^{(j)}, Y \mid W = w)$ is in reality. When ρ_N is taken to be parametric and the outcome model is correct, it is necessarily true that $\rho(X^{(j)}, Y \mid W = w) = 0$ for $X^{(j)} \in \boldsymbol{X}^{\mathcal{I}\cup\mathcal{N}}$ since $X^{(j)}$ is independent of Y conditional on W and $\boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}$.*

Theorem 1 is the main result, which states that if one is capable of correctly specifying the outcome model or collider bias is negligible (i.e. $\rho(X^{(j)}, Y \mid W = w) = o(1)$), then the outcome-adjusted balance measure is guaranteed to identify the optimal GPS model among a set of posited models in large samples.

Yang et al. (2016) show that when the true GPS has a multinomial logit form and is estimated using maximum likelihood, the GPSM estimator matching on the estimated GPS is consistent and asymptotically normal. We combine this result with Theorem 1 and state in Theorem 2 below that matching on the optimal GPS model selected by

the balance measure results in a GPSM estimator that is consistent and asymptotically normal for the average potential outcomes. Based on the empirical evidence discussed in the previous sections, we conclude that this GPSM estimator is also efficient.

Let treatment level 1 be the reference (baseline) category. We assume the following generalized linear form of the selected optimal GPS model, i.e. $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}} = \boldsymbol{x}; \boldsymbol{\beta}) = p\left(w \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_2, \ldots, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_T\right)$ for all w. Let $I_{\boldsymbol{\beta}}$ be the information matrix. We estimate $\boldsymbol{\beta}$ using maximum likelihood, and denote the estimated GPS as $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \widehat{\boldsymbol{\beta}})$. Define the GPSM estimator of $\mu(w)$ that matches on $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \widehat{\boldsymbol{\beta}})$ as $N^{-1}\sum_{i=1}^{N} Y_{m\left\{w, p(w|\boldsymbol{X}_i^{\mathcal{C}\cup\mathcal{P}}; \widehat{\boldsymbol{\beta}})\right\}}$.

For example, assume the parametric model is a multinomial logit model. That is, for $w \in \{2, ..., T\}$, we assume the following model: $p(w \mid \boldsymbol{x}; \boldsymbol{\beta}) = p\left(w \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_2, \ldots, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_T\right) = \exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_w\right) \times \left\{1 + \sum_{w'=2}^{T} \exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_{w'}\right)\right\}^{-1}$, where $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_2^{\mathrm{T}}, ..., \boldsymbol{\beta}_T^{\mathrm{T}})$ and $p(1 \mid \boldsymbol{x}; \boldsymbol{\beta}) = \left\{1 + \sum_{w'=2}^{T} \exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_{w'}\right)\right\}^{-1}$.

Assumption 4 *The optimal GPS model $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})$ has a continuous distribution with compact support $[\underline{p}, \overline{p}]$ and with a continuous density function. The conditional expectation of potential outcome $\bar{\mu}(w, p)$ is Lipschitz-continuous in p. For some $\delta > 0$, $E\{|Y|^{2+\delta} | W = w, p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) = p\}$ is uniformly bounded.*

Theorem 2 (Asymptotic normality based on optimal GPS model) *Suppose all assumptions made in Theorem 1 are satisfied. Suppose also that Assumption 4 is satisfied. Assume a generalized linear model for $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})$. Then $p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})$ is the GPS model selected by minimizing the outcome-adjusted balance measure at treatment level w, and*

$$\sqrt{N}\left[N^{-1}\sum_{i=1}^{N} Y_{m\left\{w, p(w|\boldsymbol{X}_i^{\mathcal{C}\cup\mathcal{P}}; \widehat{\boldsymbol{\beta}})\right\}} - \mu(w)\right] \xrightarrow{d} \mathcal{N}\left\{0, \sigma^2(w) - \boldsymbol{c}(w)^{T} I_{\boldsymbol{\beta}}^{-1} \boldsymbol{c}(w)\right\}$$

where

$$\begin{aligned}\sigma^2(w) =& E\left[\bar{\sigma}^2\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})\right\}\left\{\frac{3}{2}\frac{1}{p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})} - \frac{1}{2}p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})\right\}\right] \\ &+ E\left(\left[\bar{\mu}\left\{w, p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})\right\} - \mu(w)\right]^2\right).\end{aligned}$$

The first term $\sigma^2(w)$ is the asymptotic variance associated with matching on the true optimal generalized propensity score, and $\boldsymbol{c}(w)^{\mathrm{T}} I_{\boldsymbol{\beta}}^{-1} \boldsymbol{c}(w)$ corresponds to the gain in precision when matching variable is the estimated optimal GPS. We rely on empirical evidence to conclude that $\sigma^2(w) - \boldsymbol{c}(w)^{\mathrm{T}} I_{\boldsymbol{\beta}}^{-1} \boldsymbol{c}(w)$ is smaller than the asymptotic variance

corresponding to any other GPS model specifications. The precise definition of $\boldsymbol{c}(w)$ is given in the Supplementary Material.

2.5 A Simulation Study

In this section, we examine the finite sample performance of the outcome-adjusted balance measure in a simulation study with three treatment levels. For each simulated dataset, we generate nine independent standard normal covariates $\boldsymbol{X}_i = \left(X_i^{(1)}, ..., X_i^{(9)}\right)^{\mathrm{T}}$ for each individual i. The type of each covariate is specified as follows: $\boldsymbol{X}^{\mathcal{C}} = (X^{(1)}, X^{(2)})^{\mathrm{T}}$, $\boldsymbol{X}^{\mathcal{I}} = (X^{(3)}, X^{(4)}, X^{(5)})^{\mathrm{T}}$, $\boldsymbol{X}^{\mathcal{P}} = (X^{(6)}, X^{(7)}, X^{(8)})^{\mathrm{T}}$, $\boldsymbol{X}^{\mathcal{N}} = (X^{(9)})^{\mathrm{T}}$. We follow the same procedure described in Yang et al. (2016) to generate the treatment indicators $D_i(1)$, $D_i(2)$, $D_i(3)$ from a multinomial distribution with event probabilities

$$p(W_i = w|\boldsymbol{X}_i) = \exp\left\{(1, \boldsymbol{X}_i)^{\mathrm{T}} \boldsymbol{\beta}_w\right\} / \sum_{w'=1}^{3} \exp\left\{(1, \boldsymbol{X}_i)^{\mathrm{T}} \boldsymbol{\beta}_{w'}\right\} \tag{2.4}$$

for $w = \{1, 2, 3\}$. Here, $D_i(w) = 1$ if unit i belongs to treatment w. We set $\boldsymbol{\beta}_1^T = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_2^T = 0.7 \times (0, 1.5, 1.5, u, u, u, 0, 0, 0, 0)$, and $\boldsymbol{\beta}_3^T = 0.4 \times (0, 1.5, 1.5, u, u, u, 0, 0, 0, 0)$, where $u = 1, 2$ controls the strength of instrumental variables. We fix the sample size for each treatment to be $N_w = 500$, for $w \in \{1, 2, 3\}$. This is accomplished by first generating a large superpopulation based on (2.4), and then collecting a simple random sample of size 500 from each treatment group. With this constraint, it can be shown that $D_i(1)$, $D_i(2)$, $D_i(3)$ still follow a multinomial distribution with event probabilities (2.4), but with different intercept terms for $\boldsymbol{\beta}_w$'s.

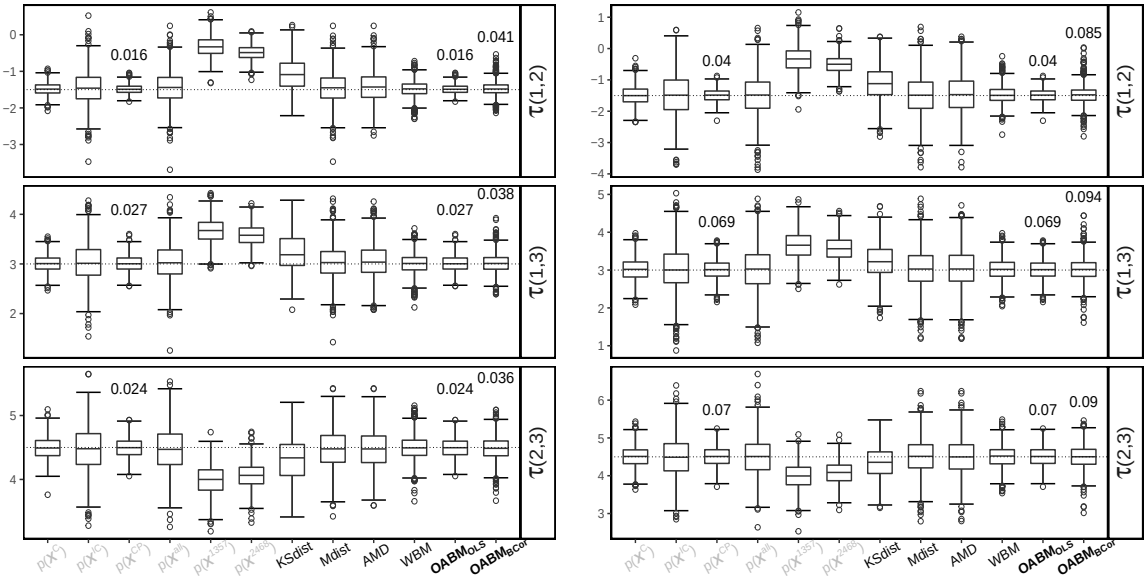
We consider two common ways potential outcomes may be related to the covariates. First, we let the mean potential outcome be a linear function of the covariates so that $Y_i(w) = (1, \boldsymbol{X}_i)^{\mathrm{T}} \boldsymbol{\theta}_w + \eta_i$ with $\eta_i \sim \mathcal{N}(0, 1)$. Second, we consider potential outcomes generated from a nonlinear function of the covariates, that is $Y_i(w) = \{1, (\boldsymbol{X}_i^2 - 1)/2\}^{\mathrm{T}} \boldsymbol{\theta}_w + \eta_i$ with $\eta_i \sim \mathcal{N}(0, 1)$. We let $\boldsymbol{\theta}_1^T = (-1.5, 1.5, 1.5, 0, 0, 0, v, v, v, 0)$, $\boldsymbol{\theta}_2^T = (-3, 1.5, 1.5, 0, 0, 0, v, v, v, 0)$, and $\boldsymbol{\theta}_3^T = (1.5, 1.5, 1.5, 0, 0, 0, -v, -v, -v, 0)$, where $v = 1, 2$ controls the strength of relationship between precision variables and potential outcomes. The true pairwise average treatment effects are $\tau(1, 2) = -1.5, \tau(1, 3) = 3, \tau(2, 3) = 4.5$.

We postulate six multinomial regression models that include different sets of covariates as linear terms for estimating the GPS. Model $\mathtt{p}(\mathtt{X}^{\mathtt{C}})$ includes only the confounders

$\boldsymbol{X}^{\mathcal{C}}$. Model $\mathtt{p}(\mathtt{X}^{\mathtt{IC}})$ includes both confounders $\boldsymbol{X}^{\mathcal{C}}$ and instrumental variables $\boldsymbol{X}^{\mathcal{I}}$. Model $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$ includes confounders as well as covariates that are only related to outcome (i.e., $\boldsymbol{X}^{\mathcal{C}}$, $\boldsymbol{X}^{\mathcal{P}}$). Model $\mathtt{p}(\mathtt{X}^{\mathtt{all}})$ includes all nine covariates. Model $\mathtt{p}(\mathtt{X}^{\mathtt{1357}})$ includes covariates $(X^{(1)}, X^{(3)}, X^{(5)}, X^{(7)})$, and model $\mathtt{p}(\mathtt{X}^{\mathtt{2468}})$ includes $(X^{(2)}, X^{(4)}, X^{(6)}, X^{(8)})$. Each model includes the same set of covariates across 1,000 simulated datasets. These six GPS models serve as benchmark models for comparing the performance of $\mathtt{OABM}_{\rho}$ to other measures. Among these six postulated models, only $\mathtt{p}(\mathtt{X}^{\mathtt{IC}})$ and $\mathtt{p}(\mathtt{X}^{\mathtt{all}})$ are correctly specified, in the sense of correctly describing the distribution of W conditional on the chosen set of covariates. All other models are misspecified, but only to a mild degree as to not have severe consequences.

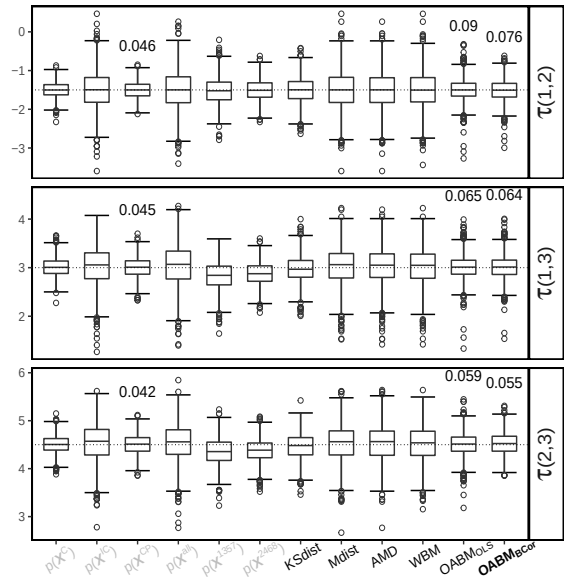
We use each postulated GPS model specification to estimate the GPS for each individual in the simulated datasets. We carry out the model selection procedure with two versions of $\mathtt{OABM}_{\rho}$ and four other existing balance measures defined in the Supplementary Material: `WBM`, `AMD`, `Mdist`, and `KSdist`. $\mathtt{OABM}_{\mathtt{BCor}}$ is a variant of the outcome-adjusted balance measure with ρ set to be the ball correlation. $\mathtt{OABM}_{\mathtt{OLS}}$ is another variant $\mathtt{OABM}_{\rho}$ with ρ chosen as the maximum likelihood estimator of the partial regression coefficient from assuming that the potential outcomes are linear in the covariates. Due to space constraints, here we only present results for 4 of the 8 total scenarios. Results of other scenarios can be found in the Supplementary Material.

Figure 2.2 summarizes the performance of the GPSM estimator based on the six benchmark models and five measures under the 4 chosen scenarios. Models $\mathtt{p}(\mathtt{X}^{\mathtt{1357}})$ and $\mathtt{p}(\mathtt{X}^{\mathtt{2468}})$ both yield biased estimates in all four scenarios due to failure to include all confounders. While all other benchmark models lead to unbiased estimates, estimates by matching on $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$ result in the smallest variance. Under the linear outcome design, $\mathtt{OABM}_{\mathtt{OLS}}$ has mean squared error closest to the optimal benchmark model, and outperforms all other measures. $\mathtt{OABM}_{\mathtt{BCor}}$ results in mean squared error smaller than `Mdist` and `AMD`, and performs similar to `WBM`. Bias occurs when matching on GPS models selected based on `KSdist`. When the mean potential outcome is nonlinear in the covariates, $\mathtt{OABM}_{\mathtt{OLS}}$ no longer dominates in performance due to the outcome model misspecification. In this case, because of the model-free property of the ball correlation, $\mathtt{OABM}_{\mathtt{BCor}}$ shows a slight advantage over all other measures in terms of MSE. Overall, the results indicate that matching on GPS models selected by our proposed methods result in efficiency gain, since the MSE of $\mathtt{OABM}_{\mathtt{OLS}}$ (when outcome model is correct) and $\mathtt{OABM}_{\mathtt{BCor}}$ are fairly close to the MSE of the most efficient benchmark model $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$. For interval estimation, despite mild cases of under-coverage for the two `OABM` measures in scenario (b), coverage rates for both

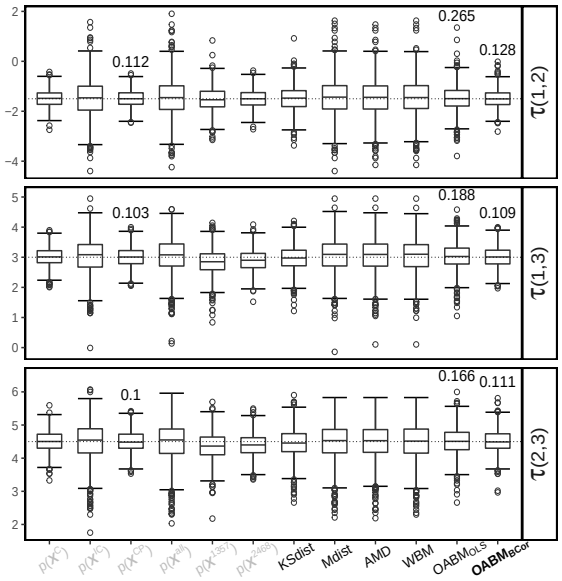


(a) $u = 2, v = 1$
Outcome *linear* in covariates

(b) $u = 2, v = 2$
Outcome *linear* in covariates

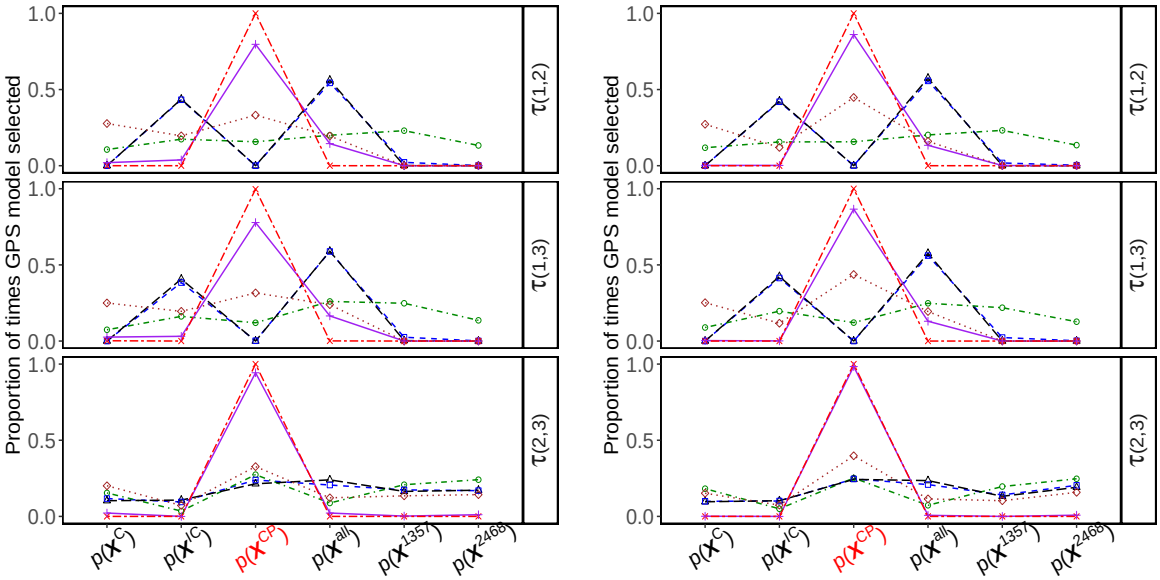


(c) $u = 2, v = 1$
Outcome *nonlinear* in covariates



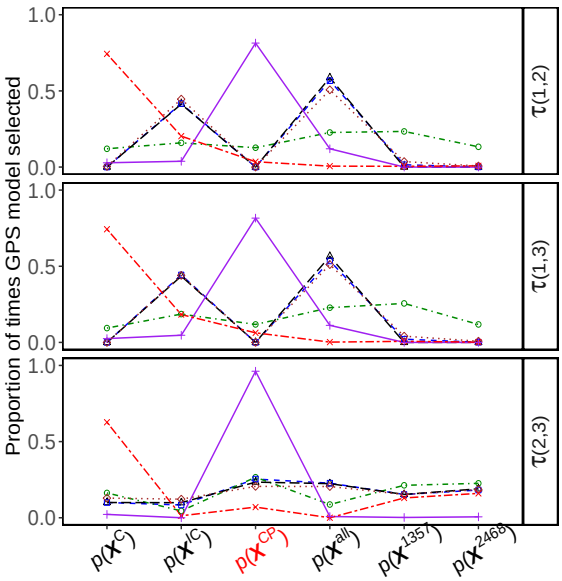
(d) $u = 2, v = 2$
Outcome *nonlinear* in covariates

Figure 2.2: Box plots of GPSM estimates. Numeric MSEs for $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$, $\mathtt{OABM}_{\mathtt{BCor}}$ and $\mathtt{OABM}_{\mathtt{OLS}}$ are explicitly shown above their corresponding box plots. True pairwise treatment effects are denoted by the horizontal dotted lines.

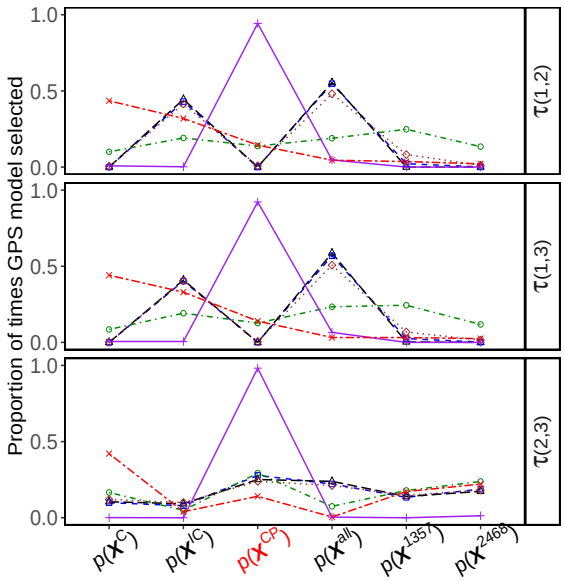


(a) $u = 2, v = 1$
Outcome *linear* in covariates

(b) $u = 2, v = 2$
Outcome *linear* in covariates



(c) $u = 2, v = 1$
Outcome *nonlinear* in covariates



(d) $u = 2, v = 2$
Outcome *nonlinear* in covariates

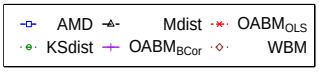


Figure 2.3: Proportion of model selection by six balance measures.

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.586	0.688	0.725
Mdist	0.908	0.907	0.929
AMD	0.897	0.900	0.911
WBM	0.958	0.963	0.965
$\texttt{OABM}_{\texttt{OLS}}$	0.964	0.935	0.950
$\texttt{OABM}_{\texttt{BCor}}$	0.954	0.936	0.948

(a) $u = 2, v = 1$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.744	0.800	0.851
Mdist	0.932	0.898	0.907
AMD	0.918	0.887	0.896
WBM	0.986	0.945	0.950
$\texttt{OABM}_{\texttt{OLS}}$	0.987	0.917	0.917
$\texttt{OABM}_{\texttt{BCor}}$	0.969	0.916	0.916

(b) $u = 2, v = 2$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.929	0.939	0.937
Mdist	0.914	0.904	0.898
AMD	0.908	0.905	0.900
WBM	0.900	0.909	0.890
$\texttt{OABM}_{\texttt{OLS}}$	0.930	0.937	0.946
$\texttt{OABM}_{\texttt{BCor}}$	0.929	0.938	0.934

(c) $u = 2, v = 1$
Outcome *nonlinear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.941	0.942	0.932
Mdist	0.905	0.908	0.900
AMD	0.903	0.911	0.903
WBM	0.898	0.920	0.900
$\texttt{OABM}_{\texttt{OLS}}$	0.927	0.918	0.935
$\texttt{OABM}_{\texttt{BCor}}$	0.941	0.942	0.946

(d) $u = 2, v = 2$
Outcome *nonlinear* in covariates

Table 2.1: Coverage rates of asymptotic 95% confidence intervals using the GPSM point estimator and Abadie and Imbens (2016) variance estimator with GPS models selected by competing balance measures; simulation study

$\texttt{OABM}_{\texttt{BCor}}$ and $\texttt{OABM}_{\texttt{OLS}}$ do not deviate much from the specified 95% probability in the other scenarios, as shown in Table 2.1.

In Figure 2.3 we present the proportion of times each of the six benchmark GPS models are selected by the measures. When outcome is linear in covariates, $\texttt{OABM}_{\texttt{OLS}}$ consistently selects $\texttt{p}(\texttt{X}^{\texttt{CP}})$, the GPS model that results in the smallest mean squared error. Following $\texttt{OABM}_{\texttt{OLS}}$, $\texttt{OABM}_{\texttt{BCor}}$ selects $\texttt{p}(\texttt{X}^{\texttt{CP}})$ with high proportions, while occasionally selecting the benchmark model that includes all covariates. `Mdist` and `AMD` tend to select GPS models that include the IVs, thereby explaining their large estimation variance. When the potential outcome is nonlinear in the covariates, $\texttt{OABM}_{\texttt{OLS}}$ no longer selects $\texttt{p}(\texttt{X}^{\texttt{CP}})$, while $\texttt{OABM}_{\texttt{BCor}}$ is still able to identify it thanks to the model-free property of the ball correlation. In this case, in addition to `Mdist` and `AMD`, `WBM` also frequently selects GPS models that include the instrumental variables. `KSdist`'s worst performance could be explained by its inability to rule out the two biased benchmark models.

2.6 Applications

We apply our proposed balance measure to two data applications. Due to space constraints, we present one application in the main article and leave the analysis of the Tutoring Data in the Supplementary Material.

The evaluation of providers' performance based on patient outcomes is a crucial step to help improve the quality of health care programs in the United States. To estimate the effect of providers on patient outcomes, random assignment of patients to different providers would be ideal but unviable in practice, since patients may choose providers according to their medical profiles. As a result, the program evaluation data can be viewed as observational data with multiple treatment levels.

The Nursing Home compiled by Scotina and Gutman (2019) contains information on Medical enrollees linked to the Minimum Data Set who were discharged from Rhode Island Hospital over a 9-year period beginning on January 2, 1999 and were assigned to one of 5 nearby skilled nursing facilities (SNFs) in Rhode Island or Massachusetts. To compare the performance and quality of the SNFs, the measured outcome of interest is whether a patient was readmitted to Rhode Island Hospital within 30 days of their initial discharge. Each patient's age, gender, and year of admission are collected as demographic characteristics. Clinical characteristics include ICD-9 diagnosis code for hospital admission. Other covariates include inflation-adjusted reimbursement for the hospital stay, which

reflects the intensiveness of the stay, as well as intensive care unit use and SNF use in the 120 days prior to hospitalization. The number of patients in these SNFs ranges from 670 to 1211. For ease of notation, we represent nursing home 1 by treatment 1, nursing home 2 by treatment 2, and so on, so that we have $\mathbb{W} = \{1, 2, 3, 4, 5\}$.

To illustrate the efficiency gain by using the propose balance measure, we first apply the group lasso (Yuan and Lin 2006) for multinomial logit regression to select strong predictors of nursing home assignment. We then apply the lasso and select strong predictors of the outcome separately at each treatment level. For estimation of $E\{Y(1)\}$, the average potential rehospitalization rate if all patients were assigned to nursing home 1, we categorize the 30 covariates into instrumental variables, confounders, and precision variables based on the lasso variable selection result for this treatment level. We follow the same procedure to partition the covariates into the 3 types separately for estimating $E\{Y(2)\}$,..., and $E\{Y(5)\}$. For each treatment level $\omega \in \{1, 2, 3, 4, 5\}$, we estimate $E\{Y(w)\}$ by positing 3 candidate GPS models:

- Model 1: multinomial logit with all 30 covariates entered linearly
- Model 2: multinomial logit with all confounders and precision variables entered linearly
- Model 3: model 2 with the addition of all first-order interaction terms between the outcome related variables

Since we lack prior knowledge in correctly specifying the parametric outcome models, we use $\texttt{OABM}_{\texttt{BCor}}$ to select the best GPS model separately for each treatment level, and the model selection results are shown in Table 2.2. Model 2 is consistently selected across all 5 treatment levels. We fit the selected models to estimate the generalized propensity scores, which serve as matching variables for the GPSM estimator to produce point estimates and standard errors of contrasts between nursing home 1 and nursing home 2 through 5. To illustrate the benefit of $\texttt{OABM}_{\texttt{BCor}}$, we compare it with naively adopting the GPS model that involves only strong predictors of treatment assignment (GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$). Table 2.3 summarizes the result of the analysis, which shows a reduction in the standard errors by using the proposed $\texttt{OABM}_{\texttt{BCor}}$ for GPS model selection for three of the four contrasts.

Among these pairwise comparisons, only the one between nursing home 1 and 5 yields a statistically significant difference in 30-day rehospitalization rates at the 0.05 level. This implies that nursing home 5 generally takes better care of patients than nursing home 1

Model	$E\{Y(1)\}$	$E\{Y(2)\}$	$E\{Y(3)\}$	$E\{Y(4)\}$	$E\{Y(5)\}$
1	1241	1341	4760	2751	19882860
2	**233**	**534**	**67**	**42**	**15**
3	603	892	186	79	20

Table 2.2: The proposed $\texttt{OABM}_{\texttt{BCor}}$ evaluated at each treatment level, for each posited model; bold number indicates the lowest $\texttt{OABM}_{\texttt{BCor}}$ value for that treatment; nursing home data

Method	$\hat{\tau}(1,2)$	$\hat{\tau}(1,3)$	$\hat{\tau}(1,4)$	$\hat{\tau}(1,5)$
GPSM-$\texttt{OABM}_{\texttt{BCor}}$	0.0418 (0.0281)	0.0223 (**0.0289**)	0.0402 (**0.0280**)	0.0721 (**0.0262**)*
GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$	0.0268 (0.0277)	0.0400 (0.0314)	0.0296 (0.0292)	0.0659 (0.0372)

Table 2.3: GPSM estimates of the pairwise ATEs between nursing home 1 and all other nursing homes; GPSM-$\texttt{OABM}_{\texttt{BCor}}$ uses $\texttt{OABM}_{\texttt{BCor}}$ to select GPS model among the three posited models; GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$ uses a multinomial logit model for the GPS specification that involves linear terms of IVs and confounders; standard errors are displayed in parenthesis; * indicates significance at the 0.05 level; nursing home data

within the first month of their discharge from the hospital, while the quality of nursing home 2, 3, and 4 are at a similar level as nursing home 1.

2.7 Discussion

In this article, we present a novel balance measure that can select the optimal generalized propensity score model from a set of candidate models for the estimation of each average potential outcome. Given a set of candidate GPS models, the outcome-adjusted balance measure works as follows. First we impute all measured covariates by matching on each candidate GPS model as if they were missing. For each candidate model, the balance measure then evaluates a weighted sum of the absolute mean differences between the imputed covariates and their original sample values. These absolute differences are each scaled by a penalty, which takes into account the covariate's relationship with the potential outcome to ensure that only the desired covariates are selected into the GPS model. Under certain conditions, we theoretically show that in large samples, the GPS specification that minimizes the proposed balance measure is the optimal GPS model. As a result, the GPSM estimator based on a GPS model selected by the outcome-adjusted balance measure is consistent and efficient for the pairwise ATEs. The proposed method hinges on the fact that the set of posited GPS models must contain the optimal model or at least

one close approximation to it. If the set of posited GPS models only consists of severely misspecified models, then the balance measure may not be able to suggest the "best" one among them.

Under the correct specification of (generalized) propensity score models, Abadie and Imbens (2006) and Yang et al. (2016) propose matching estimators to impute the missing potential outcomes for treatment effect estimation. The key innovation of our proposed framework lies in imputing the "potential" covariates for balancing assessment to gauge model specification on the GPS. Adjustment of the balance metric by the outcome information plays a key role for selecting the optimal subset of baseline covariates. We also establish the new model selection consistency results, which are absent in the literature.

Since the proposed balance measure incorporates outcome information to conduct GPS model selection, this is contradictory to Rubin's principle that modeling of the assignment mechanism should be carried out at the design stage, without accessing any outcomes (Rubin 2007). However, for efficiency considerations, leveraging outcome information is necessary. In practice, if pilot studies exist, the outcome information can be borrowed from those studies.

We focus on searching for a scalar balancing score for matching in this article, as Abadie and Imbens (2006) show that matching on a vector of variables with a dimension greater than one results in asymptotic bias for the matching estimator. In the case of binary treatment, the propensity score is by far the most popular scalar balancing score, and propensity score matching has been widely used in practice due to its intuitive appeal and simplicity. It was believed that it is infeasible to extend the advantage of propensity score matching (i.e., matching on a scalar balancing score) to the case of multi-level treatments until Yang et al. (2016). The key idea is that, by shifting the focus to estimating each potential outcome mean separately, the GPS evaluated at each corresponding treatment level serves as a scalar balancing score. Future research work may involve extending the current balance measure to settings with longitudinal or time-to-event outcomes (Tang et al. 2022), as well as to predictive mean matching (Yang and Kim 2020) in missing data, where the predictive mean function requires estimation and model selection.

CHAPTER

3

DOUBLE SCORE MATCHING IN OBSERVATIONAL STUDIES WITH MULTI-LEVEL TREATMENTS

3.1 Introduction

Observational data contain rich information for drawing causal conclusions. The propensity score (Rosenbaum and Rubin 1983), which is defined as the conditional probability of receiving treatment given covariates, plays a central role in removing confounding bias. Incorporating this notion, standard techniques such as matching, subclassification, weighting, and regression adjustment (Imbens and Rubin 2015) have been proposed to estimate the average treatment effect (ATE).

While the main body of literature focuses on estimating the ATE for a binary treatment variable, the standard techniques have been modified and extended to estimate the pairwise average treatment effects for more than two treatment levels (Linden et al. 2016). For instance, regression adjustment may be directly adapted to model the conditional mean of

each potential outcome. By replacing the propensity score with the generalized propensity score (GPS), weighting estimators such as inverse probability weighting (IPW) (Horvitz and Thompson 1952) and its augmented version (AIPW) (Robins et al. 1995; Bang and Robins 2005) can also be applied for estimation of the pairwise ATEs (Imbens 2000). In the case of matching and subclassification, direct extension to multiple treatments was thought to be infeasible due to the curse of dimensionality until Yang et al. (2016), under which Zhao and Yang (2022) provided GPS model selection strategies. For estimation of the conditional average treatment effect, Liang and Yu (2022) proposed a flexible semiparametric approach for modeling the treatment contrasts, which also allows for extension to multiple treatment levels.

In the current literature, the augmented inverse probability weighting (AIPW) estimator is perhaps the most attractive due to its *double-robustness* property, which guarantees that it consistently estimates the treatment effect if either the propensity score model or the outcome mean function is correctly specified. However, like other weighting estimators, it inevitably suffers from high variability. This is because the estimator is computed by dividing the generalized propensity scores, which are often close to zero or one. The problem exacerbates when the number of treatment levels is large. Generalized propensity score matching, on the other hand, is more robust to extreme values of the GPS than weighting. Matching is also intuitively appealing, since it seeks to recreate a randomized experiment by balancing distributions of the pre-treatment variables between treatment groups. Moreover, matching can be viewed as a special case of hot deck imputation, which admits inference procedures for more general population parameters such as quantiles.

Just as in the binary treatment setting, matching on all covariates is not an attractive procedure in the multi-level treatment setting if the number of covariates is substantial (Abadie and Imbens 2006; Imbens and Rubin 2015; Imai and Van Dyk 2004). When the dimension of covariates is large, matching on the generalized propensity score provides an effective way in reducing bias inflicted by the curse of dimensionality (Abadie and Imbens 2006, 2016). Another dimension reduction technique is called the prognostic score (Hansen 2008). Prognostic score matching seeks to balance the potential outcome distributions between the treatment groups, and is more advantageous than propensity score matching when the (generalized) propensity score distributions are strongly separated (Wyss et al. 2015; Kumamaru et al. 2016).

In practice, however, GPS matching is not without its caveats, as it suffers from generalized propensity score model dependence (King and Nielsen 2019). In settings with two treatment levels, the idea of matching jointly on propensity and prognostic scores, or

so-called double score matching (DSM), was coined by Antonelli et al. (2018). The DSM estimator also enjoys the double robustness property and therefore alleviates concerns delineated by King and Nielsen (2019). Leacy and Stuart (2014) showed empirically that DSM improves the treatment effect estimation. Yang and Zhang (2023) further formalized DSM as a robust inferential procedure for general causal estimands.

In this paper, we extend the recipe offered by Yang and Zhang (2023) to the multi-level treatment setting. Specifically, we propose a new DSM estimator for the pairwise average treatment effect based on both the generalized propensity score and the generalized prognostic score (GPGS). We estimate the average potential outcomes separately for each treatment level, which only requires adjusting for the generalized propensity score and the generalized prognostic score for that treatment level. To correct for bias due to matching discrepancy, we propose a de-biased DSM estimator. Under certain regularity conditions, we show that the de-biased DSM estimator is doubly robust, and establish its consistency and asymptotic normality based on the true double scores, as well as on the estimated scores. This implies that matching on the double score is more robust to GPS model misspecfication than the GPSM estimator. In finite samples, DSM shows superior robustness to extreme values of the GPS than the AIPW.

The remaining article proceeds as follows. In Section 2, we lay out the multi-level treatment set-up, and review the definition of the generalized propensity score matching estimator. We introduce the double score matching algorithm in Section 3. Asymptotic properties of the DSM estimator and a variance estimation procedure are discussed in Section 4. We carry out a simulation study to assess the finite sample performance of the double score matching estimator in Section 5. In Section 6, we analyze the tutoring data using our proposed method.

3.2 Background

3.2.1 Setup, notation, and assumptions

Following the potential outcomes framework (Rubin 1974; Rosenbaum and Rubin 1983), we consider an observational study with more than two unordered treatment levels. Let treatment assignment be denoted by $W_i \in \mathbb{W} = \{1, \ldots, T\}$. In the standard binary treatment case $T = 2$, the two treatments are often labeled treatment and control. For each unit i there are T potential outcomes, one for each treatment level, denoted by $Y_i(w)$, for $w \in \mathbb{W}$. Implicit in this notation is the assumption that there is no interference

between units and no different versions of each treatment level (the stable-unit-treatment-value assumption, or SUTVA, (Rubin 1978)). The observed outcome for unit i is the potential outcome corresponding to the treatment that the unit received: $Y_i = Y_i(W_i)$. We also observe a vector-valued covariate or pre-treatment variable, denoted by X_i. We assume the sequence $\{W_1, X_1, Y_1(1), \ldots, Y_1(T)\}, \ldots, \{W_n, X_n, Y_n(1), \ldots, Y_n(T)\}$ with the potential outcomes is independent and identically distributed (i.i.d.), so that the sequence of observed values $(W_1, X_1, Y_1), \ldots, (W_n, X_n, Y_n)$ is also i.i.d. Various causal estimands can be of interest. Here, we focus on the problem of estimating the pairwise average treatment effect between treatment levels w and w':

$$\tau(w, w') = \mathbb{E}\{Y(w') - Y(w)\}. \tag{3.1}$$

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

$$\mu(w \mid V) = \mathbb{E}\{Y(w) \mid V\}, \sigma^2(w \mid V) = \mathbb{V}\{Y(w) \mid V\},$$

where $\mu(w \mid V)$ is an outcome mean function, $\sigma^2(w \mid V)$ is a variance function.

In the binary treatment case, Rosenbaum and Rubin (1983) defined the propensity score as the conditional probability of receiving the active treatment: $p(x) = \mathbb{P}(W = 1 \mid X = x)$. Imbens (2000) defined the generalized propensity score as the following:

Definition 2 (Generalized propensity score) *The generalized propensity score is the conditional probability of receiving each treatment level:*

$$p(w \mid x) = \mathbb{P}(W = w \mid X = x).$$

Throughout the remaining article, we make the following two standard assumptions.

Assumption 5 (Overlap) *For all values of x the probability of receiving any level of the treatment is positive:*

$$p(w \mid x) > 0 \quad \textit{for all } w, x.$$

Assumption 5 ensures that there is sufficient overlap between the covariate distributions of any two treatment levels. Without this assumption, there will be values of x for which we cannot estimate the pairwise average treatment effect without relying on extrapolation.

Define the treatment indicator variable $D(w) \in \{0, 1\}$ as $D(w) = 1$ if $W = w$, and $D(w) = 0$ otherwise. We also assume weak unconfoundedness (Imbens 2000).

Assumption 6 (Weak unconfoundedness) *For all* $w \in \mathbb{W}$,

$$D(w) \perp\!\!\!\perp Y(w) \mid X,$$

where "$\perp\!\!\!\perp$*" means "independent of".*

Assumption 6 holds if X contains all prognostic factors that affect the treatment assignment. It has no testable implications, but can be made more plausible by collecting detailed information on individual characteristics that are related to treatment assignment and outcome.

Under Assumptions 5 and 6, weak unconfoundedness is preserved if we condition on the generalized propensity score instead.

Lemma 3 (Weak unconfoundedness given the GPS) *Suppose the assignment mechanism is weakly unconfounded. Then for all* $w \in \mathbb{W}$,

$$D(w) \perp\!\!\!\perp Y(w) \mid p(w \mid X).$$

Lemma 3 is the key result. It implies that weak unconfoundedness is sufficient for identifying the average potential outcome at treatment level w by conditioning on the generalized propensity score evaluated at the same treatment level w. That is, we have

$$\mathbb{E}\{Y(w)\} = \mathbb{E}\Big[\mathbb{E}\{Y(w) \mid p(w \mid X)\}\Big] = \mathbb{E}\Big[\mathbb{E}\{Y \mid W = w, p(w \mid X)\}\Big]. \tag{3.2}$$

This in turn leads to the identification of the pairwise ATE.

Lemma 4 *Suppose the assignment mechanism is weakly unconfounded. Then*

$$\begin{aligned}\mathbb{E}\{Y(w') - Y(w)\} &= \mathbb{E}\{Y(w')\} - \mathbb{E}\{Y(w)\} \\ &= \mathbb{E}\Big[\mathbb{E}\{Y \mid W = w', p(w' \mid X)\}\Big] - \mathbb{E}\Big[\mathbb{E}\{Y \mid W = w, p(w \mid X)\}\Big].\end{aligned}$$

Proofs for Lemma 3 and Lemma 4 can be found in Imbens (2000).

3.2.2 Matching on the generalized propensity score

In this section, we briefly review the generalized propensity score matching estimator and introduce more relevant notation. Generally speaking, matching estimators hinge on

imputing the missing potential outcomes for each unit. Throughout, we consider matching with replacement and with the number of matches for each treatment level fixed at M. To be more specific, for each unit i, the potential outcome under W_i is the observed outcome Y_i; its (counterfactual) potential outcome under a different treatment level $w \neq W_i$ is not observed but can be imputed by the average of the observed outcomes of the M nearest units from treatment level w. In practice, it is common to let $M = 1$, in which case matching becomes nearest neighbor imputation (Little and Rubin 2014; Chen and Shao 2000, 2001).

First, consider a generic variable V as the matching variable. We let $\mathcal{J}_{w,i}^V$ denote the index set of the units from treatment level w that are matched to unit i and $K_i^V = \sum_{l=1}^n \mathbf{1}(i \in \mathcal{J}_{W_i,l}^V)$ denote the number of times unit i is used as a match, where the superscript "V" in $\mathcal{J}_{w,i}^V$ and K_i^V indicates the name of the matching variable. We require that $|\mathcal{J}_{w,i}^V| = M$ if $W_i \neq w$, and $|\mathcal{J}_{w,i}^V| = 0$ if $W_i = w$. Without loss of generality, we use the Euclidean distance as the metric to determine neighbors; the discussion applies to other distances (Abadie and Imbens 2006). Following Yang et al. (2016), GPS matching first imputes the missing potential outcomes under each treatment level w separately by matching on the GPS $V = p(w \mid X)$ evaluated at w. For simplicity, we write $p(w \mid X)$ as p_w. The imputed potential outcome for unit i is $\hat{Y}_i(w) = Y_i$ if $W_i = w$, and $\hat{Y}_i(w) = M^{-1} \sum_{j \in \mathcal{J}_{w,i}^{p_w}} Y_j$ if $W_i \neq w$. Then, the GPS matching estimator of the pairwise average treatment effect between w' and w is

$$
\begin{aligned}
\hat{\tau}_{\text{gps}}(w, w') &= n^{-1} \sum_{i=1}^n \{\hat{Y}(w') - \hat{Y}(w)\} \\
&= n^{-1} \sum_{i=1}^n \left[D_i(w') Y_i + \{1 - D_i(w')\} M^{-1} \sum_{j \in \mathcal{J}_{w',i}^{p_{w'}}} Y_j \right] \\
&\quad - n^{-1} \sum_{i=1}^n \left[D_i(w) Y_i + \{1 - D_i(w)\} M^{-1} \sum_{j \in \mathcal{J}_{w,i}^{p_w}} Y_j \right] \qquad (3.3) \\
&= n^{-1} \sum_{i=1}^n D_i(w')(1 + M^{-1} K_i^{p_{w'}}) Y_i - n^{-1} \sum_{i=1}^n D_i(w)(1 + M^{-1} K_i^{p_w}) Y_i. \qquad (3.4)
\end{aligned}
$$

Notice the equivalent definition of the GPS matching estimator by using the average of the imputed potential outcomes (3.3) and the weighted average of the observed outcomes (3.4). Table 3.1 illustrates the GPS matching scheme when $T = 3$.

Table 3.1: Matching scheme to impute missing potential outcomes when $T = 3$; for estimating $\mathbb{E}\{Y(w)\}$ at each level, the average of the imputed column is equal to the average of the weighted column.

	observed data			$Y(1)$		$Y(2)$		$Y(3)$	
index	X	W	Y	imputed	weighted	imputed	weighted	imputed	weighted
1	X_1	1	Y_1	Y_1	$[1 + M^{-1}K_1^{p_1}]Y_1$	$M^{-1}\sum_{l\in\mathcal{J}_{2,1}^{p_2}} Y_l$	0	$M^{-1}\sum_{l\in\mathcal{J}_{3,1}^{p_3}} Y_l$	0
2	X_2	1	Y_2	Y_2	$[1 + M^{-1}K_2^{p_1}]Y_2$	$M^{-1}\sum_{l\in\mathcal{J}_{2,2}^{p_2}} Y_l$	0	$M^{-1}\sum_{l\in\mathcal{J}_{3,2}^{p_3}} Y_l$	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	X_9	2	Y_9	$M^{-1}\sum_{l\in\mathcal{J}_{1,9}^{p_1}} Y_l$	0	Y_9	$[1 + M^{-1}K_9^{p_2}]Y_9$	$M^{-1}\sum_{l\in\mathcal{J}_{3,9}^{p_3}} Y_l$	0
10	X_{10}	2	Y_{10}	$M^{-1}\sum_{l\in\mathcal{J}_{1,10}^{p_1}} Y_l$	0	Y_{10}	$[1 + M^{-1}K_{10}^{p_2}]Y_{10}$	$M^{-1}\sum_{l\in\mathcal{J}_{3,10}^{p_3}} Y_l$	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	X_{n-1}	3	Y_{n-1}	$M^{-1}\sum_{l\in\mathcal{J}_{1,n-1}^{p_1}} Y_l$	0	$M^{-1}\sum_{l\in\mathcal{J}_{2,n-1}^{p_2}} Y_l$	0	Y_{n-1}	$[1 + M^{-1}K_{n-1}^{p_3}]Y_{n-1}$
n	X_n	3	Y_n	$M^{-1}\sum_{l\in\mathcal{J}_{1,n}^{p_1}} Y_l$	0	$M^{-1}\sum_{l\in\mathcal{J}_{2,n}^{p_2}} Y_l$	0	Y_n	$[1 + M^{-1}K_n^{p_3}]Y_n$

In practice, the generalized propensity score is unknown, and requires model specification and estimation from the observed data. To protect against misspecification of the propensity score model, double score matching is an attractive alternative to propensity score matching in the binary treatment case (Yang and Zhang 2023). In the next section, we propose a new double robust DSM estimator for the multi-level treatment case.

3.3 New Double Score Matching Estimator

Hansen (2008) introduced the notion of the prognostic score. To generalize this notion to the multi-level treatment setting, we define the generalized prognostic score as follows.

Definition 3 (Generalized prognostic score) *The generalized prognostic score is* $\Psi(X) = \{\Psi_1(X), \ldots, \Psi_T(X)\}$*, where* $\Psi_w(X)$ *is a sufficient statistic for* $Y(w)$ *in the sense that* $Y(w) \perp\!\!\!\perp X \mid \Psi_w(X)$ *for all* $w \in \mathbb{W}$*.*

Without loss of generality, we assume that $Y(w)$ follows a location-shift family $f_w(y - \mu_w(x))$ so that the conditional mean $\Psi_w(X) = \mu(w \mid X)$ is a generalized prognostic score.

Without further assumptions, Hansen (2008) showed that conditioning on the vector of prognostic scores is required to balance both potential outcome distributions between treatment and control. However, to separately estimate the average potential outcome for a specific treatment level, we no longer need to condition on the vector of generalized prognostic scores. For any w, conditioning on the generalized prognostic score evaluated at w balances the distribution of $Y(w)$ between treatment group w and all other treatment groups collectively.

Lemma 5 (Weak unconfoundedness given the GPGS) *Under Assumptions 1 and 2,* $Y(w) \perp\!\!\!\perp D(w) \mid \mu(w \mid X)$ *for all* $w \in \mathbb{W}$.

The proof is the same as for Lemma 3, except we use the balancing property of $\mu(w \mid X)$ instead of $p(w \mid X)$. Therefore, Lemma 5 implies that

$$\mathbb{E}\{Y(w)\} = \mathbb{E}\left[\mathbb{E}\{Y(w) \mid \mu(w \mid X)\}\right] = \mathbb{E}\left[\mathbb{E}\{Y \mid W = w, \mu(w \mid X)\}\right]. \tag{3.5}$$

Define the double score as $S(w \mid X) = \{p(w \mid X), \mu(w \mid X)\}$ for all $w \in \mathbb{W}$. Combining Lemma 3 and 4, we show that conditioning on the double score is sufficient for identification of the average potential outcome at a single treatment level, and therefore the pairwise ATE.

Lemma 6 (Weak unconfoundedness given the double score) *Under Assumptions 5 and 6, we have* $Y(w) \perp\!\!\!\perp D(w) \mid \{h(X), \mu(w \mid X)\}$, *and* $Y(w) \perp\!\!\!\perp D(w) \mid \{h(X), p(w \mid X)\}$ *for any* $h(X)$ *and* $w \in \mathbb{W}$.

The proof is similar to that of Theorem 1 in Antonelli et al. (2018), except we focus on the multiple treatment level case. Lemma 6 is the key result, which suggests that

$$\mathbb{E}\{Y(w)\} = \mathbb{E}\left[\mathbb{E}\{Y(w) \mid p(w \mid X), \mu(w \mid X)\}\right] = \mathbb{E}\left[\mathbb{E}\{Y \mid W = w, p(w \mid X), \mu(w \mid X)\}\right]. \tag{3.6}$$

Given this insight, we propose a new double score matching estimator. Let $\mu(w) = \mathbb{E}\{Y(w)\}$ denote the average potential outcome at treatment level w. To fix ideas, we again consider matching with replacement with the number of matches fixed at M. That is, if $W_i \neq w$ so that $Y_i(w)$ is unobserved, we select M nearest neighbors in treatment level w as matches for i. Let the matching variable be the double score $S(w \mid X) = \{p(w \mid X), \mu(w \mid X)\}$ or simply S_w. To stabilize numerical performance, it is desirable to standardize each component in S_w to have mean zero and variance one. Without loss of generality, we use the Euclidean distance to determine neighbors. Let $\mathcal{J}_{w,i}^{S_w}$ denote the index set of units from treatment level w that are matched to unit i, and $K_i^{S_{W_i}} = \sum_{j=1}^{n} \mathbf{1}(i \in \mathcal{J}_{W_i,j}^{S_{W_i}})$ be the number of times that unit i is used as a match. We focus on estimating $\mu(w)$ separately for $w = 1, ..., T$. Define the imputed potential outcome for unit i as $\hat{Y}_i(w) = Y_i$ if $W_i = w$, and $\hat{Y}_i(w) = M^{-1} \sum_{j \in \mathcal{J}_{w,i}^{S_w}} Y_j$ if $W_i \neq w$. Then an initial double score matching estimator of $\mu(w)$ is defined as

$$\hat{\mu}_{\text{dsm}}^{(0)}(w) = n^{-1} \sum_{i=1}^{n} \hat{Y}_i(w) = n^{-1} \sum_{i=1}^{n} D_i(w)(1 + M^{-1} K_i^{S_w}) Y_i.$$

Following Abadie and Imbens (2006), we obtain the following decomposition:

$$n^{1/2}\left\{\hat{\mu}^{(0)}_{\text{dsm}}(w) - \mu(w)\right\} = D_n(w) + B_n(w),$$

where

$$\begin{aligned} D_n(w) &= n^{-1/2}\sum_{i=1}^{n}\left[\mu(w \mid S_{w,i}) - \mu(w) + D_i(w)\left(1 + M^{-1}K_i^{S_w}\right)\{Y_i - \mu(w \mid S_{w,i})\}\right], \\ B_n(w) &= n^{-1/2}\sum_{i=1}^{n}\{D_i(w) - 1\}\left[M^{-1}\sum_{j\in\mathcal{J}^{S_w}_{w,i}}\{\mu(w \mid S_{w,i}) - \mu(w \mid S_{w,j})\}\right]. \end{aligned} \tag{3.7}$$

Because of (3.2) and (3.5), we have $\mathbb{E}\{\mu(w \mid S_{w,i})\} = \mu(w)$, so $D_n(w)$ is unbiased. The difference $\mu(w \mid S_{w,i}) - \mu(w \mid S_{w,j})$ in (3.7) accounts for the matching discrepancy, and therefore $B_n(w)$ contributes to the asymptotic bias of the matching estimator. In general, if S is k-dimensional, Abadie and Imbens (2006) showed that under certain regularity conditions, $||S_{w,i} - S_{w,j}|| = O_P(n^{-1/k})$ for $j \in \mathcal{J}^{S_w}_{w,i}$ and $W_j = w$. Then if $k \geq 2$, the matching discrepancy bias $B_n(w) = O_P(n^{1/2-1/k}) \neq o_P(1)$ is no longer negligible. To correct for this bias, let $\hat{\mu}(w \mid S_w)$ be a nonparametric estimator of $\mu(w \mid S_w)$. We propose a de-biased double score matching estimator of $\mu(w)$:

$$\hat{\mu}_{\text{dsm}}(w) = \hat{\mu}^{(0)}_{\text{dsm}}(w) - n^{-1/2}\hat{B}_n(w), \tag{3.8}$$

where $\hat{B}_n(w)$ is an estimator of $B_n(w)$ by replacing $\mu(w \mid S_w)$ with $\hat{\mu}(w \mid S_w)$.

Because S_w is unknown in practice, we posit working models for both scores in S_w, for all treatment levels w.

Assumption 7 *The parametric model $p(w \mid X; \alpha)$ is a correct specification for $p(w \mid X)$; i.e., $p(w \mid X) = p(w \mid X; \alpha_0)$, where α_0 is the true model parameter.*

Assumption 8 *Suppose that $Y(w)$ follows a location-shift family and that the parametric model $\mu(w \mid X; \beta_w)$ is a correct specification for $\mu(w \mid X)$; i.e., $\mu(w \mid X) = \mu(w \mid X; \beta_{w,0})$ where $\beta_{w,0}$ is the true model parameter.*

Consistency and asymptotic normality of the de-biased DSM estimator only require that one of the two above assumptions is satisfied.

We summarize the double score matching algorithm in the following steps.

Step 1. Posit parametric models for $p(w \mid X)$ and $\mu(w \mid X)$, denoted by $p(w \mid X; \alpha_0)$ and $\mu(w \mid X; \beta_{w,0})$ with fixed unknown parameters α_0 and $\beta_{w,0}$. Let $\theta = (\alpha^{\mathrm{T}}, \beta_1^{\mathrm{T}}, \ldots, \beta_T^{\mathrm{T}})^{\mathrm{T}}$ denote the vector of nuisance parameters. Obtain an estimator $\hat{\theta} = (\hat{\alpha}^{\mathrm{T}}, \hat{\beta}_1^{\mathrm{T}}, \ldots, \hat{\beta}_T^{\mathrm{T}})^{\mathrm{T}}$ under the posited models based on the observed data. For each unit i, calculate $\hat{S}_{w,i} = S(w \mid X_i; \hat{\theta}) = \{p(w \mid X_i; \hat{\alpha}), \mu(w \mid X_i; \hat{\beta}_w)\}$. To stabilize the numerical performance, it is desirable to standardize $\hat{S}_{w,i}$ such that each component has mean zero and variance one.

Step 2. For each unit i with treatment $W_i \neq w$, find M nearest neighbors from the treatment group w based on the matching variable $\hat{S}_{w,i}$. Include their indices in $\mathcal{J}_{w,i}^{S_w}$.

Step 3. Obtain a nonparametric estimator of $\mu(w \mid S_w)$, denoted by $\hat{\mu}(w \mid S_w)$, e.g. by the method of sieves (Chen 2007) based on $[\{Y_i, \hat{S}_{w,i}\} : W_i = w]$.

Step 4. Compute the de-biased double score matching estimator of $\mu(w)$ given by (3.8) with $S_{w,i}$ replaced by $\hat{S}_{w,i}$. We denote the final estimator of $\mu(w)$ as $\hat{\mu}_{\mathrm{dsm}}(w; \hat{\theta})$ to reflect its dependence on the estimated nuisance parameters $\hat{\theta}$. Repeat the previous steps for w'. The double score matching estimator of $\tau(w, w')$ is $\hat{\tau}_{\mathrm{dsm}}(w, w'; \hat{\theta}) = \hat{\mu}_{\mathrm{dsm}}(w'; \hat{\theta}) - \hat{\mu}_{\mathrm{dsm}}(w; \hat{\theta})$.

3.4 Main Results

In this section we derive asymptotic results of the double score estimator of the pairwise ATE $\tau(w, w')$, for both known and estimated double scores. We also provide a variance estimation procedure based on weighted bootstrap.

3.4.1 Asymptotic results

Let θ^* be the probability limit of $\hat{\theta}$. Under Assumption 2, we have $\alpha^* = \alpha_0$ and under Assumption 3, we have $\beta_w^* = \beta_{w,0}$. First, we consider matching on the true double score $S(w|X; \theta^*) = \{p(w|X; \alpha^*), \mu(w|X; \beta_w^*)\}$. To study the asymptotic properties of $\hat{\tau}_{\mathrm{dsm}}(w, w'; \theta^*)$, we require some regularity assumptions. For simplicity, let $S_w = S(w|X; \theta^*)$ and let $f_{w'}(S_w)$ be the conditional density of S_w given $W = w'$.

Assumption 9 *For $w = 1, ..., T$, (i) The matching variable S_w has a compact and convex support $\mathcal{S}$, with a continuous density bounded and bounded away from zero: there exist*

constants C_{1L} *and* C_{1U} *such that* $C_{1L} \leq f_{w'}(S_w)/f_w(S_w) \leq C_{1U}$ *for any* $S_w \in \mathcal{S}$*; (ii)* $\mu(w|S_w)$ *and* $\sigma^2(w|S_w)$ *satisfy the Lipschitz continuity condition; and (iii) there exists* $\delta > 0$ *such that* $\mathbb{E}\{|Y(w)|^{2+\delta}|S_w\}$ *is uniformly bounded for any* $S_w \in \mathcal{S}$*.*

Assumption 9 has been discussed by Abadie and Imbens (2006) and Abadie and Imbens (2016) for matching estimators based on the covariates and the propensity score. Part (i) is a convenient regularity condition. (ii) imposes smoothness conditions for the outcome mean function and the variance function. (iii) is a moment condition required to invoke the central limit theorem. The following theorem establishes the double robustness and asymptotic distribution of $\hat{\tau}_{\text{dsm}}(w, w'; \theta^*)$.

Theorem 3 *Under Assumptions 5, 6, and 9, if either Assumption 7 or Assumption 8 is satisfied, we have*

$$n^{1/2}\left\{\hat{\tau}_{\text{dsm}}(w, w'; \theta^*) - \tau(w, w')\right\} \xrightarrow{d} \mathcal{N}\left(0, V_\tau\right),$$

where

$$\begin{aligned} V_\tau =& \mathbb{E}\left[\left\{\mu(w'|S_{w'}) - \mu(w|S_w) - \tau(w, w')\right\}^2\right] \\ &+ \mathbb{E}\left(\sigma^2(w'|S_{w'})\left[\frac{1}{p(w' \mid X; \alpha^*)} + \frac{1}{2M}\left\{\frac{1}{p(w' \mid X; \alpha^*)} - p(w' \mid X; \alpha^*)\right\}\right]\right) \\ &+ \mathbb{E}\left(\sigma^2(w|S_w)\left[\frac{1}{p(w \mid X; \alpha^*)} + \frac{1}{2M}\left\{\frac{1}{p(w \mid X; \alpha^*)} - p(w \mid X; \alpha^*)\right\}\right]\right). \end{aligned}$$

Theorem 3 states that the consistency and asymptotic normality of the DSM estimator is guaranteed if either the posited generalized propensity score or the prognostic score model is correctly specified. As we will show empirically in the next section, when both scores are correctly specified, the DSM estimator is more efficient than the AIPW when generalized propensity scores are close to 0 or 1.

Next, we consider matching on the estimated double score $S(w|X_i; \hat{\theta}) = \{p(w|X_i; \hat{\alpha}), \mu(w|X_i; \hat{\beta}_w)\}$. Let the log-likelihood function for α be

$$L(\alpha|W_1, X_1, \cdots, W_n, X_n) = \sum_{i=1}^{n}\sum_{w=1}^{T} D_i(w)\log\{p(w|X_i; \alpha)\},$$

where $D_i(T) = 1 - \sum_{w=1}^{T-1} D_i(w)$ and $p(T|X_i; \alpha) = 1 - \sum_{w=1}^{T-1} p(w|X_i; \alpha)$. To obtain $\hat{\tau}_{\text{dsm}}(w, w'; \hat{\theta})$, under the posited parametric models $p(w \mid X; \alpha_0)$ and $\mu(w \mid X; \beta_{w,0})$, we

obtain an estimator $(\hat{\alpha}^{\mathrm{T}}, \hat{\beta}_w^{\mathrm{T}}, \hat{\beta}_{w'}^{\mathrm{T}})^{\mathrm{T}}$ by solving the estimating equation

$$\mathcal{U}_n(\theta) = n^{-1/2} \sum_{i=1}^{n} U\left(W_i, X_i, Y_i; \theta\right) = n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} U_1\left(W_i, X_i; \alpha\right) \\ U_2\left(D_i(w), X_i, Y_i; \beta_w\right) \\ U_3\left(D_i(w'), X_i, Y_i; \beta_{w'}\right) \end{pmatrix} = 0,$$

where

$$\begin{aligned} U_1(W, X; \alpha) &= \frac{\partial}{\partial \alpha} \sum_{w=1}^{T} D(w) \log\{p(w|X; \alpha)\}, \\ U_2\left(D(w), X, Y; \beta_w\right) &= D(w) \frac{\partial \mu\left(w|X; \beta_w\right)}{\partial \beta_w} \left\{Y - \mu\left(w|X; \beta_w\right)\right\}, \\ U_3\left(D(w'), X, Y; \beta_{w'}\right) &= D(w') \frac{\partial \mu\left(w'|X; \beta_{w'}\right)}{\partial \beta_{w'}} \left\{Y - \mu\left(w'|X; \beta_{w'}\right)\right\}. \end{aligned}$$

Theorem 4 *Under Assumption 5, 6, and 9, and regularity conditions specified in the supplementary material, if either Assumption 7 or Assumption 8 holds, the approximate distribution of* $n^{1/2}\left\{\hat{\tau}_{\mathrm{dsm}}(w, w'; \hat{\theta}) - \tau(w, w')\right\}$ *is* $\mathcal{N}\left(0, V_{\tau,\mathrm{adj}}\right)$*, where*

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

where Σ_U *is the information matrix,* $\Sigma_{\theta^*} = \Gamma_{\theta^*}^{-1} \Sigma_U \left(\Gamma_{\theta^*}^{-1}\right)^{\mathrm{T}}$, $\Gamma_{\theta^*} = \mathbb{E}\left\{\partial U(W, X, Y; \theta^*)/\partial \theta^{\mathrm{T}}\right\}$, $\gamma_1^{\mathrm{T}} = \left(\gamma_{11}^{\mathrm{T}}, \gamma_{12}^{\mathrm{T}}, \gamma_{13}^{\mathrm{T}}\right)$,

$$\begin{aligned} \gamma_{11} =& \mathbb{E}\left[\left\{\mu\left(w'|S_{w'}\right) - \mu\left(w|S_w\right) - \tau(w, w')\right\} U_1(W, X; \alpha^*)\right] \\ &+ \mathbb{E}\left[\left\{\mu(w'|X) - \mu(w'|S_{w'})\right\} U_1(W, X; \alpha^*)\right] \\ &- \mathbb{E}\left[\left\{\mu(w|X) - \mu(w|S_w)\right\} U_1(W, X; \alpha^*)\right], \end{aligned}$$

$$\begin{aligned} &\gamma_{12} \\ =& - \mathbb{E}\left[\left\{\mu(w'|S_{w'}) - \mu(w|S_w) - \tau(w, w')\right\} D(w) \frac{\partial \mu\left(w|X; \beta_w^*\right)}{\partial \beta_w} \left\{\mu(w|X) - \mu\left(w|X; \beta_w^*\right)\right\}\right] \\ &- \mathbb{E}\left[\left\{\mu(w|X) - \mu(w|S_w)\right\} \frac{\partial \mu\left(w|X; \beta_w^*\right)}{\partial \beta_w} \left\{\mu(w|X) - \mu\left(w|X; \beta_w^*\right)\right\}\right] \\ &- \mathbb{E}\left\{\frac{\partial \mu\left(w|X; \beta_w^*\right)}{\partial \beta_w} \sigma^2(w|X)\right\}, \end{aligned}$$

$$
\begin{aligned}
&\gamma_{13} \\
&= -\mathbb{E}\left[\{\mu(w'|S_{w'}) - \mu(w|S_w) - \tau(w,w')\}\right. \\
&\quad \left. D(w')\frac{\partial \mu\left(w'|X;\beta^*_{w'}\right)}{\partial \beta_{w'}}\left\{\mu(w'|X) - \mu\left(w'|X;\beta^*_{w'}\right)\right\}\right] \\
&\quad -\mathbb{E}\left[\left\{\mu(w'|X) - \mu(w'|S_{w'})\right\}\frac{\partial \mu\left(w'|X;\beta^*_{w'}\right)}{\partial \beta_{w'}}\left\{\mu(w'|X) - \mu\left(w'|X;\beta^*_{w'}\right)\right\}\right] \\
&\quad -\mathbb{E}\left\{\frac{\partial \mu\left(w'|X;\beta^*_{w'}\right)}{\partial \beta_{w'}}\sigma^2(w|X)\right\},
\end{aligned}
$$

and $\gamma_2^{\mathrm{T}} = \left(\gamma_{21}^{\mathrm{T}}, \gamma_{22}^{\mathrm{T}}, \gamma_{23}^{\mathrm{T}}\right)$,

$$
\begin{aligned}
\gamma_{21} &= -\mathbb{E}\left\{\left[\frac{D(w')\left\{Y - \mu\left(w'|X;\beta^*_{w'}\right)\right\}}{p(w \mid X;\alpha^*)^2}\right.\right. \\
&\qquad \left.\left. + \frac{(1 - D(w'))\left\{Y - \mu\left(w|X;\beta^*_{w}\right)\right\}}{\left\{1 - p(w \mid X;\alpha^*)\right\}^2}\right]\frac{\partial p(w \mid X;\alpha^*)}{\partial \alpha}\right\}, \\
\gamma_{22} &= \mathbb{E}\left\{\frac{D(w) - p(w \mid X;\alpha^*)}{p(w \mid X;\alpha^*)}\frac{\partial \mu\left(w|X;\beta^*_{w}\right)}{\partial \beta_{w}}\right\}, \\
\gamma_{23} &= \mathbb{E}\left\{\frac{p(w' \mid X;\alpha^*) - D(w')}{p(w'|X;\alpha^*)}\frac{\partial \mu\left(w'|X;\beta^*_{w'}\right)}{\partial \beta_{w'}}\right\}.
\end{aligned}
$$

We discuss the implication of estimating the nuisance parameters on the matching estimator. In the binary treatment case, Abadie and Imbens (2016) showed that matching on the estimated propensity score always improves efficiency compared to matching on the true propensity score, due to the correlation between the matching estimator and the score function of the nuisance parameters. However, for DSM, the difference between the two asymptotic variances in Theorem 3 and 4 can either be positive, negative, or zero. This indicates that matching on the estimated double score can either increase, decrease, or maintain the efficiency of estimation compared to matching on the true double score. Specifically, the variance reduction term $-\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1$ stems from the fact that the matching estimator and the score function for the double score nuisance parameters are correlated. The term $\gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2$ inflates the variance if either the prognostic score or the generalized propensity score model is misspecified.

3.4.2 Resampling-based variance estimation

The conventional bootstrap replication variance estimator is invalid for the matching estimator (Chen and Shao 2001; Otsu and Rai 2017). Because of the lack of smoothness for

matching, the non-parametric bootstrap cannot preserve the distribution of the number of times that each unit is used as a match. Otsu and Rai (2017) proposed a wild bootstrap procedure for the matching estimator when matching is directly based on the covariates. Theoretical work on wild bootstrap variance estimation has been developed in the case of matching on the prognostic score (Yang and Kim 2020) and on the propensity score (Adusumilli 2018).

The double score estimator has a two-stage estimation procedure: estimation of the two scores and matching. To take into account all sources of variability, we use a parallel two-stage replication variance estimation procedure. First, we construct replicates of the nuisance parameter estimators in the double score function. Second, based on the asymptotic linearization of the matching estimators, we construct replicates of the matching estimators directly based on the linear terms with the replicated nuisance parameters (Yang and Kim 2020). In this way, the distribution of the number of times that each unit is used as a match can be retained. An outline of the replicate variance estimation procedure is given below.

VE-Step 1. Obtain a bootstrap sample, or equivalently the bootstrap replication weights $w_i^* = n^{-1}m_i^*$ with $(m_1^*, \cdots, m_n^*)$ is a multinomial random vector with n draws on n equal probability cells. Obtain a bootstrap replicate of $\hat{\theta}$, $\hat{\theta}^* = (\hat{\alpha}^{*\mathrm{T}}, \hat{\beta}_1^{*\mathrm{T}}, \ldots, \hat{\beta}_T^{*\mathrm{T}})^{\mathrm{T}}$ under the posited models with the replication weights. For each unit i, calculate $\hat{S}_{w,i}^* = S(w \mid X_i; \hat{\theta}^*) = \{p(w \mid X_i; \hat{\alpha}^*), \mu(w \mid X_i; \hat{\beta}_w^*)\}^{\mathrm{T}}$.

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

$$
\begin{aligned}
\hat{\tau}_{\mathrm{dsm}}^*(w, w'; \hat{\theta}^*) =& n^{-1} \sum_{i=1}^{n} w_i^* \{\hat{\mu}(w' | \hat{S}_{w',i}^*) - \hat{\mu}(w | \hat{S}_{w,i}^*)\} \\
&+ n^{-1} \sum_{i=1}^{n} w_i^* D_i(w')(1 + M^{-1} K_i^{\hat{S}_{w'}}) \{Y_i - \hat{\mu}(w' | \hat{S}_{w',i}^*)\} \\
&- n^{-1} \sum_{i=1}^{n} w_i^* D_i(w)(1 + M^{-1} K_i^{\hat{S}_w}) \{Y_i - \hat{\mu}(w | \hat{S}_{w,i}^*)\}.
\end{aligned}
$$

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

3.5 A Simulation Study

In this section we assess the performance of the new estimator in a Monte Carlo study relative to several previously proposed estimators: (i) the simple difference in average outcomes by treatment status (`naive`), (ii) matching on all covariates (`m.x`), (iii) the generalized propensity score estimator based on matching on each level of the generalized propensity scores (`gpsm`), (iv) the estimator based on matching on each level of the prognostic score (`m.prog`), (v) the proposed doubly robust estimator based on matching on each level of the propensity score and the prognostic score (`m.ds`), (vi) the estimator based on inverse propensity score weighting (`ipw`), and (vii) the augmented inverse propensity score weighting estimator (`aipw`).

We fix the number of matches M to be one. We let the number of treatment levels be three, and consider two designs: design (P1) produces propensity scores with a minimum in the scale of 10^{-2}; design (P2) produces propensity score with a minimum in the scale of 10^{-5}.

Under design (P1), the covariates X_{1i}, X_{2i}, and X_{3i} are multivariate normal with means zero, variances of $(2, 1, 1)$, and $Cov(X_{1i}, X_{2i}) = 1$, $Cov(X_{1i}, X_{3i}) = -1$, and $Cov(X_{2i}, X_{3i}) = -0.5$. We take $X_{4i} \sim U[-3, 3], X_{5i} \sim \chi^2_1$, and $X_{6i} \sim$ Bernoulli(0.5) to be independent from each other, and independent from (X_{1i}, X_{2i}, X_{3i}). The vector of covariates for individual i is $X_i^{\mathrm{T}} = (1, X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})$. Let a nonlinear transformation of X_i be $Z_i^{\mathrm{T}} = (1, Z_{1i}, \ldots, Z_{6i})$, where $Z_{1i} = X_{1i}^2$, $Z_{2i} = X_{2i}^2$, $Z_{3i} = \exp(X_{3i}/2)$, $Z_{4i} = \exp(X_{4i}/3)$, $Z_{5i} = \log(X_{5i})$, $Z_{6i} = X_{5i} \times X_{6i}$, which are further scaled and centered such that $\mathbb{E}(Z_i) = 0$ and $\mathbb{V}(Z_i) = 1$ for all i. The three treatment groups are generated using a multinomial regression model:

$$\{D_i(1), D_i(2), D_i(3)\} \sim \text{Multinom}\{n = 1, p(1|Z_i), p(2|Z_i), p(3|Z_i)\},$$

where $D_i(w)$ is the treatment indicator, i.e. $D_i(w) = 1$, if the unit i belongs to treatment w, and $p(w \mid Z_i) = \exp(Z_i^{\mathrm{T}}\alpha_{w,0}) / \sum_{w'=1}^{3} \exp(Z_i^{\mathrm{T}}\alpha_{w',0})$, where $\alpha_{1,0}^{\mathrm{T}} = (0, 0, 0, 0, 0, 0, 0)$, $\alpha_{2,0}^{\mathrm{T}} = 0.1 \times (0, 1, 1, 1, 1, 1, 1)$ and $\alpha_{3,0}^{\mathrm{T}} = -0.1 \times (0, 1, 1, 1, 1, 1, 1)$.

For design (P2), the covariates X_{1i}, X_{2i}, X_{3i}, and X_{4i} are multivariate normal with means zero, variances of $(1, 1, 1, 1)$ and covariance 0; with $X_i^{\mathrm{T}} = (1, X_{1i}, X_{2i}, X_{3i}, X_{4i})$. Let a nonlinear transformation of X_i be $Z_i^{\mathrm{T}} = (1, Z_{1i}, \ldots, Z_{4i})$, where $Z_{1i} = \exp(X_{1i}/2)$, $Z_{2i} = X_{2i}/(1 + \exp(X_{1i})) + 10$, $Z_{3i} = (X_{1i}X_{3i}/25 + 0.6)^3$, $Z_4 = (X_{2i} + X_{4i} + 20)^2$. The three treatment groups are generated using the same multinomial regression model, but

with different coefficients: $\alpha_{1,0}^{\mathrm{T}} = (0.419, 0, 0, 0, 0)$, $\alpha_{2,0}^{\mathrm{T}} = (0, -1, 0.5, -0.25, -0.1)$ and $\alpha_{3,0}^{\mathrm{T}} = -\alpha_{2,0}^{\mathrm{T}}$.

The outcome model is $Y_i(w) = Z_i^{\mathrm{T}}\beta_{w,0} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,1)$, and $\beta_{1,0}^{\mathrm{T}} = \beta_{2,0}^{\mathrm{T}} = \beta_{3,0}^{\mathrm{T}} = (0,1,1,1,1,1,1)/2$ for (P1), and $\beta_{1,0}^{\mathrm{T}} = \beta_{2,0}^{\mathrm{T}} = \beta_{3,0}^{\mathrm{T}} = (0,1,1,1,1)/2$ for (P2). The sample sizes are $N_w = 500$, for all w. Additional simulation results for small sample sizes ($N_1 = N_2 = N_3 = 100$), unequal sample sizes ($N_1 \neq N_2 \neq N_3$), and unequal mean potential outcomes ($\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$) are included in the Supplementary Material.

To assess the double robustness property of the double score matching estimator, we consider two model specifications for the generalized propensity score: a multinomial logistic regression model with the predictor Z (correct) and that with the predictor X (wrong). We also consider two model specifications for the prognostic score: $\beta_{w,0}^{\mathrm{T}}Z$ (correct) and $\beta_{w,0}^{\mathrm{T}}X$ (wrong). The matching estimators are implemented based on the steps given in Section 3.3: if the matching variable is X, we start with Step 2 and replace $S_i(\hat{\theta})$ with X_i; if the generalized propensity score is the matching variable, we start with Step 1 and replace $S_i(\hat{\theta})$ with $p(w|Z_i;\hat{\alpha})$. For all matching estimators, the conditional outcome mean functions are approximated using power series.

We compare four model specification scenarios (S1)–(S4) under (P1) in Figure 3.1 and those under (P2) in Figure 3.2. The `naive` estimator is biased for all ATEs. Matching directly based on all covariates X show biases for the ATEs across all four scenarios. The single score matching estimators and the `ipw` estimator are singly robust. Reliable performance of the GPS matching estimator and `ipw` estimator relies on a correct specification of the underlying score model, and matching on the prognostic score relies on a correct specification of the outcome model. The double score matching estimator and `aipw` estimator are doubly robust in that it has small biases for the ATE even if one of the scores is misspecified, as shown in scenarios (S2) and (S3). Under (P1), where the generalized propensity scores are not close to zero, the `aipw` estimator is more efficient than the double score matching estimator. However, under (P2), where the generalized propensity scores can be very close to zero, the inverse propensity score weights may suffer from larger variability. Then, the `ipw` and `aipw` estimator exhibit high variance when one or more inverse of the generalized propensity scores are enormous. Therefore, under (P2), the double score matching estimator appears to be more efficient than the `aipw` estimator. We also notice that when both models are misspecified, the `aipw` estimator performs worse than the double score matching estimator in terms of both bias and root mean square error. This implies that the `aipw` estimator is sensitive to model misspecification.

When both models are misspecified, estimation performance of the `aipw` estimator is even worse than using one single misspecified model (`ipw` estimator). In this scenario, our results indicate that the double score matching estimator is more robust than the `aipw` estimator.

Figure 3.1: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios under (P1) for the generalized propensity score (GPS) and generalized prognostic score (GPGS).

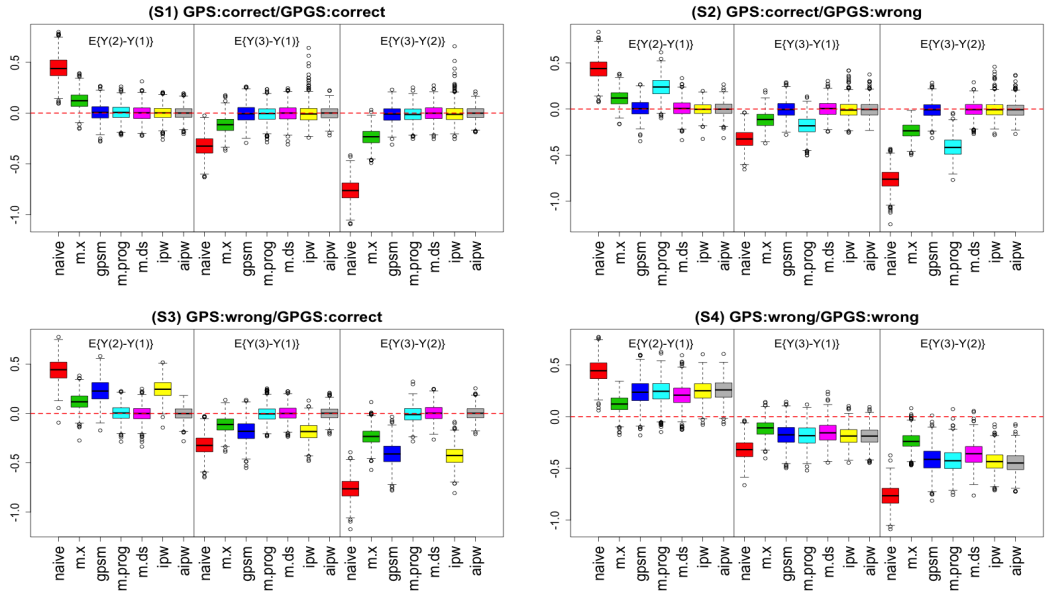


Figure 3.2: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios under (P2) for the generalized propensity score (GPS) and generalized prognostic score (GPGS).

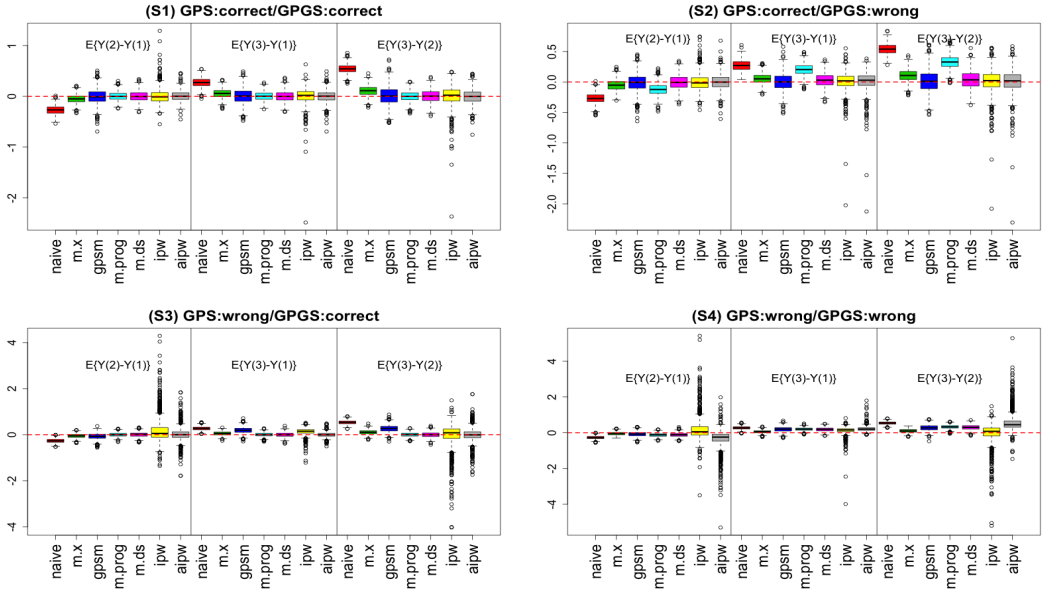


Table 3.2 reports the simulation results for the coverage probabilities for the proposed double score matching estimator using the proposed replication-based method. Under the double robustness condition (i.e., if either the propensity score model or the prognostic score model is correctly specified), for the proposed replication method, the coverage rates are close to the nominal coverage. When the propensity score model and the prognostic score model are both misspecified, the coverage rates fail to match the nominal coverage.

3.6 An Application

In this section, we apply the proposed double score matching method as well as existing methods mentioned in Section 5 to the *tutoring* dataset included in the TriMatch R package. The treatment takes one of Treat1, Treat2, and Control, which represent the type of tutoring service each student received. There are 918 controls, 134 subjects who received Treat1 and 90 subjects underwent Treat2. The outcome is course grade, which takes one of 0, 1, 2, 3, or 4. Pre-treatment variables include gender, ethnicity, military service status of the student, non-native English speaker status, education level of the subject's mother, education level of the subject's father, age of the student, employment

Table 3.2: Simulation results based on 1000 Monte Carlo simulated datasets for the coverage properties for the proposed double score matching estimators of the average treatment effects under four scenarios for the generalized propensity score (GPS) and generalized prognostic score (GPGS) models: empirical coverage rate and (empirical coverage rate ± 2×Monte Carlo standard error)

	P1		
	$\mathbb{E}\{Y(2)-Y(1)\}$	$\mathbb{E}\{Y(3)-Y(1)\}$	$\mathbb{E}\{Y(3)-Y(2)\}$
(S1) GPS:correct/GPGS:correct	95.1 (93.8,96.4)	96.6 (95.5,97.7)	95.8 (94.6,97.0)
(S2) GPS:correct/GPGS:wrong	97.0 (95.9,98.1)	95.6 (94.3,96.9)	96.6 (95.5,97.7)
(S3) GPS:wrong/GPGS:correct	95.2 (93.9,96.5)	96.0 (94.8,97.2)	96.2 (95.0,97.4)
(S4) GPS:wrong/GPGS:wrong	52.7 (49.6,55.8)	70.9 (68.1,73.7)	13.6 (11.5,15.7)
	P2		
	$\mathbb{E}\{Y(2)-Y(1)\}$	$\mathbb{E}\{Y(3)-Y(1)\}$	$\mathbb{E}\{Y(3)-Y(2)\}$
(S1) GPS:correct/GPGS:correct	95.7 (94.4,97.0)	93.8 (92.3,95.3)	94.0 (92.5,95.5)
(S2) GPS:correct/GPGS:wrong	94.9 (93.5,96.3)	92.9 (91.3,94.5)	93.6 (92.1,95.1)
(S3) GPS:wrong/GPGS:correct	94.2 (92.8,95.6)	93.7 (92.2,95.2)	93.8 (92.3,95.3)
(S4) GPS:wrong/GPGS:wrong	80.7 (78.3,83.1)	64.4 (61.4,67.4)	40.4 (37.4,43.4)

status (no, part-time, full-time), household income, number of transfer credits, and grade point average.

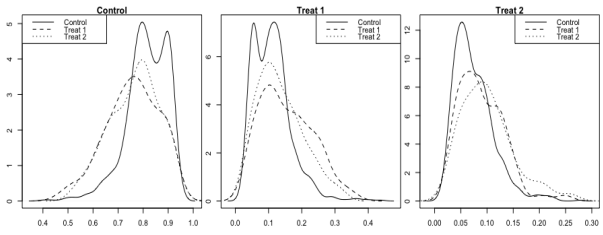
The prognostic scores are estimated by fitting a linear regression of all first-order terms of the covariates separately for the three treatment groups. Table 3.3 shows the R^2 and mean squared error (MSE) from the linear regressions for the three treatment groups. It turns out that the linear model performs poorly in fitting the outcome with the pre-treatment variables, suggesting that the pre-treatment variables can explain only a small portion of the outcome variability.

Table 3.3: Results of fitting a linear regression of all first-order terms of the covariates for the generalized prognostic score of each tutoring service.

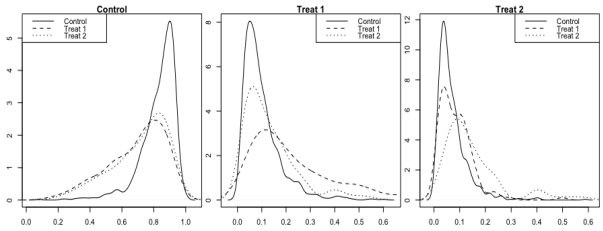
	Control	Treat 1	Treat 2
R^2	0.25	0.40	0.58
MSE	1.34	0.78	0.53

We fit two models for the propensity scores to compare the robustness of different estimators two different proposed generalized propensity score models. Model 1 is a multinomial logistic regression model that includes all first-order terms of the covariates.

Model 2 is a multinomial logistic regression model that includes first-order and second-order terms of the covariates with the Lasso penalty to select variables with different tuning parameters. Figure 3.3 shows the densities of the fitted generalized propensity scores based on the two models for each treatment group, which indicates that the overlap assumption likely holds true for this dataset.



(a) Generalized propensity scores estimated based on model 1



(b) Generalized propensity scores estimated based on model 2

Figure 3.3: Densities of fitted generalized propensity scores based on the two models.

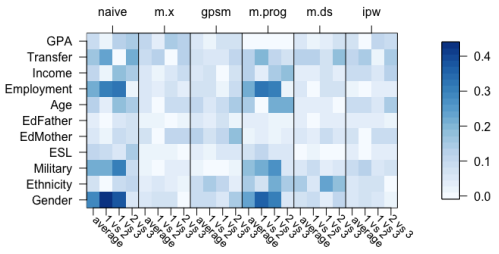
Table 3.4 shows results from fitting the two GPS models. The null deviance is 1432.44.

Table 3.4: Results of fitting two different GPS models for the three tutoring levels.

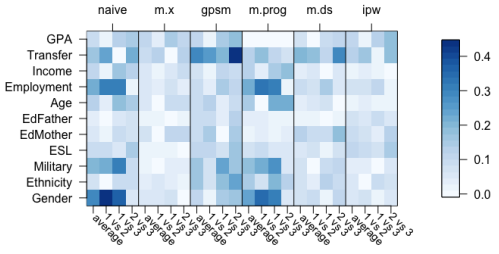
Model	Deviance	#Parameters	AUC
1	1363.28	24	0.63
2	1228.80	70	0.73

Figure 3.4 checks covariate balance before and after various adjustments. The colors represent the standardized differences in means of each covariate in the original dataset, data weighted after matching and data weighted by the inverse generalized propensity

scores.



(a) Generalized propensity scores estimated based on model 1



(b) Generalized Propensity scores estimated based on model 2

Figure 3.4: Standardized difference in means based on the two models.

The `naive` procedure (i.e. without any weighting or matching adjustment) shows significant imbalance in several covariates, suggesting that there exists large confounding bias. Matching on all covariates, double score matching and inverse propensity score weighting result in a standardized difference in means less than 0.2 for all covariates, effectively removing differences in the observed covariates among different treatment groups. Matching on the generalized propensity scores performs even worse for removing differences in covariates when the model deviance gets smaller. The reason is that few units receive extremely large weights. Prognostic score matching does not produce balanced data, due to the difficulty with positing a prognostic score model that fits the data well.

Table 3.5 shows the estimated ATEs and 95% Wald confidence intervals for the pairwise ATEs based on the seven estimators described in the previous section. Estimators with suffix $i = 1$ or 2 indicate that estimation is based on the generalized propensity score

Model i.

Table 3.5: Estimated ATEs of tutoring services and 95% Wald confidence intervals

	Treat1-Control	Treat2-Control	Treat2-Treat1
`naive`	0.39 (0.12,0.66)	0.70 (0.38,1.02)	0.31 (-0.06,0.68)
`m.x`	0.40 (0.26,0.53)	0.75 (0.64,0.86)	0.35 (0.25,0.45)
`gpsm1`	0.51 (0.36,0.66)	0.56 (0.32,0.80)	0.05 (-0.19,0.28)
`m.prog`	0.34 (0.07,0.61)	0.65 (0.43,0.86)	0.31 (-0.01,0.62)
`m.ds1`	0.39 (0.12,0.66)	0.63 (0.43,0.82)	0.23 (-0.05,0.51)
`ipw1`	0.37 (0.11,0.63)	0.52 (0.06,0.99)	0.15 (-0.36,0.67)
`aipw1`	0.39 (0.15,0.63)	0.59 (0.38,0.80)	0.20 (-0.10,0.50)
`gpsm2`	0.36 (0.20,0.52)	0.70 (0.56,0.83)	0.34 (0.20,0.48)
`m.ds2`	0.37 (0.10,0.65)	0.66 (0.42,0.90)	0.29 (-0.03,0.60)
`ipw2`	0.43 (0.13,0.73)	0.66 (0.23,1.09)	0.23 (-0.16,0.63)
`aipw2`	0.40 (0.13,0.68)	0.59 (0.33,0.84)	0.18 (-0.14,0.51)

The `naive` estimator and estimator from matching with X are biased due to the high dimensionality of the covariates, which is consistent with the simulation results. In Table 3.5, for contrast 'Treat2-Control' and 'Treat2-Treat1', the point estimates for the `naive` estimator and estimator from matching with X are quite different from other estimators but not significantly different. Point estimates for `gpsm` and `m.prog` are also different from `m.ds`, `ipw` and `aipw` estimates for one or two contrasts. We also notice that the `gpsm` and `ipw` estimators are more sensitive to model specification than the double score matching estimator and `aipw` estimator. Thus, the double score matching estimator and the `aipw` estimator are more robust.

Since none of the confidence intervals for comparing either treatment to control contains zero, there is sufficient statistical evidence that students undergoing the second form of tutoring will have better course grades than if they receive the first form of tutoring or no tutoring at all. Receiving the first form of tutoring also seems to help increase course grades compared to not receiving any form of tutoring at all.

3.7 Discussion

In this article, we propose the double score matching estimator for estimating pairwise treatment effects when there are more than two treatment levels. The DSM estimator

is an attractive alternative to existing multi-level weighting and matching methods for two main reasons: (i) it is a matching estimator, which is more robust to generalized propensity score outliers than weighting, (ii) it is more robust than the GPSM estimator by offering protection against propensity score or prognostic score model misspecification.

Given the recent development of sufficient dimension reduction techniques (Huang and Yang 2022), an extension of the proposed DSM estimator to allow nonparametric models of the GPS and GPGS is plausible. Take the GPS for example. The main idea of sufficient dimension reduction is to search for the fewest linear indices $B_w^{\mathrm{T}}x$ such that $g(w \mid x) = g_w(B_w^{\mathrm{T}}x)$, where B_w is a $p \times d_w$ matrix consisting of index coefficient, and $g_w(\cdot)$ is an unknown d_w-variate function. Although the estimator of $g_w(\cdot)$ is often not root-n consistent, the index coefficient can be estimated at the root-n rate of convergence under certain regularity conditions. The key insight is that in order to control for confounding, matching on $B_w^{\mathrm{T}}x$ is sufficient. Thus, extending the current theoretical framework to double score matching with sufficient dimension reduction of the propensity score and prognostics score is promising, which will be our future research topic.

As with GPSM, although the matching variables are well-balanced, individual covariates may not. In such cases, researchers may augment the double score further by adding important confounders to ensure balance. It is also important to note that just like other existing methods, the DSM method does not solve the problem of unmeasured confounding. Sensitivity analyses in the matching framework could be insightful.

CHAPTER

4

PROPENSITY SCORE MATCHING FOR ESTIMATION OF PAIRWISE MARGINAL HAZARD RATIOS

Estimation of treatment effects on survival outcomes using observational studies is of great importance in biomedical research. To overcome the problem of censoring and confounding, standard techniques frequently combine survival analysis approaches with propensity score (PS) methods so that credible causal conclusions can be drawn (Austin 2013). Gayat et al. (2012) apply propensity score matching to reduce confounding and Cox proportional hazards regression model to deal with a time-to-event outcome. Cole and Hernán (2004) adjust the survival curves using inverse probability weighting, while another classic approach combines Cox regression and inverse probability weighting (Hernán et al. 2000).

Observational survival data frequently arise from settings that inherently involve more than two (unordered) treatment levels. For instance, in real world oncology clinical settings, multiple subsequent treatment options are often introduced to patients with

non-small-cell lung cancer after initial chemotherapy. However, only a small number of existing methods are suitable for estimating treatment effects on survival outcomes in such settings (Zeng et al. 2021; Hu et al. 2022b,a). In particular, two recent works apply inverse probability weighting to adjust for confounding. Zeng et al. (2021) extend the balancing weights to multiple treatment levels by constructing pseudo-observations. When treatment and confounding are time-varying, Hu et al. (2022b) incorporate normalized inverse probability weights into a joint marginal structural model to estimate the causal effects.

An alternative class of PS methods considers matching on the PS. Matching is less sensitive to misspecification and extreme values of the PS compared to inverse probability weighting because it avoids inverting an estimated probability (Frölich 2004). Moreover, matching is intuitive in nature as it seeks to emulate the ideal of a randomized experiment (Rosenbaum 1989; Stuart 2010). In a seminal work, Yang et al. (2016) show that the advantage of matching is retained in multi-level treatment settings. That is, under a weaker version of the no unmeasured confounding assumption, matching on a single component of the generalized propensity scores (GPS) suffices to minimize confounding for estimation of the pairwise average treatment effects. Despite the validity of GPS matching (GPSM) when the outcome is continuous, its extension to censored survival outcome remains limited, with the exception of the work of Tang et al. (2019) for a binary treatment.

Following a large thread of literature on survival analysis, we focus on estimating the pairwise marginal hazard ratios among different treatment levels. While other causal estimands are available, the marginal hazard ratios are the most widely used measure for summarizing treatment effect on survival outcome (Austin 2014; Tchetgen Tchetgen and Robins 2012; VanderWeele and Ding 2017; Shu et al. 2021). Motivated by Yang et al. (2016), under no unmeasured confounding, we consider matching on the GPS to impute the unobserved potential survival outcomes at all treatment levels for all study subjects. To model the hazard rates among different treatment levels, the marginal Cox proportional hazard regression is applied to the imputed outcomes. Consistency and asymptotic normality of the proposed estimator for the pairwise causal hazard ratios are established when the GPS is known. In practice, the GPS is commonly estimated with a generalized linear model, e.g., a multinomial logistic regression model. We also derive the limiting distribution of the GPS estimator with estimated GPS, using similar arguments from Andreou and Werker (2012) and Abadie and Imbens (2016). The asymptotic variance suggests a variance estimator similar to the one for continuous outcome in Abadie and

Imbens (2016). We evaluate the finite sample performance of the GPSM estimator with a simulation study. Finally, we apply GPSM to analyze the IQVIA electronic medical records (EMR) data.

The outline of the article is as follows. In Section 4.1, we formulate the research problem by introducing basic notation, model and assumptions. In Section 4.2, we discuss the estimation procedure of causal pairwise hazard ratios by integrating GPS matching with a marginal Cox proportional hazard model. The asymptotic properties of the estimator are investigated in two theorems, and an estimator of asymptotic variance is established in the remainder of Section 4.2. To support our theoretical claims, numerical studies are conducted in Section 4.3. We illustrate the method with the analysis of the EMR data in Section 4.4.

4.1 Notation, Model, and Assumptions

4.1.1 Data structure

We begin under the potential outcomes framework with the multilevel treatment setting. A sample of n units is drawn from the population. We indicate the assigned treatment level for individual i by $W_i \in \mathbb{W} = \{0, 1, ..., J\}$, where $J \geq 1$, so that there are a total of $J + 1$ treatment levels. Here $\omega = 0$ refers to the reference treatment level or the control group.

Each unit has a set of potential survival times $\{T_i^{(\omega)} : \omega \in \mathbb{W}\}$ and a set of potential censoring times $\{C_i^{(\omega)} : \omega \in \mathbb{W}\}$. Given the above definitions, for $\omega \in \mathbb{W}$, we can further define $U_i^{(\omega)} = \min(T_i^{(\omega)}, C_i^{(\omega)})$ as the potential time to clinical event or censoring, $\Delta_i^{(\omega)} = I(T_i^{(\omega)} \leq C_i^{(\omega)})$ as the potential clinical event indicator, $N_i^{(\omega)}(t) = I(U_i^{(\omega)} \leq t, \Delta_i^{(\omega)} = 1)$ as the counting process for the potential clinical event, and $Y_i^{(\omega)}(t) = I(U_i^{(\omega)} \geq t)$ as the potential at-risk process. The above notation implicitly assumes no interference, where each subject's potential quantities are unaffected by treatment received by others.

Let T_i and C_i denote the observed survival and censoring time. We also assume consistency, which requires $T_i = T_i^{(W_i)}$ and $C_i = C_i^{(W_i)}$. Throughout, we assume independent censoring, i.e. $T_i^{(\omega)} \perp\!\!\!\perp C_i^{(\omega)}$ for all $\omega \in \mathbb{W}$, where"$\perp\!\!\!\perp$" means independent of. This is often reasonable under administrative censoring, where the censored times occur at the end of the study follow-up τ. Due to right-censoring, we only observe the smaller of the survival time and the censoring time for each unit. Therefore, we define the observed time to a clinical event or censoring as $U_i = \min(T_i, C_i)$, the clinical event indicator as $\Delta_i = I(T_i \leq C_i)$,

the observed counting process as $N_i(t) = I(U_i \leq t, \Delta_i = 1)$, and the observed at-risk process as $Y_i(t) = I(U_i \geq t)$. The consistency assumption also implies that $U_i, \Delta_i, N_i(t)$ and $Y_i(t)$ are all equal to their potential counterparts under the assigned treatment. In addition, we observe a set of time-invariant baseline covariates $\boldsymbol{X}_i \in \mathcal{R}^d$. Define the generalized propensity score $e_\omega(\boldsymbol{X}_i) = P(W_i = \omega \mid \boldsymbol{X}_i)$ as the probability of receiving treatment level ω given pre-treatment variables $\boldsymbol{X}_i$. Let $\boldsymbol{e}(\boldsymbol{X}_i) = (e_0(\boldsymbol{X}_i), ..., e_J(\boldsymbol{X}_i))^{\mathrm{T}}$ be the GPS vector for unit i. Here $^{\mathrm{T}}$ denotes the transpose. We assume the full data $\{\boldsymbol{X}_i, W_i, T_i^{(0)}, ..., T_i^{(J)}, C_i^{(0)}, ..., C_i^{(J)}\}_{i=1}^n$ are independent and identically distributed (i.i.d), so that the observed data $\{\boldsymbol{X}_i, W_i, U_i, \Delta_i\}_{i=1}^n$ are also i.i.d.

Some new notation will be needed later to simplify the presentation. For a vector $\boldsymbol{v}$, we will use the shorthand $\boldsymbol{v}^{\otimes 2}$ to denote the outer product $\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$. We define $\mathrm{diag}(\boldsymbol{v})$ as the square matrix that has $\boldsymbol{v}$ on the diagonal and zeroes everywhere else. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we will use $\boldsymbol{A} \otimes \boldsymbol{B}$ to denote the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$. We will use $\rightarrow_p$ and $\rightarrow_d$ to denote convergence in probability and in distribution, respectively.

4.1.2 Causal proportional hazard model

Let $\lambda_\omega(t) = \lim_{\delta_t \to 0} {\delta_t}^{-1} P\left(t \leq T^{(\omega)} < t + \delta_t \mid T^{(\omega)} \geq t\right)$ represent the hazard function for the distribution of $T^{(\omega)}$. In this article, our objective is to estimate the pairwise hazard ratios, which are ratios of the hazard rates corresponding to two different treatment levels. These quantities can be estimated using a marginal proportional hazard (PH) model (Tchetgen Tchetgen and Robins 2012) adapted to the multilevel treatment setting. We define a J-dimensional treatment indicator vector as

$$\boldsymbol{A}_\omega = (0, \cdots, 1, \cdots, 0)^{\mathrm{T}}, \tag{4.1}$$

for $\omega = 1, ..., J$, where the ω_{th} component is 1 and all other components are 0; for $\omega = 0$, we let $\boldsymbol{A}_0 = (0, \cdots, 0)^{\mathrm{T}}$.

Definition 4 (Causal proportional hazard model) *The marginal structural model for comparing treatment ω and treatment 0 is*

$$\lambda_\omega(t) = \lambda_0(t) \exp(\boldsymbol{A}_\omega^{\mathrm{T}} \boldsymbol{\beta}) \tag{4.2}$$

where $\lambda_0(t)$ is the population hazard rate if all individuals were assigned the reference treatment level $\omega = 0$. Here $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)^{\mathrm{T}}$, where β_ω denotes the log hazard ratio of

the ω_{th} treatment level to the reference level.

This is a causal model because it compares the outcomes under different treatments for the same group of individuals, i.e., the population of all individuals. The parameter β_ω describes the relative hazard of having a clinical event if all individuals received treatment ω compared to if all received treatment 0. Moreover, the log hazard ratios between any pair of treatment levels ω and ω' can be expressed as

$$\tau(\omega, \omega') = \log\left\{\frac{\lambda_{\omega'}(t)}{\lambda_\omega(t)}\right\} = \beta_{\omega'} - \beta_\omega. \tag{4.3}$$

Remark 4 *We note that treatment effect estimands other than the marginal hazard ratios are available, such as the pairwise difference of the restricted mean survival times (Yang et al. 2021). However, in this article, we focus on estimating the marginal hazard ratios, as they are commonly used to assess the effects of treatments on time-to-event outcomes by comparing the hazard functions of failure times of individuals in different treatment groups.*

4.1.3 Identification assumptions

In order to estimate the causal hazard ratios, we make the following identification assumptions.

Assumption 10 (Unconfoundedness) $W_i \perp\!\!\!\perp \{T_i^{(0)}, T_i^{(1)}, \cdots, T_i^{(J)}\} \mid \boldsymbol{X}_i$.

Assumption 11 (Positivity) *With probability* 1, $0 < \underline{c} < e_\omega(\boldsymbol{x}) < \bar{c} < 1$ *for all values of* $\boldsymbol{x}$ *and* $\omega \in \mathbb{W}$.

Assumptions 10 and 11 are the usual identification assumptions. The unconfoundedness assumption is untestable but can be made more plausible by collecting detailed information on individual characteristics that are related to treatment assignment and survival outcome. Assumption 11 ensures that there is sufficient overlap between the covariate distributions of any two treatment levels. Without this assumption, there will be subjects for which we cannot estimate the pairwise treatment effects without relying on extrapolation.

4.2 Methodology

4.2.1 Matching

If we were to observe all potential outcome processes, we can fit a Cox PH model to $\{N_i^{(\omega)}(t), Y_i^{(\omega)}(t) : \omega = 0, 1, ..., J; i = 1, \ldots, n\}$ to obtain an estimator for $\boldsymbol{\beta}$ and the causal log hazard ratios. However, the fundamental problem in the potential outcomes framework is that for a particular unit i, we can observe $\{N_i^{(\omega)}(t), Y_i^{(\omega)}(t)\}$ under only one of $J+1$ treatment levels. From this point of view, causal inference is inherently a missing data problem.

We use matching as a tool to impute the missing potential outcomes. We consider matching with replacement with the number of matches fixed at $M \geq 1$, ignoring ties. In typical applications, M is small and is often fixed at $M = 1$ (Abadie and Imbens 2016). We first consider the case when the generalized propensity score is known and matching is carried out based on the true GPS. We define $\mathcal{J}_M(\omega, p_\omega)$ as the collection of indices of M subjects such that (i) they are from treatment level ω and (ii) their generalized propensity scores under ω are closest to p_ω. Formally, that is, $\mathcal{J}_M(\omega, p_\omega) = \{j : W_j = \omega,\ \sum_{i=1}^n I(W_i = \omega) I\{|p_\omega - e_\omega(\boldsymbol{X}_j)| \leq |p_\omega - e_\omega(\boldsymbol{X}_i)|\} \leq M\}$. Without loss of generality, we use the Euclidean distance to determine neighbors. For each individual i, we define the imputed potential outcome process as

$$\left\{\overline{N}_i^{*(\omega)}(t), \overline{Y}_i^{*(\omega)}(t)\right\} = \begin{cases} \{N_i(t), Y_i(t)\} & \text{if } \omega = W_i, \\ M^{-1} \sum_{j \in \mathcal{J}_M\{\omega, e_\omega(\boldsymbol{X}_i)\}} \{N_j(t), Y_j(t)\} & \text{if } \omega \neq W_i, \end{cases} \tag{4.4}$$

for $\omega = 0, 1, ..., J$. To be precise, for individual i, the potential outcome process under $\omega = W_i$ is $\{N_i(t), Y_i(t)\}$; the potential outcome process under $\omega \neq W_i$ is not observed but can be imputed based on the observed outcome processes of the nearest individuals with ω.

As an illustrative example when $J = 2$ and $M = 2$, Table 4.1 shows the full imputed dataset, as well as an equivalent representation given by the weighted observed dataset, where each individual i is assigned a weight of $1 + k_i(W_i)/M$. Here $k_i(W_i)$ is the number of times individual i is used as a match for individuals who received treatment levels other than W_i. We refer to $k_i(W_i)/M$ as the *matching weight* with which individual i can represent for individuals in all other treatment groups if treatment were randomly assigned.

Table 4.1: GPS matching for imputation of the missing potential outcomes when the number of treatment levels is 3 and the number of matches is 2; the matching index set $\mathcal{J}_2\{1, e_1(\boldsymbol{X}_n)\}$, for instance, denotes the set containing the indices of 2 subjects who received treatment level $\omega = 1$ and whose generalized propensity scores evaluated at 1 are closest to $e_1(\boldsymbol{X}_n)$, the GPS of n-th unit evaluated at $\omega = 1$; the full imputed dataset can be equivalently viewed as a weighted dataset, where $k_i(W_i)$ is the number of times individual i is used as a match.

Id	W	Full Imputed Dataset							Weighted Dataset	
			$\omega = 0$		$\omega = 1$		$\omega = 2$		Weight	(U, Δ)
		Weight	Matching idx	$(U^{(0)}, \Delta^{(0)})$	Matching idx	$(U^{(1)}, \Delta^{(1)})$	Matching idx	$(U^{(2)}, \Delta^{(2)})$		
1	1	$\frac{1}{2}$	$\mathcal{J}_2\{0, e_0(\boldsymbol{X}_1)\}$	$\left(U_{j_{1,0}^{(1)}}, \Delta_{j_{1,0}^{(1)}}\right)$		(U_1, Δ_1)	$\mathcal{J}_2\{2, e_2(\boldsymbol{X}_1)\}$	$\left(U_{j_{1,2}^{(1)}}, \Delta_{j_{1,2}^{(1)}}\right)$	$1 + \frac{k_1(1)}{2}$	(U_1, Δ_1)
		$\frac{1}{2}$	$= \{j_{1,0}^{(1)}, j_{1,0}^{(2)}\}$	$\left(U_{j_{1,0}^{(2)}}, \Delta_{j_{1,0}^{(2)}}\right)$		(U_1, Δ_1)	$= \{j_{1,2}^{(1)}, j_{1,2}^{(2)}\}$	$\left(U_{j_{1,2}^{(2)}}, \Delta_{j_{1,2}^{(2)}}\right)$		
2	0	$\frac{1}{2}$		(U_1, Δ_1)	$\mathcal{J}_2\{1, e_1(\boldsymbol{X}_2)\}$	$\left(U_{j_{2,1}^{(1)}}, \Delta_{j_{2,1}^{(1)}}\right)$	$\mathcal{J}_2\{2, e_2(\boldsymbol{X}_2)\}$	$\left(U_{j_{2,2}^{(1)}}, \Delta_{j_{2,2}^{(1)}}\right)$	$1 + \frac{k_2(0)}{2}$	(U_2, Δ_2)
		$\frac{1}{2}$		(U_2, Δ_2)	$= \{j_{2,1}^{(1)}, j_{2,1}^{(2)}\}$	$\left(U_{j_{2,1}^{(2)}}, \Delta_{j_{2,1}^{(2)}}\right)$	$= \{j_{2,2}^{(1)}, j_{2,2}^{(2)}\}$	$\left(U_{j_{2,2}^{(2)}}, \Delta_{j_{2,2}^{(2)}}\right)$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	2	$\frac{1}{2}$	$\mathcal{J}_2\{0, e_0(\boldsymbol{X}_n)\}$	$\left(U_{j_{n,0}^{(1)}}, \Delta_{j_{n,0}^{(1)}}\right)$	$\mathcal{J}_2\{1, e_1(\boldsymbol{X}_n)\}$	$\left(U_{j_{n,1}^{(1)}}, \Delta_{j_{n,1}^{(1)}}\right)$		(U_n, Δ_n)	$1 + \frac{k_n(2)}{2}$	(U_n, Δ_n)
		$\frac{1}{2}$	$= \{j_{n,0}^{(1)}, j_{n,0}^{(2)}\}$	$\left(U_{j_{n,0}^{(2)}}, \Delta_{j_{n,0}^{(2)}}\right)$	$= \{j_{n,1}^{(1)}, j_{n,1}^{(2)}\}$	$\left(U_{j_{n,1}^{(2)}}, \Delta_{j_{n,1}^{(2)}}\right)$		(U_n, Δ_n)		

In observational data, the generalized propensity score is typically unknown and needs to be estimated from the observed data. Following most of the empirical literature, we assume that the GPS follows a generalized linear model $\{e_\omega(\boldsymbol{X}_i; \boldsymbol{\theta}^*) : \omega = 0, ..., J\}$ with $e_\omega(\boldsymbol{X}_i; \boldsymbol{\theta}^*) = \boldsymbol{p}\left(w \mid \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}_J^*\right)$ and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^{*\mathrm{T}}, ..., \boldsymbol{\theta}_J^{*\mathrm{T}})^{\mathrm{T}}$. The matching procedure can then be carried out based on the estimated GPS $e_\omega(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}^*$. We will denote $k_i(\omega)$ as $k_{\widehat{\boldsymbol{\theta}},i}(\omega)$ to reflect its dependence on $\widehat{\boldsymbol{\theta}}$.

4.2.2 Estimating equations

Define $\Lambda_0(t) = \int_0^t \lambda_0(v)\mathrm{d}v$ as the cumulative hazard function for $\omega = 0$ at time t. In what follows, we derive our estimators for $\boldsymbol{\beta}^*$ and $\Lambda_0(t)$, $t \geq 0$. Based on the imputed dataset, we can fit the causal PH model directly. Specifically, the estimating functions for $\boldsymbol{\beta}^*$ and $\Lambda_0(t)$, $t \geq 0$, are

$$\sum_{i=1}^{n}\sum_{\omega=0}^{J}\left\{\mathrm{d}\overline{N}_i^{*(\omega)}(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta})\overline{Y}_i^{*(\omega)}(t)\right\}, \tag{4.5}$$

$$\sum_{i=1}^{n}\sum_{\omega=0}^{J}\int_0^\tau \boldsymbol{A}_\omega\left\{\mathrm{d}\overline{N}_i^{*(\omega)}(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta})\overline{Y}_i^{*(\omega)}(t)\right\}, \tag{4.6}$$

where $\overline{N}_i^{*(\omega)}(t)$ and $\overline{Y}_i^{*(\omega)}(t)$ are defined in (4.4).

Because the full imputed dataset can be equivalently represented by a weighted observed dataset, we can also write (4.5) and (4.6) as

$$\sum_{i=1}^{n} \left\{1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M\right\} \left\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_{W_i}^{\mathrm{T}}\boldsymbol{\beta})Y_i(t)\right\}, \tag{4.7}$$

$$\sum_{i=1}^{n} \left\{1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M\right\} \int_0^{\tau} \boldsymbol{A}_{W_i} \left\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_{W_i}^{\mathrm{T}}\boldsymbol{\beta})Y_i(t)\right\}, \tag{4.8}$$

respectively. Setting (4.7) equal to zero, we can obtain the estimator for $\mathrm{d}\Lambda_0(t)$ for fixed $\boldsymbol{\beta}$ as

$$\mathrm{d}\widehat{\Lambda}_0(t) = \frac{\sum_{i=1}^{n}\{1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M\}\mathrm{d}N_i(t)}{\sum_{i=1}^{n}\{1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M\}\exp(\boldsymbol{A}_{W_i}^{\mathrm{T}}\boldsymbol{\beta})Y_i(t)}. \tag{4.9}$$

Substituting (4.9) into (4.8), we are able to obtain the estimating equation for $\boldsymbol{\beta}$,

$$\boldsymbol{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \{1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M\} \left\{\boldsymbol{A}_{W_i} - \widehat{\boldsymbol{Q}}(\boldsymbol{\beta}, t)\right\} \mathrm{d}N_i(t) = \boldsymbol{0}, \tag{4.10}$$

where

$$\widehat{\boldsymbol{Q}}(\boldsymbol{\beta}, t) = \frac{\sum_{j=1}^{n} \left\{1 + k_{\widehat{\boldsymbol{\theta}},j}(W_j)/M\right\} Y_j(t)\exp(\boldsymbol{A}_{W_j}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{A}_{W_j}}{\sum_{j=1}^{n} \left\{1 + k_{\widehat{\boldsymbol{\theta}},j} j(W_j)/M\right\} Y_j(t)\exp(\boldsymbol{A}_{W_j}^{\mathrm{T}}\boldsymbol{\beta})}. \tag{4.11}$$

Equation (4.10) is the partial score equation of a Cox PH model with covariates $\boldsymbol{A}_{W_i}$ and weighted by $1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M$. Therefore, an estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$ can be calculated using standard software. The steps to compute the GPS matching estimator of $\boldsymbol{\beta}^*$ can be summarized as follows.

Step 1. Fit a generalized propensity score model, and obtain an estimate $\widehat{\boldsymbol{\theta}}$.

Step 2. As in Table 4.1, match on the estimated GPS to create an imputed dataset. Compute the matching weight $1 + k_{\widehat{\boldsymbol{\theta}},i}(W_i)/M$ for each individual.

Step 3. Obtain the GPS matching estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}^*$ by solving (4.10) using standard software.

In the following two subsections, we divide the investigation of the asymptotic properties of the GPS matching estimator into two steps. First, we establish the results for $\widehat{\boldsymbol{\beta}}$ when the true value of the generalized propensity score parameters $\boldsymbol{\theta}^*$ are known in Section

4.2.3, and second, building on the step one result, we quantify the impact of the estimation of $\boldsymbol{\theta}^*$ on the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ in Section 4.2.4.

4.2.3 Asymptotic results with known GPS

Under regularity conditions in Assumption S1, we show that there exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ and a function $\boldsymbol{Q}(\boldsymbol{\beta}, t)$ such that for all $(\boldsymbol{\beta}, t) \in \mathcal{B} \times [0, \tau]$, $\widehat{\boldsymbol{Q}}(\boldsymbol{\beta}, t) \to_p \boldsymbol{Q}(\boldsymbol{\beta}, t)$, as $n \to \infty$, where

$$\boldsymbol{Q}\left(\boldsymbol{\beta}, t\right) = \frac{\left(\exp\left(\beta_1\right) E\{Y^{(1)}(t)\}, \cdots, \exp\left(\beta_J\right) E\{Y^{(J)}(t)\}\right)^{\mathrm{T}}}{E\{Y^{(0)}(t)\} + \sum_{\omega=1}^{J} \exp\left(\beta_\omega\right) E\{Y^{(w)}(t)\}}.$$

We then define

$$\boldsymbol{H}_i(\omega) = \int_0^\tau \{\boldsymbol{A}_\omega - \boldsymbol{Q}(\boldsymbol{\beta}^*, t)\} \left\{ \mathrm{d}N_i^{(\omega)}(t) - \mathrm{d}\Lambda_0(t) \exp(\boldsymbol{A}_\omega^{\mathrm{T}} \boldsymbol{\beta}^*) Y_i^{(\omega)}(t) \right\}, \tag{4.12}$$

and its conditional mean and variance as $\boldsymbol{\mu}_{\boldsymbol{H}}(\omega, \boldsymbol{X}) = E\{\boldsymbol{H}(\omega) \mid \boldsymbol{X}\}$ and $\boldsymbol{\sigma}^2_{\boldsymbol{H}}(\omega, \boldsymbol{X}) = \mathrm{var}\{\boldsymbol{H}(\omega) \mid \boldsymbol{X}\}$, respectively.

Because $\widehat{\boldsymbol{\beta}}$ is the solution to the estimating equation (4.10), the main step is to characterize the asymptotic properties of $\boldsymbol{S}_n(\boldsymbol{\beta}^*)$. With a known $\boldsymbol{\theta}^*$, we can show that

$$n^{-1/2} \boldsymbol{S}_n(\boldsymbol{\beta}^*) = n^{-1/2} \sum_{i=1}^{n} \left\{ 1 + \frac{k_{\boldsymbol{\theta}^*, i}(W_i)}{M} \right\} \boldsymbol{H}_i(W_i) + \boldsymbol{o}_p(1). \tag{4.13}$$

Based on (4.13) and the M-estimation theory, we derive the asymptotic results for $\widehat{\boldsymbol{\beta}}$ as follows.

Theorem 5 *Under Assumptions 10, 11 and the regularity conditions in Assumption S1 presented in the supplementary material, with the known generalized propensity score,*

$$n^{1/2}\{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\} \to_d \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}),$$

as $n \to \infty$, *where* $\boldsymbol{V} = \{\boldsymbol{A}(\boldsymbol{\beta}^*)\}^{-1}\boldsymbol{V}_s\{\boldsymbol{A}(\boldsymbol{\beta}^*)\}^{-1}$,

$$\begin{aligned}\boldsymbol{V}_s =& E\left(\left[\sum_{\omega=0}^{J}\boldsymbol{\mu}_{\boldsymbol{H}}\left\{\omega, e_\omega(\boldsymbol{X};\boldsymbol{\theta}^*)\right\}\right]^{\otimes 2}\right) \\ &+\sum_{\omega=0}^{J}E\left[\boldsymbol{\sigma}^2_{\boldsymbol{H}}\left\{\omega, e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\left\{\frac{2M+1}{2Me_\omega(\boldsymbol{X};\boldsymbol{\theta}^*)}-\frac{e_\omega(\boldsymbol{X};\boldsymbol{\theta}^*)}{2M}\right\}\right]\end{aligned}$$

and

$$\boldsymbol{A}(\boldsymbol{\beta}^*) = E\left[\int_0^\tau\left\{\operatorname{diag}\left(\boldsymbol{Q}\left(\boldsymbol{\beta}^*,t\right)\right)-\boldsymbol{Q}\left(\boldsymbol{\beta}^*,t\right)^{\otimes 2}\right\}\left\{\sum_{\omega=0}^{J}\mathrm{d}N^{(\omega)}(t)\right\}\right].$$

Therefore, for any pairwsie comparison between treatment ω and ω', the estimated log hazard ratio follows

$$n^{1/2}\left\{\widehat{\tau}\left(\omega,\omega'\right)-\tau\left(\omega,\omega'\right)\right\}\to_d\mathcal{N}\left\{0,\left(\boldsymbol{A}_{\omega'}-\boldsymbol{A}_{\omega}\right)^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{A}_{\omega'}-\boldsymbol{A}_{\omega}\right)\right\}$$

as $n \to \infty$.

4.2.4 Asymptotic results with estimated GPS

We now study the asymptotic properties of the GPS matching estimator based on the estimated generalized propensity score. The technique we will use is based on Andreou and Werker (2012). The main idea is to apply Le Cam's third lemma (Le Cam et al. 2000) to locally asymptotically normal models. Let $P^{\boldsymbol{\theta}^*}$ be the true probability measure of n copies of the random variables, $\boldsymbol{\theta}_n$ be contiguous to $\boldsymbol{\theta}^*$, and $P^{\boldsymbol{\theta}_n}$ be the probability measure with the local parameter $\boldsymbol{\theta}_n$. Under $P^{\boldsymbol{\theta}_n}$, denote the true parameter value as $\boldsymbol{\beta}^*(\boldsymbol{\theta}_n)$, the GPSM estimator based on the true parameter $\boldsymbol{\theta}_n$ as $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_n)$, and the log likelihood of $P^{\boldsymbol{\theta}^*}$ with respect to $P^{\boldsymbol{\theta}_n}$ as $\Lambda_n(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n)$. If we can show that *under* $P^{\boldsymbol{\theta}_n}$,

$$\left(n^{1/2}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_n)-\boldsymbol{\beta}^*(\boldsymbol{\theta}_n)\}, n^{1/2}\{\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_n\}, \Lambda_n(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}_n)\right)^{\mathrm{T}} \tag{4.14}$$

has a limiting normal distribution, Le Cam's third lemma states that *under* $P^{\boldsymbol{\theta}^*}$, $n^{1/2}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_n)-\boldsymbol{\beta}^*(\boldsymbol{\theta}_n)\}$ has a limiting normal distribution. Then heuristically by replacing $\boldsymbol{\theta}_n$ with $\widehat{\boldsymbol{\theta}}$, one can then approximate the asymptotic distribution of $n^{1/2}\{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\}$ as shown in Theorem 6.

Assuming a generalized linear model for the GPS, for $\omega \in \{0,...,J\}$ and $\omega' \in \{1,...,J\}$, we define $f\left(\omega;\omega' \mid \boldsymbol{X};\boldsymbol{\theta}\right)$ as satisfying $f\left(\omega;\omega' \mid \boldsymbol{X};\boldsymbol{\theta}\right)\boldsymbol{X} = \frac{\partial}{\partial\boldsymbol{\theta}_{w'}}e_\omega(\boldsymbol{X};\boldsymbol{\theta})$. We

let $\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}\right)=\left(f\left(\omega;1 \mid \boldsymbol{X};\boldsymbol{\theta}\right),...,f\left(\omega;J \mid \boldsymbol{X};\boldsymbol{\theta}\right)\right)^{\mathrm{T}}$.

Theorem 6 *Under Assumptions 10, 11 and the regularity conditions in Assumption S1 presented in the supplementary material, and assuming a correctly specified generalized propensity score model,*

$$n^{1/2}\{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*}\}\rightarrow_{d}\mathcal{N}(\mathbf{0},\boldsymbol{V}_{adj}),$$

as $n\rightarrow\infty$*, where* $\boldsymbol{V}_{adj}=\{\boldsymbol{A}(\boldsymbol{\beta}^{*})\}^{-1}\widetilde{\boldsymbol{V}}_{s}\{\boldsymbol{A}(\boldsymbol{\beta}^{*})\}^{-1}$ *and* $\widetilde{\boldsymbol{V}}_{s}=\boldsymbol{V}_{s}-\boldsymbol{C}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}^{*}}^{-1}\boldsymbol{C}$*, where* $\boldsymbol{V}_{s}$ *and* $\boldsymbol{A}(\boldsymbol{\beta}^{*})$ *are defined in Theorem 5,* $\boldsymbol{I}_{\boldsymbol{\theta}^{*}}$ *is the Fisher information of* $\boldsymbol{\theta}^{*}$*, and*

$$\boldsymbol{C}=\sum_{\omega=0}^{J}E\left[\frac{\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^{*}\right)}{e_{\omega}\left(\boldsymbol{X};\boldsymbol{\theta}^{*}\right)}\otimes\operatorname{cov}\left\{\boldsymbol{X},\boldsymbol{\mu}_{\boldsymbol{H}}\left(\omega,\boldsymbol{X}\right) \mid e_{\omega}\left(\boldsymbol{X};\boldsymbol{\theta}^{*}\right)\right\}\right]. \tag{4.15}$$

Theorem 6 shows that matching based on the estimated GPS improves the efficiency of the matching estimator compared to matching based on the true GPS when it is known (Theorem 5). This phenomenon is in line with that in the setting with a continuous outcome (Abadie and Imbens 2016; Yang et al. 2016).

4.2.5 Estimation of asymptotic variance

In this section, we discuss estimation of the large sample variances of $\widehat{\boldsymbol{\beta}}$ adjusting for first step estimation of the generalized propensity score. From the previous section, we have that

$$\boldsymbol{V}_{\mathrm{adj}}=\{\boldsymbol{A}(\boldsymbol{\beta}^{*})\}^{-1}\{\boldsymbol{V}_{s}-\boldsymbol{C}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}^{*}}^{-1}\boldsymbol{C}\}\{\boldsymbol{A}(\boldsymbol{\beta}^{*})\}^{-1}.$$

We will estimate each component on the right hand side of the equation separately. We first use the observed information to estimate the Fisher information $\boldsymbol{I}_{\boldsymbol{\theta}^{*}}$:

$$\widehat{\boldsymbol{I}}_{\boldsymbol{\theta}^{*}}=\frac{1}{n}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\frac{I\left(W_{i}=\omega\right)}{e_{\omega}^{2}(\boldsymbol{X}_{i};\widehat{\boldsymbol{\theta}})}\boldsymbol{f}\left(\omega \mid \boldsymbol{X}_{i};\widehat{\boldsymbol{\theta}}\right)^{\otimes 2}\otimes\boldsymbol{X}_{i}\boldsymbol{X}_{i}^{\mathrm{T}}.$$

Let $m_{k}\{\omega,e_{\omega}(\boldsymbol{X}_{i};\widehat{\boldsymbol{\theta}})\}$ to denote the index of the k-th nearest neighbor matched to unit i based on the estimated GPS. For estimation of $\boldsymbol{V}_{s}$, we first create an imputed dataset $\{\boldsymbol{H}_{i1}^{*}(\omega),\boldsymbol{H}_{i2}^{*}(\omega)\}_{i=1}^{n}$ where

$$\boldsymbol{H}_{i1}^{*}(\omega)=\begin{cases}\boldsymbol{H}_{i}(\omega) & \text{if } W_{i}=\omega\\ \boldsymbol{H}_{m_{1}\left\{\omega,e_{\omega}(\boldsymbol{X}_{i};\widehat{\boldsymbol{\theta}})\right\}}(\omega) & \text{if } W_{i}\neq\omega\end{cases}$$

and

$$\boldsymbol{H}^*_{i2}(\omega) = \begin{cases} \boldsymbol{H}_{m_1\left\{\omega, e_\omega(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}})\right\}}(\omega) & \text{if } W_i = \omega \\ \boldsymbol{H}_{m_2\left\{\omega, e_\omega(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}})\right\}}(\omega) & \text{if } W_i \neq \omega. \end{cases}$$

Then, $\boldsymbol{V}_s$ can be estimated by

$$\widehat{\boldsymbol{V}}_s = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\omega=0}^{J}\boldsymbol{H}^*_{i1}(\omega)\right\}^{\otimes 2} + \frac{1}{n}\sum_{i=1}^{n}\left\{k_{\widehat{\boldsymbol{\theta}},i}\left(W_i\right) + k_{\widehat{\boldsymbol{\theta}},i}\left(W_i\right)^2\right\}\widehat{\boldsymbol{\sigma}}^2_i,$$

where $\widehat{\boldsymbol{\sigma}}^2_i = \sum_{k=1}^{2}\left[\boldsymbol{H}^*_{ik}(\omega) - \frac{1}{2}\left\{\boldsymbol{H}^*_{i1}(\omega) + \boldsymbol{H}^*_{i2}(\omega)\right\}\right]^{\otimes 2} = \frac{1}{2}\left\{\boldsymbol{H}^*_{i1}(\omega) - \boldsymbol{H}^*_{i2}(\omega)\right\}^{\otimes 2}$.

Next we can construct an estimator of $\boldsymbol{C}$ by averaging over the sample:

$$\widehat{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\frac{\boldsymbol{f}(\omega \mid \boldsymbol{X}_i;\widehat{\boldsymbol{\theta}})}{e_\omega(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}})} \otimes \widehat{\operatorname{cov}}\left\{\boldsymbol{X}_i, \boldsymbol{\mu}_{\boldsymbol{H}}\left(\omega, \boldsymbol{X}_i\right) \mid e_\omega(\boldsymbol{X}_i;\boldsymbol{\theta}^*)\right\}.$$

For estimation of the conditional covariance, we follow the same matching procedure to create an imputed dataset $\{\boldsymbol{X}^*_{i1}(\omega), \boldsymbol{X}^*_{i2}(\omega)\}_{i=1}^n$. Then $\widehat{\operatorname{cov}}\left\{\boldsymbol{X}_i, \boldsymbol{\mu}_{\boldsymbol{H}}\left(\omega, \boldsymbol{X}_i\right) \mid e_\omega(\boldsymbol{X}_i;\boldsymbol{\theta}^*)\right\}$ can be estimated by $\sum_{l=1}^{2}\left\{\boldsymbol{X}^*_{il}(\omega) - \frac{1}{2}\sum_{k=1}^{2}\boldsymbol{X}^*_{ik}(\omega)\right\}\left\{\boldsymbol{H}^*_{il}(\omega) - \frac{1}{2}\sum_{k=1}^{2}\boldsymbol{H}^*_{ik}(\omega)\right\}^{\mathrm{T}} = \frac{1}{2}\left\{\boldsymbol{X}^*_{i1}(\omega) - \boldsymbol{X}^*_{i2}(\omega)\right\}\left\{\boldsymbol{H}^*_{i1}(\omega) - \boldsymbol{H}^*_{i2}(\omega)\right\}^{\mathrm{T}}$.

Finally, to estimate $\boldsymbol{A}(\boldsymbol{\beta}^*)$, we use

$$\widehat{\boldsymbol{A}}(\boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\left\{\operatorname{diag}(\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\beta}}, t_k)) - \widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\beta}}, t_k)^{\otimes 2}\right\}\left\{\sum_{\omega=0}^{J}\mathrm{d}\overline{N}_i^{*(\omega)}\left(t_k\right)\right\}\right],$$

where $\{t_1, ...t_K\}$ are distinct observed time points. Putting everything together, our final estimator of the asymptotic variance is

$$\widehat{\boldsymbol{V}}_{\mathrm{adj}} = \{\widehat{\boldsymbol{A}}(\boldsymbol{\beta}^*)\}^{-1}\{\widehat{\boldsymbol{V}}_s - \widehat{\boldsymbol{C}}^{\mathrm{T}}\widehat{\boldsymbol{I}}^{-1}_{\boldsymbol{\theta}^*}\widehat{\boldsymbol{C}}\}\{\widehat{\boldsymbol{A}}(\boldsymbol{\beta}^*)\}^{-1}.$$

4.3 Simulations

The objective of this section is to assess the finite sample performance of the proposed GPSM estimator against existing estimators of the pairwise log hazard ratios. Previous simulations (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 is inferior to matching estimators in terms of root mean

squared error . On the other hand, it has been found that matching on higher-dimensional covariates may inflict bias for commonly found sample sizes (Abadie and Imbens 2006). These results motivate us to compare the weighting and matching estimators in the multi-level treatment setting, where the generalized propensity scores are more likely to be close to zero or one. We consider the following estimators of:

1. the naive estimator (`Naive`) obtained by fitting a marginal Cox proportional hazard model with the treatment indicator vector being the only set of covariates.

2. the generalized inverse probability weighting estimator (`GIPW`) obtained by fitting a weighted Cox proportional hazard model with the treatment indicator vector being the only set of covariates, where each observation is weighted by the inverse of the estimated probability of receiving the actual treatment level.

3. the regression estimator (`REG`) obtained by fitting a Cox proportional hazard model with treatment assignment vector and pre-treatment variables as the covariates.

4. the proposed GPSM estimator (`GPSM`) based on the estimated GPS.

5. the matching estimator (`COVM`) based on the set of baseline covariates.

6. the matching estimator (`PSSM`) based on the vector of estimated GPS.

We set the number of matches to $M = 1$ for the three matching estimators: `GPSM`, `COVM` and `PSSM`. For variance estimation, we use the Abadie and Imbens (2012) variance estimator for `COVM` and `PSSM`. The proposed asymptotic variance estimator is used for `GPSM`. For variance estimation of `Naive`, `GIPW`, and `REG`, the conventional robust variance estimator from standard software is used.

We consider a sample size of $n = 5000$ with $J + 1 = 5$ treatment levels. For each unit, we simulate five baseline covariates $\{X_1, ..., X_5\}$, each independently simulated from an exponential distribution with mean 1/2. We use $\boldsymbol{X}_i^{\mathrm{T}} = (1, X_{1i}, ..., X_{5i})$ to denote the full covariate vector. Treatment assignment W_i is generated from a multinomial distribution with parameter $(p\,(0 \mid \boldsymbol{X}_i), ..., p\,(4 \mid \boldsymbol{X}_i))$, where $\boldsymbol{p}\,(w \mid \boldsymbol{X}_i) = \exp\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}_\omega\right)/\{1 + \sum_{j=1}^{4} \exp\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}_j\right)\}$ for $w = 1, ..., 4$ and $\boldsymbol{p}\,(0 \mid \boldsymbol{X}_i) = 1/\{1+\sum_{j=1}^{4} \exp\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}_j\right)\}$. We set $\boldsymbol{\theta}_1 = (6, -3, -2, -1, -1, -3)^{\mathrm{T}}$, $\boldsymbol{\theta}_2 = (6, -1, -3, -2, -1, -3)^{\mathrm{T}}$, $\boldsymbol{\theta}_3 = (6, -1, -1, -3, -2, -3)^{\mathrm{T}}$, and $\boldsymbol{\theta}_4 = (6, -2, -1, -1, -3, -3)^{\mathrm{T}}$ to simulate strong separation in covariate distributions, which makes it fundamentally difficult in removing biases in all estimating 10 treatment effects simultaneously. The algorithm used to generate the potential survival outcomes

$T_i^{(\omega)}$ can be found in the supplementary material. The observed survival time is computed by $T_i = \sum_{\omega=0}^{J} I\left(W_i = \omega\right) T_i^{(\omega)}$. The censoring times are generated from a uniform distribution such that between 30% to 50% of the individuals are censored. Here we let $\boldsymbol{\beta}^* = (0, 0, 0, 0)^{\mathrm{T}}$ so that treatment effect is absent. The results for $\boldsymbol{\beta}^* \neq 0$ are similar and can be found in the supplementary material.

We compare the performance of the above estimators over 1000 simulated datasets under two scenarios: (i) the generalized propensity scores are estimated using a correctly specified multinomial logistic regression model, and (ii) the generalized propensity scores are estimated using a multinomial logistic regression model with linear predictor $\boldsymbol{X}_i'^{\mathrm{T}}\boldsymbol{\theta}_\omega$, where $\boldsymbol{X}_i'^{\mathrm{T}} = (1, X_{1i}^2, \log(X_{2i}), \sqrt{X_{3i}}, X_{4i}, X_{5i})$, so that the GPS model is misspecified.

Figure 4.1 shows the performance of the estimators in estimating the pairwise log hazard ratios $\tau(\omega, \omega')$ measured by the absolute bias, root mean squared error (RMSE), and coverage of 95% confidence intervals under scenario (i), i.e. when the GPS model is correctly specified. Among all the competing estimators, the `GPSM` and `GIPW` estimators are the most effective at minimizing bias. The root mean squared error further reveals that the `GPSM` estimator is more advantageous than `GIPW` at reducing variance. The `COVM` and `PSSM` estimators introduce larger bias, which is most likely attributed to matching on a multi-dimensional variable (Abadie and Imbens 2006). The `REG` estimator shows bias because by construction it targets the *conditional* log hazard ratios rather than the marginal ones. Moreover, the proposed asymptotic variance estimator shows reasonably accurate coverage; `COVM` leads to undercoverage, which is consistent with the findings in Abadie and Imbens (2006). In sum, `GPSM` shows the smallest bias and most reliable coverage rates among all the competing estimators.

Figure 4.2 shows the results under scenario (ii), i.e. when the GPS model is misspecified. As expected, the performance of all GPS-based methods deteriorates compared to (i). Overall, weighting inversely by the GPS (`GIPW`) yields large bias, RMSE, and poor coverage. Its bias and RMSE are even larger than the `Naive` estimator for estimating most treatment contrasts. The `GPSM` estimator shows better robustness than `GIPW` as `GPSM` maintains a much smaller RMSE and better coverage with respect to all treatment contrasts. In the meantime, the `GPSM` estimator is inferior to `COVM` in terms of bias, RMSE, as well as coverage.

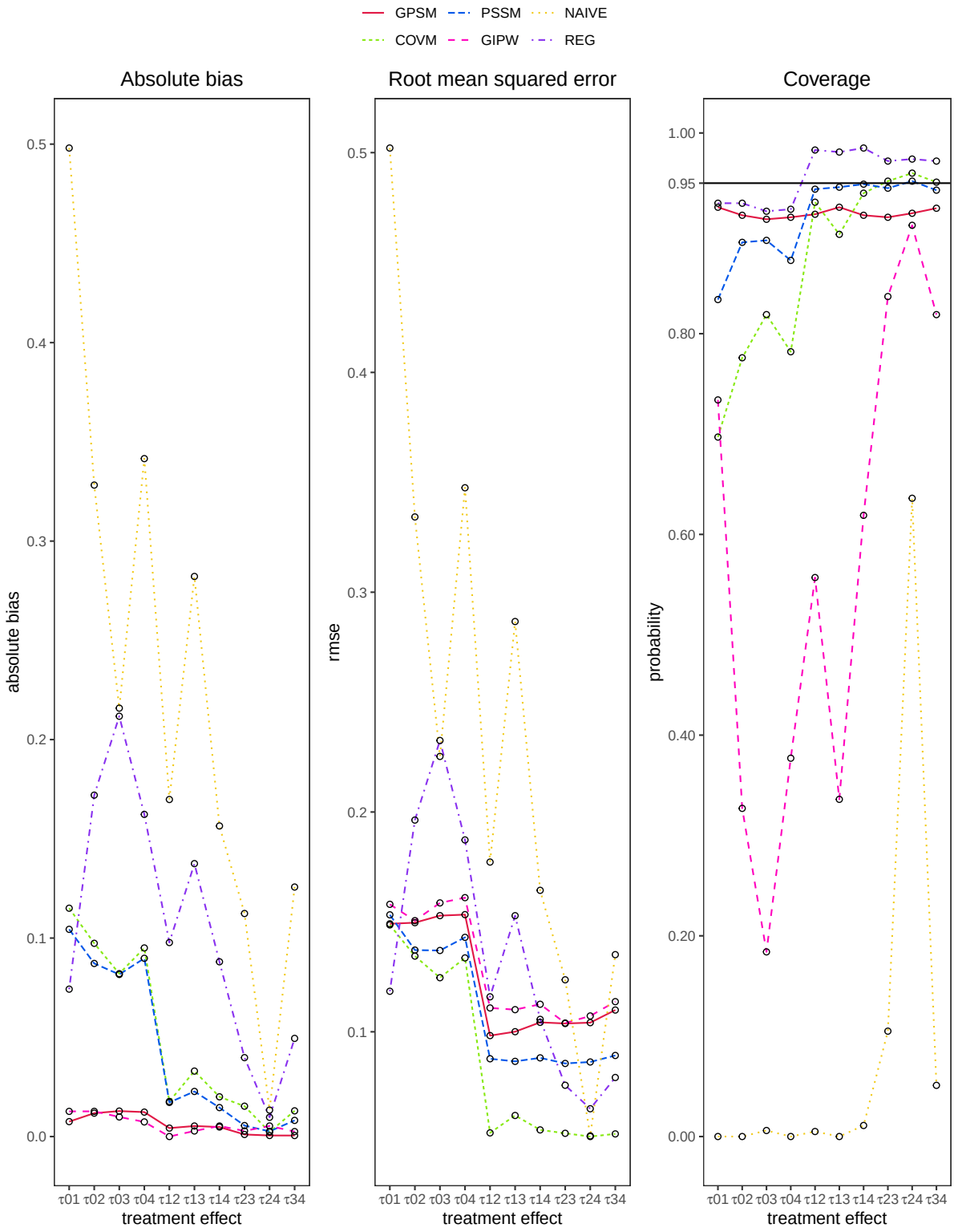


Figure 4.1: Simulation results when the GPS model is *correctly* specified.

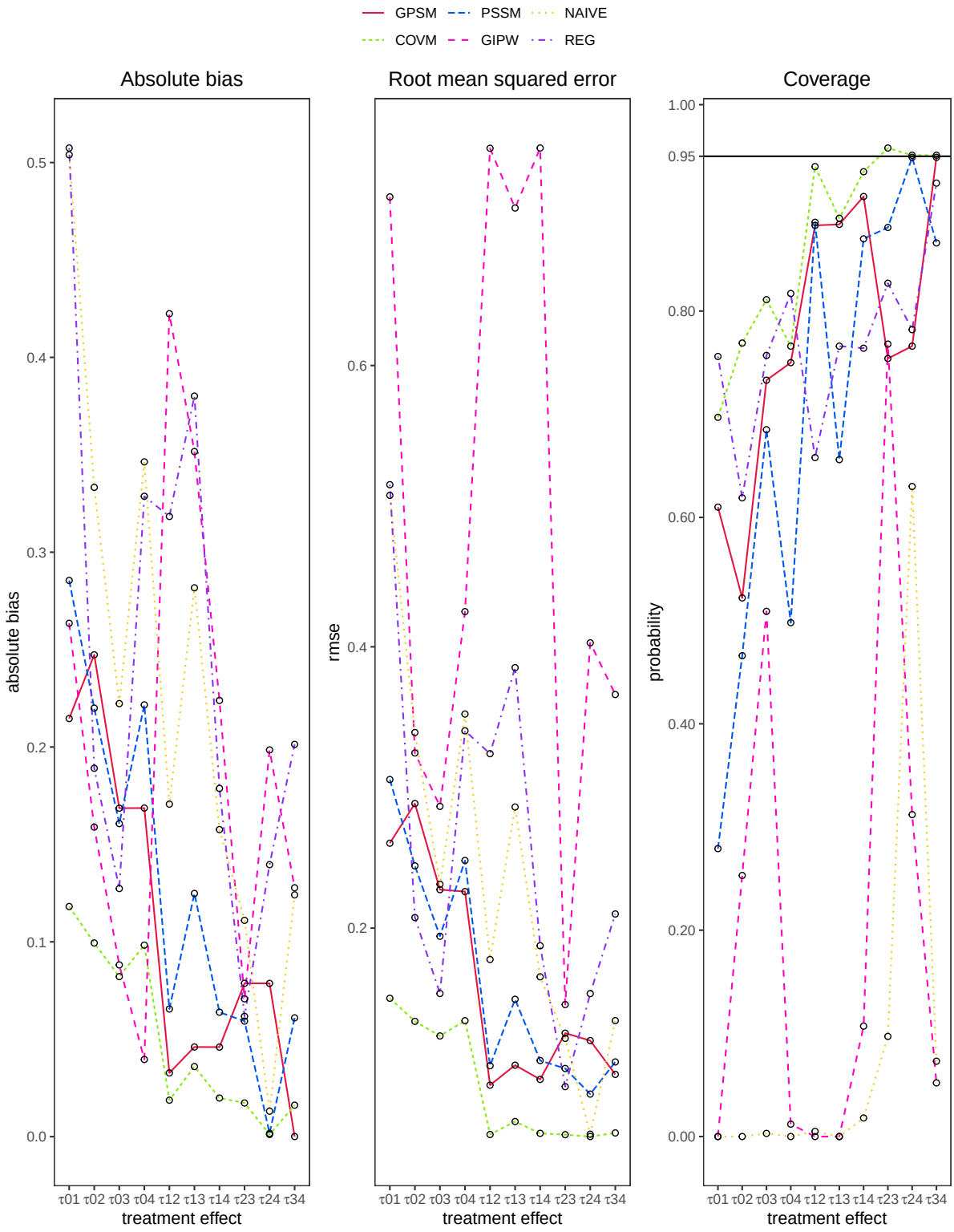


Figure 4.2: Simulation results when the GPS model is *incorrectly* specified.

4.4 Real Data Analysis

The most commonly diagnosed type of lung cancer is Non-small-cell lung cancer (NSCLC). Usually, around half the NSCLC patients receiving initial chemotherapy will go on to receive additional treatment in the post-progression setting, i.e., second-line treatment setting. In this setting, five treatments are historically commonly used, namely "pemetrexed", "carboplatin + paclitaxel", "erlotinib", "docetaxel", and "gemcitabine"; see Cui et al. (2018). The objective of this section is to compare the GPSM estimator with existing estimators in the evaluation of the comparative effectiveness of these treatment options in the second-line treatment setting.

The IQVIA electronic medical records (EMR) data are deidentified observational patient-level clinical data with demographic and baseline clinical characteristics collected from medium and large community-based oncology practices across 50 states of the USA. The analysis data set contains a retrospective cohort of 10,634 eligible patients who were at least 18 years old and received at least two lines of therapy from 1 January 2007 to 31 December 2014; see Cui et al. (2018) for details. For ease of illustration in this application, we use a subset of the eligible patients who received the five most common treatments. This subset contains a cohort of 5222 eligible patients. Overall survival is defined as the time from the start date of second-line therapy to the death date. Due to the limited availability of reported death dates (12.8%), a proxy death date using the last visit in the EMR data is determined for each study patient. Patients alive at the end of the time period are censored at the end date of the dataset. In addition, patients who do not have sufficient time between their last appointment and the conclusion of the data are censored at their last visit. Missing data are classified into a new category for each variable of interest and no imputation is applied; see de Rooij (2018). Among the 5,222 eligible patients in the data set, pemetrexed is assigned to 1,770 patients, 1,241 patients receive carboplatin + paclitaxel, 895 patients receive single-agent erlotinib, 739 receive docetaxel, and 577 receive gemcitabine as second-line therapy.

For estimation of the generalized propensity scores, we use a multinomial logistic regression model with the following baseline covariates used as predictors: age at the initiation of second-line therapy, gender, race, region, disease stage at initial diagnosis, Eastern Cooperative Oncology Group performance status score at the initiation of second-line therapy, facility types of academic or community cancer centers, year of index diagnosis, and days from index diagnosis to initiation of second-line therapy. To improve overlap in the covariate distributions for more credible matching, we apply the trimming method

described in Yang et al. (2016) to drop patients with extremely high or low GPS from the analysis. The remaining sample consists of 1495 patients receiving pemetrexed, 1035 receiving carboplatin + paclitaxel, 873 receiving erlotinib, 647 receiving docetaxel, and 509 receiving gemcitabine. We continue with remaining patients in the five cohorts for comparative effectiveness analysis on the population represented by the trimmed sample.

Table 4.2 reports the pairwise hazard ratio estimates and their corresponding 95% confidence intervals among the five treatment levels based on the unadjusted estimator `Naive`, the regression estimator `REG`, the inverse probability weighting estimator `GIPW`, matching estimator based on the full set of covariates `COVM`, and matching estimator based on the GPS vector `PSSM`, and our proposed `GPSM` estimator. The number of matches equals $M = 1$ for all the matching estimators. For `Naive`, `REG`, and `GIPW`, the robust variance estimation from the standard software is used for constructing Wald confidence intervals. Variance estimation for the `GPSM` estimator is the one proposed in Section 4.2.5. Based on the 95% confidence intervals, the `GPSM` estimator reports that six of ten (60%) pairwise comparisons show statistically significant differences. The unadjusted naive method determines that all contrasts are statistically significant, which likely hints at its poor credibility. For most contrasts, `COVM`, `REG`, and `PSSM` also result in point estimates similar to the naive approach, which suggests their tendency to incur bias.

4.5 Discussion and Future Studies

In this article, we propose the generalized propensity score matching estimator of the causal pairwise hazard ratios and study its asymptotic properties. We also propose an asymptotic variance estimator that takes into account the adjustment for estimating the GPS. We note that our approach assumes no unmeasured confounding and informative censoring, which are not testable, so it would be interesting to develop appropriate sensitivity analyses.

The methodology introduced in this paper is based on the generalized propensity score matching approach formalized by Yang et al. (2016). The credibility of our proposed approach, as with other GPS-based methods, relies on a correct specification of the GPS model, which is typically unknown to researchers. To mitigate this issue, model selection tools tailored to GPS matching have been developed for optimizing the performance of the GPS matching estimator (Zhao et al. 2022). When the outcome is continuous, multi-level double score matching provides a more robust alternative to GPS matching,

Table 4.2: IQVIA EMR data analysis results. Pem: Pemetrexed; CP: Carboplatin + paclitaxel; Erl: Erlotinib; Doc: Docetaxel; Gem: Gemcitabine. "Pem vs CP", for instance, is the hazard ratio CP/Pem. Thus, a value smaller than one implies that receiving Carboplatin + paclitaxel results in better survival than Pemetrexed. 95% confidence intervals are calculated using the same methods as described in the simulation study section.

	Pairwise hazard ratios (95% Confidence Intervals)				
Method	Pem vs CP	Pem vs Erl	Pem vs Doc	Pem vs Gem	CP vs Erl
`Naive`	0.79 (0.76, 0.83)	0.90 (0.86, 0.94)	1.05 (1.00, 1.09)	1.23 (1.18, 1.28)	1.13 (1.09, 1.18)
`REG`	0.80 (0.72, 0.88)	0.85 (0.53, 1.35)	1.01 (0.86, 1.18)	1.20 (1.05, 1.38)	1.06 (0.67, 1.68)
`COVM`	0.78 (0.70, 0.87)	0.85 (0.72, 1.01)	0.92 (0.81, 1.06)	1.18 (1.03, 1.36)	1.09 (0.92, 1.29)
`PSSM`	0.77 (0.69, 0.86)	0.89 (0.74, 1.07)	0.97 (0.75, 1.27)	1.28 (1.05, 1.57)	1.15 (0.95, 1.39)
`GPSM`	0.73 (0.63, 0.85)	0.86 (0.72, 1.02)	0.95 (0.84, 1.08)	1.40 (1.19, 1.65)	1.17 (0.97, 1.42)
`GIPW`	0.79 (0.72, 0.86)	0.75 (0.68, 0.83)	0.98 (0.88, 1.10)	1.20 (1.07, 1.34)	0.95 (0.86, 1.05)
	CP vs Doc	CP vs Gem	Erl vs Doc	Erl vs Gem	Doc vs Gem
`Naive`	1.32 (1.26, 1.37)	1.54 (1.48, 1.61)	1.16 (1.12, 1.21)	1.36 (1.31, 1.42)	1.17 (1.12, 1.22)
`REG`	1.26 (1.08, 1.47)	1.51 (1.32, 1.73)	1.19 (0.74, 1.91)	1.42 (0.88, 2.28)	1.20 (0.99, 1.44)
`COVM`	1.18 (1.03, 1.36)	1.52 (1.32, 1.75)	1.08 (0.90, 1.31)	1.39 (1.15, 1.68)	1.28 (1.09, 1.51)
`PSSM`	1.26 (0.94, 1.68)	1.66 (1.35, 2.05)	1.09 (0.80, 1.49)	1.44 (1.12, 1.87)	1.32 (0.97, 1.80)
`GPSM`	1.30 (1.11, 1.52)	1.91 (1.57, 2.32)	1.11 (0.92, 1.33)	1.63 (1.32, 2.01)	1.47 (1.23, 1.75)
`GIPW`	1.25 (1.12, 1.39)	1.52 (1.36, 1.70)	1.31 (1.18, 1.46)	1.60 (1.42, 1.81)	1.22 (1.08, 1.38)

while maintaining the protection against extreme values of the GPS (Yang and Zhang 2023; Zhao et al. 2022). Additionally, insufficient overlap between covariate distributions in different treatment levels could also hinder the performance of our proposed method. To address this issue, trimming the GPS has been shown to reduce variance when the outcome variable is continuous (Crump et al. 2009; Yang et al. 2016; Yang and Ding 2018). It would be an interesting future direction to adapt the trimming criterion to the survival context.

The current study focuses on the causal pairwise hazard ratios, which summarize the pairwise treatment effects among different treatment levels over a certain period of time. In the same point-exposure multi-level treatment setting, there are many other estimands of interest, such as the survival probability causal effect, the restricted average causal effect, the unrestricted mean potential survival times, and the conditional average treatment effect between treatment levels (Zeng et al. 2021; Hu et al. 2022a). In our future work, we will focus on extending our current matching framework to target some of these other causal estimands.

CHAPTER

5

DISCUSSION AND FUTURE RESEARCH

In Chapter 2, the optimal of set of covariates to be included in the GPS model is determined mostly based on empirical observations (Brookhart et al. 2006) and the theoretical result established for the augmented inverse probability weighting estimator (Hahn 2004; Tang et al. 2022). In the future, it would be useful to show the asymptotic variance of the GPS matching estimator adjusting for only outcome related covariates is smaller than adjusting for any other set of covariates.

The main results established in the previous chapters assume that the number of matches per unit M is fixed. Therefore a natural curiosity is how to choose M optimally. The choice of M corresponds to a bias-variance trade-off. Choosing a small M reduces finite sample biases caused by matching discrepancy though larger values of M produce lower asymptotic variances. In practice, it is often recommended in the literature to set M to a small number (e.g. $M = 1$) so that bias is minimized. This convenient choice also avoids the potential increase in computational cost associated with selecting M when the data volume is significant. Nonetheless, the optimal choice of M remains an open question and a direction for future work. Data-driven methods such as cross-validation could be used to select M with appropriately chosen criteria that optimize for estimation of the causal effect estimand. A recent work by Liu and Qin (2022) proposed a tuning free

PSM based on the non-parametric maximum-likelihood estimation of the propensity score under the monotonicity constraint. This approach does not require tuning M because all matches are guaranteed to be exact under the piece-wise constant estimated propensity score. A similar idea also appeared in Xu and Otsu (2022).

Another direction of future research is to generalize the proposed matching estimators to settings where the treatment or exposure is continuous. One important application is to inform air pollution policy, where the goal is to assess whether and in what magnitude exposure to air pollution is causally linked to adverse health outcomes. The main difference in the continuous exposure setting is that, there is no explicit way to distinguish units by treatment levels. Recently, Wu et al. (2022) proposed a matching algorithm for estimation of the population average exposure-response function (ERF). Their main idea is to divide the exposure values into blocks of equal size, and for each individual, GPS matching is used to find closest matches with observed treatment values in all other blocks. By doing so, the missing potential outcomes in different treatment blocks could be imputed. In addition, they showed that when the exposure levels are forced to be discrete and with appropriately chosen hyperparamters, their method collapses to the GPS matching estimator under the categorical treatment setting (Abadie and Imbens 2006; Yang et al. 2016). Under their matching framework, we conjucture that OABM may still be helpful for improving efficiency when treating each block as a distinct treatment level, provided the number of blocks is not prohibitively large. Generalizing the double score matching estimator to the continuous treatment setting would require showing that the local weak unconfoundedness assumption holds conditioning jointly on the GPS and the generalized prognostic score.

The parametric assumptions of the GPS could be restrictive. Besides the methods proposed in Chapter 2 and 3, machine learning algorithms also have the potential for improving robustness against misspecification of the GPS model. Using data-adaptive algorithms has the advantage that variable and model selection often occur automatically without the need to propose multiple candidate models, although more tuning parameters are introduced. In the empirical literature, nonparametric machine learning algorithms such as generalized boosted models (GBM), recursive partitioning, neural nets, and super learners have been proposed (Ju et al. 2019a; McCaffrey et al. 2013, 2004; Setoguchi et al. 2008). When outcome information is available, the loss function could be adapted to optimize covariate balance rather than prediction accuracy of treatment assignment, resulting in more precise treatment effect estimates (Pirracchio and Carone 2018). However, when the GPS is estimated by generic machine learning models, the asymptotic properties

of GPS matching estimators considered in this dissertation are no longer guaranteed, since the asymptotic properties of most machine learning algorithms are unknown.

Even though GPS matching is more robust to extreme values of the estimated propensity scores than inverse probability weighting, it still requires the true propensity scores to be bounded away from zero (ie, the positivity assumption must hold). When there is strong evidence that there is lack of overlap among the covariate distributions, a common remedy is to drop/trim the units with the extreme propensity scores and restrict analysis to the trimmed sample. As a result, trimming generally alters the estimand by changing the reference population. When treatment is binary, Crump et al. (2009) proposed a systematic approach to trim the sample using a threshold that minimizes the variance of the estimated average treatment effect on the trimmed sample. Yang et al. (2016) extended Crump's approach to the multi-level treatment case. However, the nonsmooth nature of trimming via a threshold renders the target causal estimand not root-n estimable. To address this issue, Yang and Ding (2018) proposed to use a smooth weight function to approximate the existing sample trimming, which allows for characterization of asymptotic properties of weighting estimators adjusting for the effect of trimming. In the future, it would also be interesting to quantify the limiting distributions of propensity score matching estimators adjusting for trimming of the propensity score.

REFERENCES

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

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

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

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843.

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

Adusumilli, K. (2018). Bootstrap inference for propensity score matching. Technical report, Working paper.

Ali, M. S., Groenwold, R. H., Pestman, W. R., Belitser, S. V., Roes, K. C., Hoes, A. W., de Boer, A., and Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology and drug safety*, 23(8):802–811.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10:1100–1120.

Andreou, E. and Werker, B. J. (2012). An alternative asymptotic analysis of residual-based statistics. *Review of Economics and Statistics*, 94(1):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. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107.

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

Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in medicine*, 33(7):1242–1258.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4):734–753.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H., De Boer, A., and Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and drug safety*, 20(11):1115–1129.

Bennett, M., Vielma, J. P., and Zubizarreta, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *Journal of computational and graphical statistics*, 29(4):744–757.

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-Interscience.

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.

Brown, D. W., DeSantis, S. M., Greene, T. J., Maroufy, V., Yaseen, A., Wu, H., Williams, G., and Swartz, M. D. (2020). A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study. *Statistics in medicine*, 39(17):2308–2323.

Bryer, J. (2017). *TriMatch: Propensity Score Matching of Non-Binary Treatments*. R package version 0.9.9.

Caruana, E., Chevret, S., Resche-Rigon, M., and Pirracchio, R. (2015). A new weighted balance measure helped to select the variables to be included in a propensity score model. *Journal of clinical epidemiology*, 68(12):1415–1422.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of official statistics*, 16(2):113.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96(453):260–269.

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

Cole, S. R. and Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

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

Cui, Z. L., Hess, L. M., Goodloe, R., and Faries, D. (2018). Application and comparison of generalized propensity score matching versus pairwise propensity score matching. *Journal of comparative effectiveness research*, 7(09):923–934.

de Rooij, M. (2018). Transitional modeling of experimental longitudinal data with missing values. *Advances in Data Analysis and Classification*, 12:107–130.

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

Fleming, T. R. and Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Ford, B. L. (1983). An overview of hot-deck procedures: Incomplete data in sample surveys. *Vol II: theory and bibliographies*, pages 185–206.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1):77–90.

Gayat, E., Resche-Rigon, M., Mary, J.-Y., and Porcher, R. (2012). Propensity score applied to survival data analysis through proportional hazards models: a monte carlo study. *Pharmaceutical statistics*, 11(3):222–229.

Greifer, N. and Stuart, E. A. (2021). Matching methods for confounder adjustment: an addition to the epidemiologist's toolbox. *Epidemiologic reviews*, 43(1):118–129.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics*, 86(1):73–76.

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.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pages 561–570.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

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

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

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. (2020). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical methods in medical research*, 29(11):3218–3234.

Hu, L., Ji, J., Ennis, R. D., and Hogan, J. W. (2022a). A flexible approach for causal inference with multiple treatments and clustered survival outcomes. *Statistics in Medicine*, 41(25):4982–4999.

Hu, L., Li, F., Ji, J., Joshi, H., and Scott, E. (2022b). Estimating the causal effects of multiple intermittent treatments with application to covid-19. *ArXiv*.

Huang, M.-Y. and Yang, S. (2022). Robust inference of conditional average treatment effects using dimension reduction. *Statistica Sinica*, 32:547–567.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

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

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

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

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology*, 150(4):327–333.

Ju, C., Combs, M., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., and van der Laan, M. J. (2019a). Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *Journal of Applied Statistics*, 46(12):2216–2236.

Ju, C., Wyss, R., Franklin, J. M., Schneeweiss, S., Häggström, J., and van der Laan, M. J. (2019b). Collaborative-controlled lasso for constructing propensity score-based estimators in high-dimensional data. *Statistical methods in medical research*, 28(4):1044–1063.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4).

King, G. and Nielsen, R. (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(1):1–10.

Le Cam, L., LeCam, L. M., and Yang, G. L. (2000). *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media.

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.

Lechner, M. (2001). *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*. Springer.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389 – 2415.

Liang, M. and Yu, M. (2022). A semiparametric approach to model effect modification. *Journal of the American Statistical Association*, 117(538):752–764.

Linden, A., Uysal, S. D., Ryan, A., and Adams, J. L. (2016). Estimating causal effects for multivalued treatments: a comparison of approaches. *Statistics in Medicine*, 35(4):534–552.

Little, R. J. A. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. Wiley & Sons, Incorporated, John.

Liu, Y. and Qin, J. (2022). Tuning-parameter-free optimal propensity score matching approach for causal inference. *arXiv preprint arXiv:2205.13200*.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

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

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

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.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2020). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*, 115(529):307–317.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., and Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and drug safety*, 20(6):551–559.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, 40(1):75–149.

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

Pirracchio, R. and Carone, M. (2018). The balance super learner: a robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical methods in medical research*, 27(8):2504–2518.

Rassen, J. A., Shelat, A. A., Franklin, J. M., Glynn, R. J., Solomon, D. H., and Schneeweiss, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*, pages 401–409.

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):409–424.

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.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

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

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *The Journal of Machine Learning Research*, 21(1):7642–7727.

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

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

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

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36.

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

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., and D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, 33(2):155–161.

Schuler, M. S. and Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1):65–73.

Scotina, A. D., Beaudoin, F. L., and Gutman, R. (2020). Matching estimators for causal effects of multiple treatments. *Statistical methods in medical research*, 29(4):1051–1066.

Scotina, A. D. and Gutman, R. (2019). Matching algorithms for causal inference with multiple treatments. *Statistics in medicine*, 38(17):3139–3167.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Shu, D., Han, P., Wang, R., and Toh, S. (2021). Estimating the marginal hazard ratio by simultaneously using a set of propensity score models: A multiply robust approach. *Statistics in Medicine*, 40(5):1224–1242.

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

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794.

Tang, D., Kong, D., Pan, W., and Wang, L. (2022). Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics*.

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*.

Tchetgen Tchetgen, E. J. and Robins, J. (2012). On parametrization, robustness and sensitivity analysis in a marginal structural cox proportional hazards model for point exposure. *Statistics & Probability Letters*, 82(5):907–915.

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

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

van der Vaart (2000). *Asymptotic Statistics*, volume 3. Cambridge university press, Cambridge: Cambridge University Press.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

Vansteelandt, S. and Daniel, R. M. (2014). On regression adjustment for the propensity score. *Statistics in medicine*, 33(23):4053–4072.

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

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., and Stürmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety*, 20(3):317–320.

Wu, X., Mealli, F., Kioumourtzoglou, M.-A., Dominici, F., and Braun, D. (2022). Matching on generalized propensity scores with continuous exposures. *Journal of the American Statistical Association*, pages 1–29.

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.

Xu, M. and Otsu, T. (2022). Isotonic propensity score matching. *arXiv preprint arXiv:2207.08868*.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):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(4):1055–1065.

Yang, S. and Kim, J. K. (2020). Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scandinavian Journal of Statistics*, 47(3):839–861.

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.

Yang, S. and Zhang, Y. (2023). Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*, 50(1):235–265.

Yang, S., Zhang, Y., Liu, G. F., and Guan, Q. (2021). Smim: A unified framework of survival sensitivity analysis using multiple imputation and martingale. *Biometrics*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73.

Zeng, S., Li, F., and Hu, L. (2021). Propensity score weighting analysis of survival outcomes using pseudo-observations. *arXiv preprint arXiv:2103.00605*.

Zhang, Y., Yang, S., Ye, W., Faries, D. E., Lipkovich, I., and Kadziola, Z. (2021). Practical recommendations on double score matching for estimating causal effects. *Statistics in Medicine*, 41(8):1421–1445.

Zhao, H. and Yang, S. (2022). Outcome-adjusted balance measure for generalized propensity score model selection. *Journal of Statistical Planning and Inference*, 221:188–200.

Zhao, H., Zhang, X., and Yang, S. (2022). Double score matching in observational studies with multi-level treatments. *Communications in Statistics-Simulation and Computation*, pages 1–17.

APPENDICES

APPENDIX

A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1 Notation and Assumptions

For simplification of illustration, in this document, we slightly abuse the notation by dropping the superscript $\mathcal{S}$ from $\boldsymbol{X}^{\mathcal{S}}$, so now $\boldsymbol{X}$ refers to a generic subset of all baseline covariates. Define the conditional mean and variance of a single covariate X given $p\left(w \mid \boldsymbol{X}\right)$:

$$\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}\right)\right\} = E\left\{X \mid p(w \mid \boldsymbol{X})\right\},$$

$$\bar{\sigma}^2_X\left\{p\left(w \mid \boldsymbol{X}\right)\right\} = Var\left\{X \mid p(w \mid \boldsymbol{X})\right\}.$$

Define the conditional means and variances of $Y(w)$ given covariates $\boldsymbol{X}$ and given $p\left(w \mid \boldsymbol{X}\right)$:

$$\mu(w, \boldsymbol{x}) = E\left\{Y(w) \mid W = w, \boldsymbol{X} = \boldsymbol{x}\right\},$$

$$\bar{\mu}(w, p) = E\left\{Y(w) \mid W = w, p\left(w \mid \boldsymbol{X}\right) = p\right\},$$

$$\sigma^2(w, \boldsymbol{x}) = Var\left\{Y(w) \mid W = w, \boldsymbol{X} = \boldsymbol{x}\right\},$$

$$\bar{\sigma}^2(w, p) = Var\left\{Y(w) \mid W = w, p\left(w \mid \boldsymbol{X}\right) = p\right\}.$$

We let $\rho_{w,N}(X, Y) = \rho_N\left(X, Y \mid W = w\right)$ denote the measure of correlation between X and Y conditional on $W = w$. Let $p^* = p(w \mid \boldsymbol{X}^{\mathcal{C}}, \boldsymbol{X}^{\mathcal{P}})$ denote the optimal generalized propensity score model. To show dependence of $\widehat{X}_{\text{gpsm}}(w)$ on a GPS model $p(w|\boldsymbol{X})$ explicitly and to simplify notation, we let $\widehat{X}(w, p(w|\boldsymbol{X})) = \widehat{X}_{\text{gpsm}}(w)$. For convenience, also let $\rho_{w,N}(X, Y)$ denote $\rho_N(X, Y|W = w)$.

Assumption 12 *$p(w \mid \boldsymbol{X})$ has a continuous distribution with compact support $[\underline{p}, \overline{p}]$ with a continuous density function. $\bar{\mu}_X(w, p)$ is Lipschitz-continuous in p. For some $\delta > 0$, $E\{|X|^{2+\delta}|W = w, p(w \mid \boldsymbol{X}) = p\}$ is uniformly bounded.*

Assumption 13 *We have a random sample of size N from a large population.*

We include a lemma and a theorem, which will be useful for proving Lemma 2. The lemma is proven by Abadie and Imbens (2016). Theorem 7 is the Central Limit Theorem for martingale arrays (e.g. Billingsley (1995)).

Lemma 7 *Suppose that $(W_1, X_1), ..., (W_N, X_N)$ are independent and identically distributed, where the density of X is continuous on $[a, b]$, and $Pr(W = w) > 0$ for $w \in \mathbb{W}$. Assume also that $\sigma^2(w, X_i)$ is uniformly bounded. For a given w, let $p^* = Pr(W = w)$, we have*

$$\frac{1}{N_w}\sum_{i=1}^{N} D_i(w)\sigma^2\left(w, X_i\right) K(i, w) \xrightarrow{p} E\left\{\sigma^2\left(w, X_i\right)\left(\frac{p^*}{1-p^*}\right)^{1-2D_i(w)} \mid W_i = w\right\}$$

and

$$\frac{1}{N_w}\sum_{i=1}^{N} D_i(w)\sigma^2\left(w, X_i\right) K(i, w)^2 \xrightarrow{p} E\left\{\sigma^2\left(w, X_i\right)\left(\frac{p^*}{1-p^*}\right)^{1-2D_i(w)} \mid W_i = w\right\} + \frac{3}{2}E\left\{\sigma^2\left(w, X_i\right)\left(\frac{p^*}{1-p^*}\right)^{2(1-2D_i(w))} \mid W_i = w\right\}$$

Theorem 7 *Let $X_{n,k}$, $1 \leq k \leq m_n$ be a martingale array with respect to $F_{n,k}$ and let $S_{n,k} = \sum_{i=1}^{k} X_{n,i}$. If $E_{\max_{1\leq j\leq m_n}}|X_{n,j}| \to 0$ or $\sum_{j=1}^{m_n} E|X_{n,j}|^{2+\delta} \to 0$ for some $\delta > 0$ and $\sum_{j=1}^{m_n} X_{n,j}^2 \xrightarrow{p} \sigma^2$, then $S_{n,m_n} \xrightarrow{d} N(0, \sigma^2)$.*

A.2 Proof of Lemma 2

We first prove Lemma 2. Recall that $D_i(w)$ is the indicator of whether unit i received treatment w, and $K(i,w)$ is the number of times unit i is used as a match when each unit is matched with replacement to one closest unit at treatment level w. By definition of the matching function, we can express the sum of imputed quantities in terms of $D_i(w)$, $K(i,w)$ and the original values, and vice versa:

$$\sum_{i=1}^{N} X_{m\{w,p(w|\boldsymbol{X}_i)\}} = \sum_{i=1}^{N} D_i(w)\left\{1+K(i,w)\right\}X_i \tag{A.1}$$

$$\sum_{i=1}^{N}\left\{1-D_i(w)\right\}\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_{m\{w,p(w|\boldsymbol{X}_i)\}}\right)\right\} = \sum_{i=1}^{N} D_i(w)K(i,w)\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\} \tag{A.2}$$

We follow a similar argument given by Abadie and Imbens (2016) and Yang et al. (2016). Using the two properties shown in Equation (A.1) and (A.2), we obtain

$$\begin{aligned}
&\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_{m\{w,p(w|\boldsymbol{X}_i)\}} - \overline{X}\right)\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i(w)\left\{1+K(i,w)\right\}\left[X_i - \bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\}\right]\\
&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\} - X_i\right]\\
&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{1-D_i(w)\right\}\left[\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_{m\{w,p(w|\boldsymbol{X}_i)\}}\right)\right\} - \bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\}\right]\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[D_i(w)\left\{1+K(i,w)\right\}-1\right]\left[X_i - \bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\}\right]\\
&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{1-D_i(w)\right\}\left[\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_{m\{w,p(w|\boldsymbol{X}_i)\}}\right)\right\} - \bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\}\right]\\
=&A_N + R_N
\end{aligned}$$

where

$$A_N = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[D_i(w)\left\{1+K(i,w)\right\}-1\right]\left[X_i - \bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_i\right)\right\}\right]$$

and

$$R_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{1 - D_i(w)\} \left[\bar{\mu}_X \left\{ p\left(w \mid \boldsymbol{X}_{m\{w,p(w|\boldsymbol{X}_i)\}} \right) \right\} - \bar{\mu}_X \left\{ p\left(w \mid \boldsymbol{X}_i \right) \right\} \right].$$

We rewrite A_N as

$$A_N = \sum_{k=1}^{N} \xi_{N,k},$$

where

$$\xi_{N,k} = \frac{1}{\sqrt{N}} \left[D_k(w) \left\{ 1 + K(k,w) \right\} - 1 \right] \left[X_k - \bar{\mu}_X \left\{ p\left(w \mid \boldsymbol{X}_k \right) \right\} \right],$$

for $1 \leq k \leq N$. Consider the σ-fields $F_{N,k} = \sigma\left\{ D_1(w), \ldots, D_N(w), p\left(w \mid \boldsymbol{X}_1 \right), \ldots, p\left(w \mid \boldsymbol{X}_N \right), X_1, ..., X_k \right\}$ for $1 \leq k \leq N$. Since we assumed that $D_i(w) \perp\!\!\!\perp X_i \mid p(w \mid \boldsymbol{X}_i)$, we have

$$E\left(\sum_{k=1}^{m} \xi_{N,k} \mid F_{N,m} \right) = \sum_{k=1}^{m-1} \xi_{N,k}$$

and so for each $N \geq 1$,

$$\left\{ \sum_{j=1}^{i} \xi_{N,j}, F_{N,i}, 1 \leq i \leq N \right\}$$

is a martingale. To obtain the asymptotic normality of A_N, we apply Theorem 7. The following two conditions are sufficient:

$$\sum_{k=1}^{N} E\left(|\xi|_{N,k}^{2+\delta} \right) \to 0 \text{ for some } \delta > 0 \tag{A.3}$$

and

$$\sum_{k=1}^{N} E\left(\xi_{N,k}^2 \mid F_{N,k-1} \right) \xrightarrow{p} E\left[\bar{\sigma}_X^2 \left\{ p\left(w \mid \boldsymbol{X} \right) \right\} \left\{ \frac{3}{2} \frac{1}{p\left(w \mid \boldsymbol{X} \right)} - \frac{1}{2} p\left(w \mid \boldsymbol{X} \right) - 1 \right\} \right]. \tag{A.4}$$

Equation (A.3) is the Lyapounov condition (which is sufficient for the usual Lindeberg condition to hold). Because of Assumption 12, $\bar{\mu}_X \left\{ p\left(w \mid \boldsymbol{X}_k \right) \right\}$ are continuous on a compact support, these functions are also bounded. Let $C_{\bar{\sigma}_X^{2+\delta}}$ be a bound on $E[|X_i - \bar{\mu}_X \left\{ W_i, p\left(w \mid \boldsymbol{X}_i \right) \right\}|^{2+\delta} \mid W_i, p\left(w \mid \boldsymbol{X}_i \right)]$ for $w \in \mathbb{W}$ and $p \in [\underline{p}, \overline{p}]$. Using the Law of Iterated Expectation and the fact that $K(k,w)$ has bounded moments (see Lemma

A.8 in Abadie and Imbens (2016)), we obtain

$$\sum_{k=1}^{N} E\left[|\xi_{N,k}|^{2+\delta}\right] = N^{-\delta/2} E\left[\left|D_k(w)\left\{1+K(k,w)\right\}-1\right|^{2+\delta}\left|X_k-\bar{\mu}_X\left\{p\left(w \mid \boldsymbol{X}_k\right)\right\}\right|^{2+\delta}\right]$$
$$\leq \frac{C_{\bar{\sigma}_X^{2+\delta}} E\left[\left\{1+K(k,w)\right\}^{2+\delta}\right]}{N^{\delta/2}} \to 0.$$

To prove equation (A.4) notice that

$$\sum_{k=1}^{N} E\left(\xi_{N,k}^2 \mid F_{N,k-1}\right) = \frac{1}{N}\sum_{k=1}^{N}\left[D_k(w)\left\{1+K(k,w)\right\}-1\right]^2 \bar{\sigma}_X^2\left\{p\left(w \mid \boldsymbol{X}_k\right)\right\}.$$

Then, Lemma 7 implies equation (A.4) after some algebra.

The term $R_N/\sqrt{N}$ is the conditional bias for covariate due to matching discrepancy. Theorem 1 from Abadie and Imbens (2016) states that under appropriate conditions, the conditional bias for potential outcome $Y(w)$ due to matching discrepancy $R_N(w)/\sqrt{N} = O_p\left(N^{-1/k}\right)$ where k is the number of continuous matching variables. In our case, $k=1$ and so we have $R_N = O_p\left(N^{-1/2}\right)$. This is sufficient for showing $R_N = o_p(1)$.

A.3 Proof of Theorem 1

We first prove Part (i). Without loss of generality, let $\boldsymbol{X} = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)})^{\mathrm{T}}$ be the vector of baseline covariates, where $X^{(1)}$ is a confounder, $X^{(2)}$ is an instrumental variable, $X^{(3)}$ is a precision variable, and $X^{(4)}$ is a null variable. By construction, we have $p^* = p(w \mid X^{(1)}, X^{(3)})$. For any specification $p_k \in \mathscr{P}$, we want to show

$$\mathbb{P}\{\texttt{OABM}_{\texttt{CAN}}(w,p^*) \leq \texttt{OABM}_{\texttt{CAN}}(w,p_k)\} \to 1.$$

It suffices to show

$$\mathbb{P}\left\{\sum_{j=1}^{4}\zeta_j(w,p_k)\left|\widehat{X}^{(j)}(w,p_k)-\overline{X}^{(j)}\right| \leq \sum_{i=1}^{4}\zeta_j(w,p^*)\left|\widehat{X}^{(j)}(w,p^*)-\overline{X}^{(j)}\right|\right\} \tag{A.5}$$

converges to 0 as $N \to \infty$.

By Lemma 2, we can quickly obtain the following convergence in probability results

for $\texttt{OABM}_{\texttt{CAN}}(w, p^*) = A_1 + A_2 + A_3 + A_4$, where:

$$A_1 = \zeta_1(w, p^*)\left|\widehat{X}^{(1)}(w, p^*) - \overline{X}^{(1)}\right| = \left|\widehat{X}^{(1)}(w, p^*) - \overline{X}^{(1)}\right| / \rho_{w,N}(X^{(1)}, Y) \xrightarrow{p} 0$$

$$A_2 = \zeta_2(w, p^*)\left|\widehat{X}^{(2)}(w, p^*) - \overline{X}^{(2)}\right| = N^{1/3}\rho_{w,N}(X^{(2)}, Y)\left|\widehat{X}^{(2)}(w, p^*) - \overline{X}^{(2)}\right| \xrightarrow{p} 0$$

$$A_3 = \zeta_3(w, p^*)\left|\widehat{X}^{(3)}(w, p^*) - \overline{X}^{(3)}\right| = \left|\widehat{X}^{(3)}(w, p^*) - \overline{X}^{(3)}\right| / \rho_{w,N}(X^{(3)}, Y) \xrightarrow{p} 0$$

$$A_4 = \zeta_4(w, p^*)\left|\widehat{X}^{(4)}(w, p^*) - \overline{X}^{(4)}\right| = N^{1/3}\rho_{w,N}(X^{(4)}, Y)\left|\widehat{X}^{(4)}(w, p^*) - \overline{X}^{(4)}\right| \xrightarrow{p} 0.$$

The rest of the proof is divided into four parts, with each part corresponding to a type of misspecified model $p_k \neq p^*$.

Part 1: (Penalize the inclusion of IV) Consider model $p_k = p(w \mid X^{(1)}, X^{(2)}, X^{(3)})$, which results from adding the instrumental variable to p^*. Then $\texttt{OABM}_{\texttt{CAN}}(w, p_k) = B_1 + B_2 + B_3 + B_4$ where

$$B_1 = \zeta_1(w, p_k)\left|\widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)}\right| = \left|\widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)}\right| / \rho_{w,N}(X^{(1)}, Y) \xrightarrow{p} 0$$

$$B_3 = \zeta_3(w, p_k)\left|\widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)}\right| = \left|\widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)}\right| / \rho_{w,N}(X^{(3)}, Y) \xrightarrow{p} 0$$

$$B_4 = \zeta_4(w, p_k)\left|\widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)}\right| = N^{1/3}\rho_{w,N}(X^{(4)}, Y)\left|\widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)}\right| \xrightarrow{p} 0,$$

with the exception of

$$B_2 = \zeta_2(w, p_k)\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| = \left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| / \rho_{w,N}(X^{(2)}, Y).$$

We rewrite (A.5) as

$$\begin{aligned}
&\mathbb{P}\left\{B_2 \leq o_p(1)\right\} \\
=&\mathbb{P}\left\{\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| / \rho_{w,N}(X^{(2)}, Y) \leq o_p(1)\right\} \\
=&\mathbb{P}\left\{\sqrt{N}\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| - \sqrt{N}\rho_{w,N}(X^{(2)}, Y)o_p(1) \leq 0\right\}. \qquad \text{(A.6)}
\end{aligned}$$

By Lemma 2 and the continuous mapping theorem, we know that the first term

$$\sqrt{N}\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| \xrightarrow{d} \text{folded normal.}$$

By Slutsky's theorem, we have that the second term

$$\sqrt{N}\rho_{w,N}(X^{(2)},Y)o_p(1) \xrightarrow{p} 0.$$

Therefore (A.6)$\to$ 0 as N tends to infinity, implying that the balance measure will discourage the inclusion of IV in the GPS.

Part 2: (Penalize the inclusion of null variable) Proof for the case $p_k = p(w \mid X^{(1)}, X^{(3)}, X^{(4)})$ is almost identical to part 1.

Part 3: (Penalize the exclusion of precision variable) Consider model $p_k = p(w \mid X^{(1)})$, which results from dropping the precision variable from p^*. Then $\texttt{OABM}_{\texttt{CAN}}(w, p_k) = C_1 + C_2 + C_3 + C_4$ where

$$C_1 = \zeta_1(w,p_k)\left|\widehat{X}^{(1)}(w,p_k) - \overline{X}^{(1)}\right| = \left|\widehat{X}^{(1)}(w,p_k) - \overline{X}^{(1)}\right| / \rho_{w,N}(X^{(1)},Y) \xrightarrow{p} 0$$

$$C_2 = \zeta_2(w,p_k)\left|\widehat{X}^{(2)}(w,p_k) - \overline{X}^{(2)}\right| = N^{1/3}\rho_{w,N}(X^{(2)},Y)\left|\widehat{X}^{(2)}(w,p_k) - \overline{X}^{(2)}\right| \xrightarrow{p} 0$$

$$C_3 = \zeta_3(w,p_k)\left|\widehat{X}^{(3)}(w,p_k) - \overline{X}^{(3)}\right| = N^{1/3}\rho_{w,N}(X^{(3)},Y)\left|\widehat{X}^{(3)}(w,p_k) - \overline{X}^{(3)}\right| \xrightarrow{p} 0$$

$$C_4 = \zeta_4(w,p_k)\left|\widehat{X}^{(4)}(w,p_k) - \overline{X}^{(4)}\right| = N^{1/3}\rho_{w,N}(X^{(4)},Y)\left|\widehat{X}^{(4)}(w,p_k) - \overline{X}^{(4)}\right| \xrightarrow{p} 0.$$

Now multiplying $N^{1/6}$ to both sides of the inequality in (A.5) we get

$$\begin{aligned}&\mathbb{P}\left\{N^{1/6}C_3 \le o_p(1)\right\}\\ =&\mathbb{P}\left\{N^{1/2}\rho_{w,N}(X^{(3)},Y)\left|\widehat{X}^{(3)}(w,p_k) - \overline{X}^{(3)}\right| \le o_p(1)\right\}.\end{aligned} \tag{A.7}$$

This is because the remaining terms $A_1, ..., A_4, C_1, ..., C_3$, after multiplying by $N^{1/6}$, still converge in probability to 0. By Lemma 2, continuous mapping theorem, consistency of $\rho_{w,N}(X^{(3)},Y)$ and Slutsky's theorem, we get

$$N^{1/2}\rho_{w,N}(X^{(3)},Y)\left|\widehat{X}^{(3)}(w,p_k) - \overline{X}^{(3)}\right| \xrightarrow{d} \text{folded normal.}$$

Therefore, (A.7)$\to$ 0 as N tends to infinity, implying that the balance measure will discourage the exclusion of precision variables in the GPS.

Part 4: (Penalize the exclusion of confounder) Finally, consider $p_k = p(w \mid X^{(3)})$, which results from dropping the confounder from p^*. Then $\texttt{OABM}_{\texttt{CAN}}(w, p_k) = D_1 + D_2 +$

$D_3 + D_4$ where

$$D_2 = \zeta_2(w, p_k) \left| \widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)} \right| = N^{1/3} \rho_{w,N}(X^{(2)}, Y) \left| \widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)} \right| \xrightarrow{p} 0$$

$$D_3 = \zeta_3(w, p_k) \left| \widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)} \right| = \left| \widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)} \right| / \rho_{w,N}(X^{(3)}, Y) \xrightarrow{p} 0$$

$$D_4 = \zeta_4(w, p_k) \left| \widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)} \right| = N^{1/3} \rho_{w,N}(X^{(4)}, Y) \left| \widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)} \right| \xrightarrow{p} 0,$$

with the exception of

$$D_1 = \zeta_1(w, p_k) \left| \widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)} \right| = N^{1/3} \rho_{w,N}(X^{(1)}, Y) \left| \widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)} \right|.$$

The term D_1 is strictly positive and does not converge. Therefore $\mathbb{P}\{D_1 \leq o_p(1)\} \to 0$ as $N \to \infty$, which implies that the balance measure will discourage the exclusion of confounders in the GPS.

Any other the GPS model specification (e.g. $p_k = p(w \mid X^{(2)}, X^{(3)})$) can be represented as a combination of the four types of misspecification. In such cases, the result can be proven using similar techniques.

We now prove Part (ii). Without loss of generality, let $\boldsymbol{X} = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)})$ be the baseline covariates, where $X^{(1)}$ is a confounder, $X^{(2)}$ is an instrumental variable, $X^{(3)}$ is a precision variable, and $X^{(4)}$ is a null variable. Again, we want to show

$$\mathbb{P}\{\texttt{OABM}_{\texttt{BCor}}(w, p_k) \leq \texttt{OABM}_{\texttt{BCor}}(w, p^*)\} \to 0.$$

By the properties of ball correlation and Lemma 2, we can quickly obtain the following convergence in probability results for $\texttt{OABM}_{\texttt{BCor}}(w, p^*) = A_1 + A_2 + A_3 + A_4$, where

$$A_1 = BCor^{-1}_{w,N}(X^{(1)}, Y) \left| \widehat{X}^{(1)}(w, p^*) - \overline{X}^{(1)} \right| \xrightarrow{p} 0$$

$$A_2 = N^{1/3} BCor_{w,N}(X^{(2)}, Y) \left| \widehat{X}^{(2)}(w, p^*) - \overline{X}^{(2)} \right| \xrightarrow{p} 0$$

$$A_3 = BCor^{-1}_{w,N}(X^{(3)}, Y) \left| \widehat{X}^{(3)}(w, p^*) - \overline{X}^{(3)} \right| \xrightarrow{p} 0$$

$$A_4 = N^{1/3} BCor_{w,N}(X^{(4)}, Y) \left| \widehat{X}^{(4)}(w, p^*) - \overline{X}^{(4)} \right| \xrightarrow{p} 0$$

up to a multiplicative constant.

Part 1: (Exclusion of IV) Consider model $p_k = p(w|X^{(1)}, X^{(2)}, X^{(3)})$, which results

from adding instrumental variable to the optimal model p^*. Then $\texttt{OABM}_{\texttt{BCor}}(w, p_k) = B_1 + B_2 + B_3 + B_4$ where

$$B_1 = BCor_{w,N}^{-1}(X^{(1)}, Y)\left|\widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)}\right| \xrightarrow{p} 0$$

$$B_3 = BCor_{w,N}^{-1}(X^{(3)}, Y)\left|\widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)}\right| \xrightarrow{p} 0$$

$$B_4 = N^{1/3} BCor_{w,N}(X^{(4)}, Y)\left|\widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)}\right| \xrightarrow{p} 0,$$

with the exception of

$$B_2 = BCor_{w,N}^{-1}(X^{(2)}, Y)\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right|.$$

We can rewrite (A.5) as

$$\begin{aligned}
&\mathbb{P}\left\{B_2 \leq o_p(1)\right\} \\
=&\mathbb{P}\left\{\frac{\sqrt{BCov_{w,N}(X^{(2)}, X^{(2)})BCov_{w,N}(Y, Y)}}{BCov_{w,N}(X^{(2)}, Y)}\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| \leq o_p(1)\right\} \\
=&\mathbb{P}\{\sqrt{N}\sqrt{BCov_{w,N}(X^{(2)}, X^{(2)})BCov_{w,N}(Y, Y)}\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| \qquad \text{(A.8)} \\
&\leq \sqrt{N} BCov_{w,N}(X^{(2)}, Y) o_p(1)\}. \qquad \text{(A.9)}
\end{aligned}$$

Since $BCov_{w,N}(X^{(2)}, X^{(2)})BCov_{w,N}(Y, Y) \xrightarrow{p} BCov(X^{(2)}, X^{(2)})BCov(Y, Y)$, by Lemma 2 and continuous mapping theorem, we have

$$\sqrt{N}\sqrt{BCov_{w,N}(X^{(2)}, X^{(2)})BCov_{w,N}(Y, Y)}\left|\widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)}\right| \xrightarrow{d} \text{folded normal.}$$

By assumption, we also have $\sqrt{N} BCov_{w,N}(X^{(2)}, Y) = o_p(1)$. Therefore (A.9)$\to 0$ as N tends to infinity, implying that the instrumental variable will be excluded from the PS model.

Part 2: (Exclusion of null variable) Proof for the case $p_k = p(w|X^{(1)}, X^{(2)}, X^{(4)})$ is almost identical to that for part 1.

Part 3: (Inclusion of precision variable) Consider model $p_k = p(w|X^{(1)})$, which results from dropping the precision variable from the optimal model p^*. Then $\texttt{OABM}_{\texttt{BCor}}(w, p_k) = C_1 + C_2 + C_3 + C_4$ where

$$C_1 = BCor_{w,N}^{-1}(X^{(1)}, Y)\left|\widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)}\right| \xrightarrow{p} 0$$

$$C_2 = N^{1/3} BCor_{w,N}(X^{(2)}, Y) \left| \widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)} \right| \xrightarrow{p} 0$$

$$C_3 = N^{1/3} BCor_{w,N}(X^{(3)}, Y) \left| \widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)} \right| \xrightarrow{p} 0$$

$$C_4 = N^{1/3} BCor_{w,N}(X^{(4)}, Y) \left| \widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)} \right| \xrightarrow{p} 0,$$

again up to a multiplicative constant. Now multiplying $N^{1/6}$ to both sides of the inequality in (A.5), we get

$$\begin{aligned} &\mathbb{P}\left\{ N^{1/6} C_3 \leq o_p(1) \right\} \\ =&\mathbb{P}\left\{ \sqrt{N} BCor_{w,N}(X^{(3)}, Y) \left| \widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)} \right| \leq o_p(1) \right\}. \end{aligned} \tag{A.10}$$

This is because the remaining terms $A_1, ..., A_4, C_1, C_2, C_4$, after multiplying by $N^{1/6}$, still converge in probability to 0. By Lemma 2 and continuous mapping theorem again, we know

$$\sqrt{N} BCor_{w,N}(X^{(3)}, Y) \left| \widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)} \right| \xrightarrow{d} \text{folded normal.}$$

Therefore, (A.10)$\to 0$ as N tends to infinity, implying that the precision variable will be included in the PS model.

Part 4: (Inclusion of confounder) Consider model $p_k = p(w|X^{(3)})$, which results from dropping the confounder from the optimal model p^*. Then $\texttt{OABM}_{\texttt{BCor}}(w, p_k) = B_1 + B_2 + B_3 + B_4$ where

$$D_2 = N^{1/3} BCor_{w,N}(X^{(2)}, Y) \left| \widehat{X}^{(2)}(w, p_k) - \overline{X}^{(2)} \right| \xrightarrow{p} 0$$

$$D_3 = BCor^{-1}_{w,N}(X^{(3)}, Y) \left| \widehat{X}^{(3)}(w, p_k) - \overline{X}^{(3)} \right| \xrightarrow{p} 0$$

$$D_4 = N^{1/3} BCor_{w,N}(X^{(4)}, Y) \left| \widehat{X}^{(4)}(w, p_k) - \overline{X}^{(4)} \right| \xrightarrow{p} 0,$$

with the exception of

$$D_1 = N^{1/3} BCor_{w,N}(X^{(1)}, Y) \left| \widehat{X}^{(1)}(w, p_k) - \overline{X}^{(1)} \right|.$$

The term D_1 is strictly positive and does not converge. Therefore $\mathbb{P}\{D_1 \leq o_p(1)\} \to 0$ as $N \to \infty$, which implies that the confounder will be included in the PS model.

Any other the GPS model specification (e.g. $p_k = p(w \mid X^{(2)}, X^{(3)})$) can be represented as a combination of the four types of misspecification. In such cases, the result can be

proven using similar techniques.

A.4 Definition of $c(w)$ & Variance Estimation

For any generalized linear model of the optimal GPS, suppose $\boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}$ consists of b covariates. We let $p_{\boldsymbol{\beta}_{w'}}(w)$ be such that

$$\frac{\partial p}{\partial \boldsymbol{\beta}_{w'}}(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}} = \boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x} \times p_{\boldsymbol{\beta}_{w'}}(w|\boldsymbol{x}; \boldsymbol{\beta}),$$

and define

$$\begin{aligned}\boldsymbol{c}\,(w) =& E\left[\operatorname{cov}\left\{\boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}, \mu\left(w, \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}\right) \mid p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})\right\}_{b\times 1}\right. \\ & \left. \otimes \begin{Bmatrix} p_{\boldsymbol{\beta}_2}(w|\boldsymbol{x}; \boldsymbol{\beta})/p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) \\ \vdots \\ p_{\boldsymbol{\beta}_w}(w|\boldsymbol{x}; \boldsymbol{\beta})/p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) \\ \vdots \\ p_{\boldsymbol{\beta}_T}(w|\boldsymbol{x}; \boldsymbol{\beta})/p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) \end{Bmatrix}_{(T-1)\times 1}\right].\end{aligned}$$

For example, if the optimal GPS follows a multinomial logit model, then $\boldsymbol{c}\,(w)$ takes the form:

$$\begin{aligned}&\boldsymbol{c}\,(w) \\ =& E\left[\operatorname{cov}\left\{\boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}, \mu\left(w, \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}\right) \mid p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta})\right\}_{b\times 1} \otimes \begin{Bmatrix} -p(2 \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) \\ \vdots \\ 1 - p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) \\ \vdots \\ -p(T \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}}; \boldsymbol{\beta}) \end{Bmatrix}_{(T-1)\times 1}\right].\end{aligned}$$

Let $p(w \mid \boldsymbol{X}, \boldsymbol{\beta})$ be the GPS specification chosen by the outcome-adjusted balance measure. Let $p(w \mid \boldsymbol{X}, \hat{\boldsymbol{\beta}})$ denote the corresponding estimated GPS. We break down the variance estimation into two parts.

First, we consider estimation of the asymptotic variance $\sigma^2(w)$ corresponding to matching on the true generalized propensity scores $p(w \mid \boldsymbol{X}, \boldsymbol{\beta})$. An estimator of $\sigma^2(w)$ is

given by

$$\begin{aligned}\hat{\sigma}^2(w) =& \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{Y}_i(w) - \hat{\mu}\,(w)\right\}^2 \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left[D_i(w)\left\{1+K(i,w)\right\}K(i,w)\right]\hat{\bar{\sigma}}^2\left\{W_i, p(w \mid \boldsymbol{X}, \hat{\boldsymbol{\beta}})\right\},\end{aligned}$$

where

$$\hat{\bar{\sigma}}^2\left\{W_i, p(w \mid \boldsymbol{X}, \hat{\boldsymbol{\beta}})\right\} = \frac{L}{L+1}\left\{Y_i - \frac{1}{L}\sum_{j\in N(i,w)} Y_j\right\}^2.$$

Here, $N(i,w)$ are $L \geq 1$ units whose estimated generalized propensity scores are closest to unit i with the same level of treatment w. Typically, L is set to equal 1.

Next, we estimate the adjustment term $\boldsymbol{c}\,(w)^{\mathrm{T}}\, I_{\boldsymbol{\beta}}^{-1}\boldsymbol{c}\,(w)$. The vector $\boldsymbol{c}(w)$ can be estimated by

$$\hat{\boldsymbol{c}}(w) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathrm{cov}}\left\{\boldsymbol{X}_i, \mu(w,\boldsymbol{X}_i) \mid p(w \mid \boldsymbol{X}_i, \boldsymbol{\beta})\right\} \otimes \begin{Bmatrix} p_{\boldsymbol{\beta}_2}(w|\boldsymbol{x};\hat{\boldsymbol{\beta}})/p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}};\hat{\boldsymbol{\beta}}) \\ \vdots \\ p_{\boldsymbol{\beta}_w}(w|\boldsymbol{x};\hat{\boldsymbol{\beta}})/p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}};\hat{\boldsymbol{\beta}}) \\ \vdots \\ p_{\boldsymbol{\beta}_T}(w|\boldsymbol{x};\hat{\boldsymbol{\beta}})/p(w \mid \boldsymbol{X}^{\mathcal{C}\cup\mathcal{P}};\hat{\boldsymbol{\beta}}) \end{Bmatrix},$$

where each conditional covariance vector can be estimated by

$$\begin{aligned}&\widehat{\mathrm{cov}}\left\{\boldsymbol{X}_i, \mu(w,\boldsymbol{X}_i) \mid p(w \mid \boldsymbol{X}_i, \boldsymbol{\beta})\right\} \\ =& \frac{1}{L-1}\sum_{j\in N(i,w)}\left\{\boldsymbol{X}_j - \frac{1}{L}\sum_{k\in N(i,w)}\boldsymbol{X}_k\right\}\left\{Y_j - \frac{1}{L}\sum_{k\in N(i,w)} Y_k\right\}.\end{aligned}$$

Here, $N(i,w)$ has the same definition as the above, but L is typically set to equal 2 or larger. The inverse of information matrix can be estimated by the variance-covariance matrix of the fitted GPS model parameters $\widehat{Var}(\hat{\boldsymbol{\beta}})$.

A.5 Steps to Perform Model Selection

Step 0. Standardize every covariate to have mean zero and variance one. If $\boldsymbol{X}$ is high-dimensional, we recommend applying regularized regression methods such as the Lasso (Tibshirani 1996) to screen out irrelevant covariates.

Step 1. Posit K separate parametric models for $p(w \mid \boldsymbol{X})$, denoted by $\{p(w \mid \boldsymbol{X}; \boldsymbol{\beta}_{w_k}) : k = 1, ..., K\}$ with fixed unknown parameter vectors $\boldsymbol{\beta}_{w_1}, ..., \boldsymbol{\beta}_{w_K}$. Obtain estimates $\widehat{\boldsymbol{\beta}}_{w_1}, ..., \widehat{\boldsymbol{\beta}}_{w_K}$ based on $[\{D_i(w), \boldsymbol{X}_i\} : i = 1, \ldots, N]$. For each unit i, estimate the generalized propensity scores by calculating $\{p(w \mid \boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_{w_k}) : i = 1, \ldots, N; k = 1, ..., K\}$.

Step 2. Choose a metric of correlation ρ following the discussion above. At each treatment level w, compute the outcome-adjusted balance measure for all K posited models:

$$\left[\texttt{OABM}_\rho \left\{w, p(w|\boldsymbol{X}; \widehat{\boldsymbol{\beta}}_{w_k})\right\} : k = 1, ..., K\right].$$

Select the GPS model that minimizes the outcome-adjusted balance measure for each w.

Step 3. At each treatment level w, obtain an estimate of $E\{Y(w)\}$ by matching on the estimated GPS based on the selected model from Step 2. Compute estimates of the pairwise ATEs $\tau(w, w')$ using the GPSM estimator defined in Equation (2.2).

A.6 Existing Balance Measures

- Absolute mean difference:

$$\texttt{AMD}\{w, p(w \mid \boldsymbol{X})\} = \sum_{j=1}^{d} \left| \widehat{X}^{(j)}_{\text{gpsm}}(w) - \overline{X}^{(j)} \right|.$$

- Kolmogorov-Smirnov distance:

$$\texttt{KSdist}\{w, p(w \mid \boldsymbol{X})\} = \sum_{j=1}^{d} \sup_{x^{(j)}} \left| \widehat{F}_N(w, x^{(j)}) - F_N(x^{(j)}) \right|,$$

where $\widehat{F}_N(w, x) = N^{-1} \sum_{i=1}^{N} I_{[-\infty, x]} \left\{ \widehat{X}_i(w) \right\}$ and $F_N(x^{(j)}) = N^{-1} \sum_{i=1}^{N} I_{[-\infty, x]} (X_i)$ are

the empirical CDFs of the imputed covariate and original sample covariate respectively.

- Mahalanobis distance is a better measure of overall balance when covariates are correlated:

$$\texttt{Mdist}\left\{w, p(w \mid \boldsymbol{X})\right\} = \left[\widehat{\boldsymbol{X}}_{\text{gpsm}}(w) - \overline{\boldsymbol{X}}\right]^{\text{T}} S^{-1} \left[\widehat{\boldsymbol{X}}_{\text{gpsm}}(w) - \overline{\boldsymbol{X}}\right],$$

where $\widehat{\boldsymbol{X}}_{\text{gpsm}}(w) = \left[\widehat{X}^{(1)}_{\text{gpsm}}(w), ..., \widehat{X}^{(d)}_{\text{gpsm}}(w)\right]^{\text{T}}$, $\overline{\boldsymbol{X}} = \left(\overline{X}^{(1)}, ..., \overline{X}^{(d)}\right)^{\text{T}}$, and S denotes the sample covariance matrix of $\left(X^{(1)}, ..., X^{(d)}\right)^{\text{T}}$,

- Weighted balance measure:

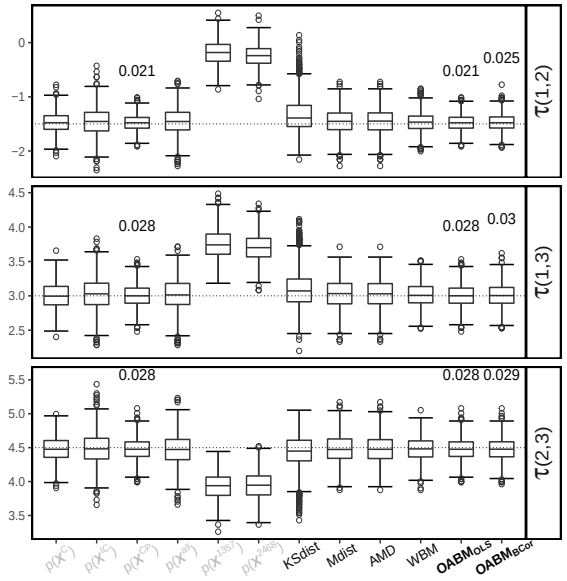
$$\texttt{WBM}\left\{w, p(w \mid \boldsymbol{X})\right\} = \sum_{j=1}^{d} \left|\widehat{\theta}^{(j)} \left[\widehat{X}^{(j)}_{\text{gpsm}}(w) - \overline{X}^{(j)}\right]\right|,$$

where $\widehat{\theta}^{(j)}$ is the coefficient of $X^{(j)}$ from a multivariate maximum likelihood regression of the observed outcome Y on $\boldsymbol{X}$.

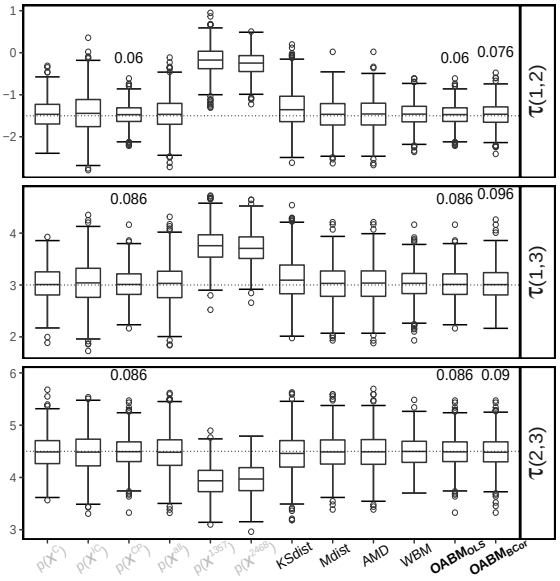
A.7 Remaining Scenarios of the Simulation Study

For the simulation study, we compare the performance of balance measures under the combinations of (strong/weak) instrumental variables, (strong/weak) precision variables, and potential outcome model that is a (linear/nonlinear) function of the covariates. In this section, we focus on the remaining scenarios where IV's are weak predictors of the treatment assignment.

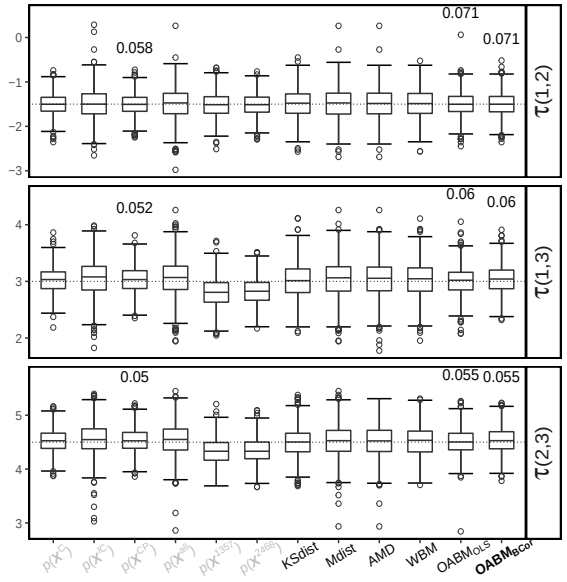
In Figure A.1, we summarize the estimation performance of the GPSM estimator with GPS selected based on different balance measures. Figure A.2 shows the percentage of the times (out of 1000 datasets) each benchmark model is selected by the balance measures. Table A.1 presents the coverage rates of the asymptotic 95% confidence intervals.



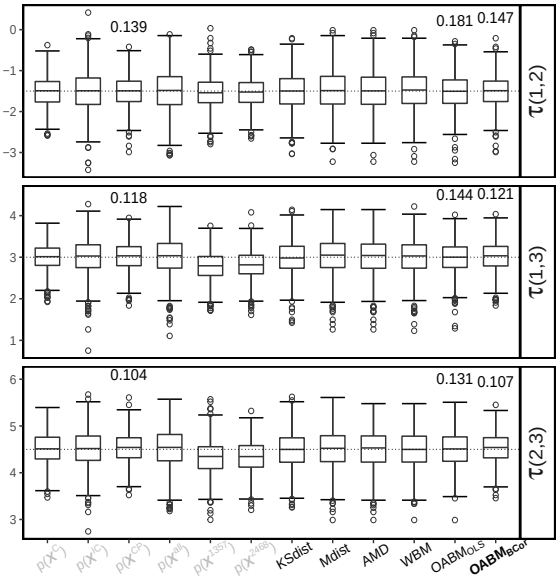
(a) $u = 1, v = 1$
Outcome *linear* in covariates



(b) $u = 1, v = 2$
Outcome *linear* in covariates



(c) $u = 1, v = 1$
Outcome *nonlinear* in covariates

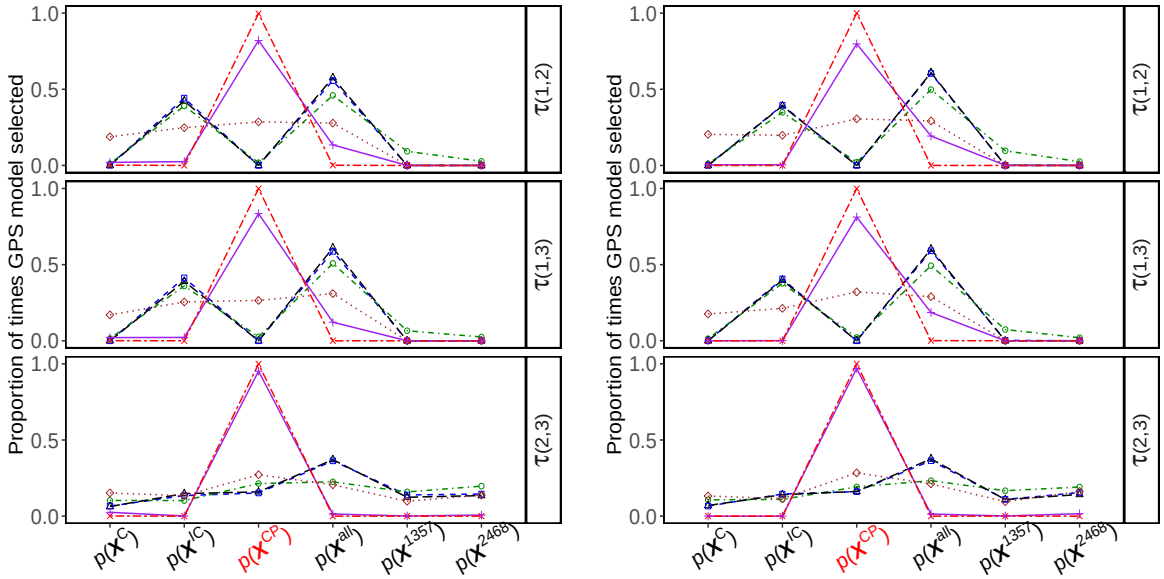


(d) $u = 1, v = 2$
Outcome *nonlinear* in covariates

Figure A.1: Box plots of 1000 generalized propensity score matching estimates under scenarios with weak IVs. The 6 benchmark GPS models are greyed out on the x-axis. Numeric MSEs for $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$, $\mathtt{OABM}_{\mathtt{OLS}}$, and $\mathtt{OABM}_{\mathtt{BCor}}$ are explicitly shown above their corresponding box plots. True pairwise treatment effects are denoted by the horizontal dotted lines.

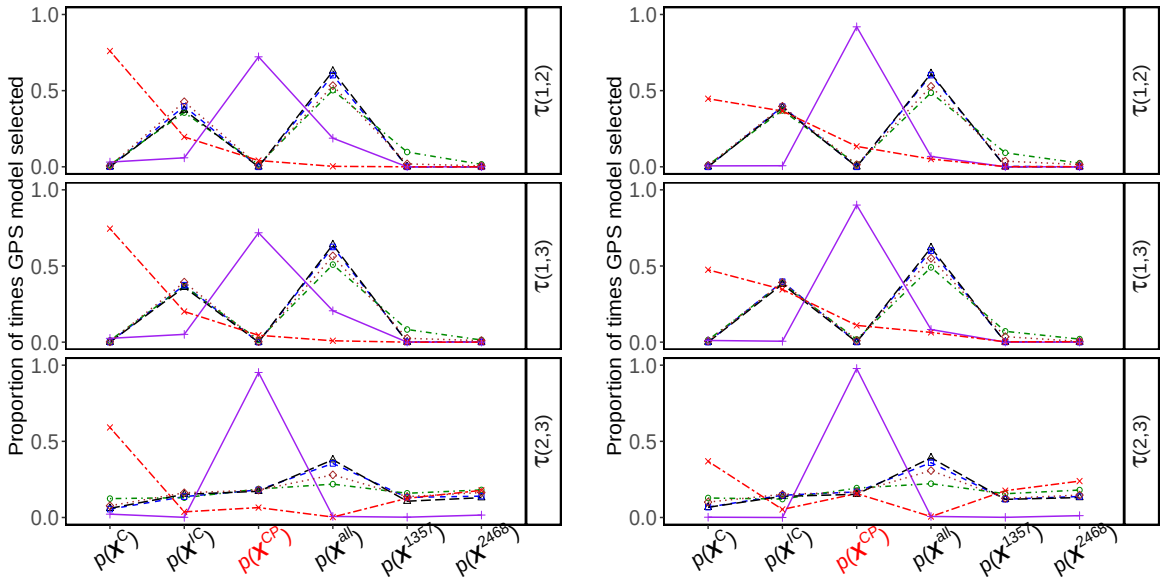
As indicated by Figure A.1a and A.1b, when the covariates are linear predictors of the potential outcomes, among all six balance measures, estimation with $\texttt{OABM}_{\texttt{OLS}}$ has mean squared errors almost matching that of the optimal benchmark model $\texttt{p}(\texttt{X}^{\texttt{CP}})$. This is accompanied by the fact that $\texttt{OABM}_{\texttt{OLS}}$ overwhelmingly selects the optimal benchmark model, as demonstrated by the purple line the Figure A.2a and A.2b. Since the penalizing weights of $\texttt{OABM}_{\texttt{OLS}}$ are computed from a correctly specified outcome model, this result is expected and is consistent with Theorem 1. The coverage rates for all balance measures are within acceptable range except for `KSdist`, which fall short of the nominal 95% rate by a sizable amount. Comparing to the strong IV scenarios from the main article, the advantage of using $\texttt{OABM}_{\texttt{OLS}}$ is slightly smaller when IVs are weaker, in which case including these IVs in the GPS will result in less efficiency loss.

When the potential outcome has a nonlinear relationship with the covariates, $\texttt{OABM}_{\texttt{BCor}}$ results in slightly superior estimation performance in terms of mean squared error, as illustrated in Figure A.2c and A.2d. Combining the model selection results shown in Figure 2.3 and A.2, we notice that $\texttt{OABM}_{\texttt{BCor}}$ is slightly less consistent in selecting the optimal model than $\texttt{OABM}_{\texttt{OLS}}$ in the linear cases, possibly due to the existence of collider bias. Regarding the coverage rates, $\texttt{OABM}_{\texttt{BCor}}$ also yields slightly greater coverage than non-penalizing based balance measures, coming close to the nominal 95%.



(a) $u = 1, v = 1$
Outcome *linear* in covariates

(b) $u = 1, v = 2$
Outcome *linear* in covariates



(c) $u = 1, v = 1$
Outcome *nonlinear* in covariates

(d) $u = 1, v = 2$
Outcome *nonlinear* in covariates

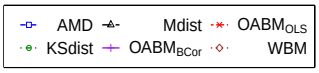


Figure A.2: Proportion of model selection over 1000 simulations under scenarios with weak IVs. The optimal benchmark model $\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$ is colored in red.

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.805	0.843	0.885
Mdist	0.959	0.950	0.957
AMD	0.956	0.951	0.956
WBM	0.980	0.977	0.978
$\texttt{OABM}_{\texttt{OLS}}$	0.968	0.950	0.942
$\texttt{OABM}_{\texttt{BCor}}$	0.970	0.955	0.941

(a) $u = 1, v = 1$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.861	0.857	0.910
Mdist	0.963	0.931	0.945
AMD	0.960	0.926	0.936
WBM	0.985	0.952	0.955
$\texttt{OABM}_{\texttt{OLS}}$	0.974	0.920	0.913
$\texttt{OABM}_{\texttt{BCor}}$	0.968	0.916	0.923

(b) $u = 1, v = 2$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.928	0.919	0.941
Mdist	0.925	0.909	0.922
AMD	0.923	0.903	0.925
WBM	0.925	0.904	0.933
$\texttt{OABM}_{\texttt{OLS}}$	0.942	0.941	0.948
$\texttt{OABM}_{\texttt{BCor}}$	0.936	0.934	0.939

(c) $u = 1, v = 1$
Outcome *nonlinear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
KSdist	0.945	0.942	0.941
Mdist	0.944	0.929	0.927
AMD	0.946	0.933	0.933
WBM	0.940	0.937	0.937
$\texttt{OABM}_{\texttt{OLS}}$	0.938	0.951	0.951
$\texttt{OABM}_{\texttt{BCor}}$	0.941	0.947	0.956

(d) $u = 1, v = 2$
Outcome *nonlinear* in covariates

Table A.1: Coverage rates of asymptotic 95% confidence intervals under scenarios with weak IVs

A.8 Nursing Home Data Codebook

- Reimbursement—inflation-adjusted reimbursement (US $2000)
- ICU—any ICU during hospitalization
- NH Stay—nursing home stay within 120 days before inpatient admission
- SNF Admission—SNF admission (yes/no)
- Gender—beneficiary sex (1=male,2=female)
- Age—beneficiary age

- Infectious—primary ICD-9 diagnosis 001-139: infectious and parasitic diseases
- Neoplasms—primary ICD-9 diagnosis 140-239: neoplasms
- Blood—primary ICD-9 diagnosis 280-289: diseases of the blood and blood-forming organs
- Mental—primary ICD-9 diagnosis 290-319: mental disorders
- Nervous—primary ICD-9 diagnosis 320-389: diseases of the nervous system and sense organs
- Circulatory—primary ICD-9 diagnosis 390-459: diseases of the circulatory system
- Respiratory—primary ICD-9 diagnosis 460-519: diseases of the respiratory system
- Digestive—primary ICD-9 diagnosis 520-579: diseases of the digestive system
- Genitourinary—primary ICD-9 diagnosis 580-629: diseases of the genitourinary system
- Skin—primary ICD-9 diagnosis 680-709: diseases of the skin and subcutaneous tissue
- Muscle—primary ICD-9 diagnosis 710-739: diseases of the musculoskeletal system and connective tissue
- Ill Defined—primary ICD-9 diagnosis 780-799: symptoms,signs,and ill-defined conditions
- Poisoning—primary ICD-9 diagnosis 800-999: injury and poisoning
- External—primary ICD-9 diagnosis E: external causes of injury
- Supplementary—primary ICD-9 diagnosis V: supplemental classification
- Admission 1999—nursing home year of admission in 1999
- Admission 2000—nursing home year of admission in 2000
- Admission 2001—nursing home year of admission in 2001
- Admission 2002—nursing home year of admission in 2002
- Admission 2003—nursing home year of admission in 2003

- Admission 2004—nursing home year of admission in 2004
- Admission 2005—nursing home year of admission in 2005
- Admission 2006—nursing home year of admission in 2006
- Admission 2007—nursing home year of admission in 2007

A.9 Application: Tutoring Data Analysis

We apply the model selection method to the Tutoring dataset from the *TriMatch* R package (Bryer 2017), which contains results from a study examining the effects of tutoring services on course grades. A total of 1142 observations consist of 918 students who did not receive any form of tutoring (Control), 134 students who received the first form of tutoring (Treat1), and 90 students who received the second form of tutoring (Treat2). The course grade the student earn is the outcome variable and takes on one of five numeric values: 4=A, 3=B, 2=C, 1=D, 0=F or W. Information on gender, ethnicity, military service status of the student, non-native English speaker status, education level of the student's mother, education level of the student's father, age of the student, employment status, household income, number of transfer credits, and grade point average are collected as baseline covariates. The objective in this analysis is to further showcase the efficiency gain by using the outcome-adjusted balance measure. The fitted generalized propensity scores (probabilities of a student receiving each of the three tutoring services) will serve as matching variables for the GPSM estimator to estimate and make inference about the pairwise average treatment effects.

Following the same analysis procedure as for the nursing home data, we first apply the group lasso (Yuan and Lin 2006) to select strong predictors for tutoring service assignment. We also apply the lasso to select strong predictors for $Y(\text{Control})$, $Y(\text{Treat1})$, and $Y(\text{Treat2})$. For estimation of each average potential outcome, we again categorize the relevant covariates into instrumental variables, confounders, and precision variables based on the lasso variable selection results. For each treatment level w, we estimate the average potential outcome $E\{Y(w)\}$ by positing the following three candidate GPS models:

- Model 1: multinomial logit with all 12 covariates entered linearly
- Model 2: multinomial logit with all confounders and precision variables entered linearly

- Model 3: model 2 with the addition of first-order interaction terms of outcome-related variables

We consider using $\texttt{OABM}_{\texttt{BCor}}$ for GPS model selection, as it represents our lack of confidence in describing the true relationship between the full set of covariates and the course grades conditional on the choice of tutoring service. The model selection results are shown in Table A.2. Model 2 is consistently selected for estimating all three potential outcome means. The selected model specifications are subsequently used to fit the data and compute the estimated generalized propensity scores, which serve as matching variables for the GPSM estimator to produce point estimates of the pairwise ATEs and standard errors. To show the benefit of our proposed method, we compare the result to one based on the GPS specification that involves only confounders and IVs. Table A.3 shows a notable reduction in standard error by using $\texttt{OABM}_{\texttt{BCor}}$ to select the GPS model compared to naively modeling the GPS using strong treatment predictors. Based on the result of the GPSM estimates, we conclude that there is sufficient statistical evidence that students who undergo either form of tutoring will lead to better course grades than if they receive no tutoring at all. The second form of tutoring service also improves course grades compared to the first form of tutoring.

Model	$E\{Y(1)\}$	$E\{Y(2)\}$	$E\{Y(3)\}$
1	75	83	39
2	**36**	**5**	**25**
3	45	8	41

Table A.2: The proposed $\texttt{OABM}_{\texttt{BCor}}$ evaluated at each treatment level, for each posited model; bold number indicates the lowest $\texttt{OABM}_{\texttt{BCor}}$ value for that treatment; tutoring data

Method	$\hat{\tau}(1,2)$	$\hat{\tau}(1,3)$	$\hat{\tau}(2,3)$
GPSM-$\texttt{OABM}_{\texttt{BCor}}$	0.319 (**0.089**)*	0.559 (**0.039**)*	0.240 (**0.082**)*
GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$	0.384 (0.120)*	0.633 (0.126)*	0.249 (0.154)

Table A.3: GPSM estimates of the pairwise ATEs among three types of tutoring services; GPSM-$\texttt{OABM}_{\texttt{BCor}}$ uses $\texttt{OABM}_{\texttt{BCor}}$ to select GPS model among the three posited models; GPSM-$p(\boldsymbol{X}^{\mathcal{I}\cup\mathcal{C}})$ uses a multinomial logit model for the GPS specification that involves linear terms of IVs and confounders; standard errors are displayed in parenthesis; * indicates significance at the 0.05 level; tutoring data

A.10 Simulation Result Details

We display in the tables below the MSE of GPSM estimator based on GPS models selected using different balance measures from the simulation study.

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\texttt{p}(\texttt{X}^{\texttt{CP}})$	0.021	0.028	0.028
KSdist	0.157	0.105	0.072
Mdist	0.054	0.047	0.042
AMD	0.055	0.046	0.044
WBM	0.033	0.031	0.029
$\texttt{OABM}_{\texttt{OLS}}$	0.021	0.028	0.028
$\texttt{OABM}_{\texttt{BCor}}$	0.025	0.030	0.029

(a) $u = 1, v = 1$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\texttt{p}(\texttt{X}^{\texttt{CP}})$	0.060	0.086	0.086
KSdist	0.241	0.190	0.140
Mdist	0.142	0.132	0.113
AMD	0.148	0.136	0.118
WBM	0.079	0.090	0.083
$\texttt{OABM}_{\texttt{OLS}}$	0.060	0.086	0.086
$\texttt{OABM}_{\texttt{BCor}}$	0.076	0.096	0.090

(b) $u = 1, v = 2$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\texttt{p}(\texttt{X}^{\texttt{CP}})$	0.016	0.027	0.024
KSdist	0.362	0.178	0.139
Mdist	0.182	0.119	0.092
AMD	0.186	0.120	0.094
WBM	0.051	0.043	0.040
$\texttt{OABM}_{\texttt{OLS}}$	0.016	0.027	0.024
$\texttt{OABM}_{\texttt{BCor}}$	0.041	0.038	0.036

(c) $u = 2, v = 1$
Outcome *linear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\texttt{p}(\texttt{X}^{\texttt{CP}})$	0.040	0.069	0.070
KSdist	0.427	0.251	0.201
Mdist	0.388	0.286	0.220
AMD	0.382	0.291	0.229
WBM	0.083	0.088	0.086
$\texttt{OABM}_{\texttt{OLS}}$	0.040	0.069	0.070
$\texttt{OABM}_{\texttt{BCor}}$	0.085	0.094	0.090

(d) $u = 2, v = 2$
Outcome *linear* in covariates

Table A.4: Simulation results of the MSE of GPSM estimator using competing balance measures for GPS model selection when the outcome model is nonlinear in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$	0.058	0.052	0.050
KSdist	0.109	0.089	0.072
Mdist	0.129	0.102	0.087
AMD	0.126	0.102	0.088
WBM	0.116	0.099	0.080
$\mathtt{OABM}_{\mathtt{OLS}}$	0.071	0.060	0.055
$\mathtt{OABM}_{\mathtt{BCor}}$	0.071	0.060	0.055

(a) $u = 1, v = 1$
Outcome *nonlinear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$	0.139	0.118	0.104
KSdist	0.206	0.160	0.148
Mdist	0.238	0.187	0.167
AMD	0.239	0.187	0.166
WBM	0.233	0.174	0.161
$\mathtt{OABM}_{\mathtt{OLS}}$	0.181	0.144	0.131
$\mathtt{OABM}_{\mathtt{BCor}}$	0.147	0.121	0.107

(b) $u = 1, v = 2$
Outcome *nonlinear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$	0.046	0.045	0.042
KSdist	0.118	0.080	0.077
Mdist	0.248	0.158	0.149
AMD	0.242	0.157	0.142
WBM	0.244	0.151	0.147
$\mathtt{OABM}_{\mathtt{OLS}}$	0.090	0.065	0.059
$\mathtt{OABM}_{\mathtt{BCor}}$	0.076	0.064	0.055

(c) $u = 2, v = 1$
Outcome *nonlinear* in covariates

	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
$\mathtt{p}(\mathtt{X}^{\mathtt{CP}})$	0.112	0.103	0.100
KSdist	0.238	0.165	0.183
Mdist	0.555	0.326	0.341
AMD	0.525	0.325	0.321
WBM	0.534	0.311	0.322
$\mathtt{OABM}_{\mathtt{OLS}}$	0.265	0.188	0.166
$\mathtt{OABM}_{\mathtt{BCor}}$	0.128	0.109	0.111

(d) $u = 2, v = 2$
Outcome *nonlinear* in covariates

Table A.5: Simulation results of the MSE of GPSM estimator using competing balance measures for GPS model selection when the outcome model is nonlinear in covariates

APPENDIX

B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

We first summarize all relevant conditions and then outline the proofs for Theorems 3 and 4. The main structure of the proofs follows directly from Yang and Zhang (2023) and Yang et al. (2016).

We list all the relevant assumptions:

Assumption 14 *We have a random sample of size n from a large population.*

Assumption 15 *For all values of x the probability of receiving any level of the treatment is positive: $p(w \mid x) > 0 \quad$ for all w, x.*

Assumption 16 *For all $w \in \mathbb{W}$, we have $D_i(w) \perp\!\!\!\perp Y_i(w) \mid X_i$.*

Assumption 17 *The parametric model $p(w \mid X; \alpha)$ is a correct specification for $p(w \mid X)$; i.e., $p(w \mid X) = p(w \mid X; \alpha_0)$, where α_0 is the true model parameter.*

Assumption 18 *Suppose that $Y(w)$ follows a location-shift family and that the parametric model $\mu(w \mid X; \beta_w)$ is a correct specification for $\mu(w \mid X)$; i.e., $\mu(w \mid X) = \mu(w \mid X; \beta_{w,0})$ where $\beta_{w,0}$ is the true model parameter.*

Assumption 19 *For $w = 1, ..., T$, (i) The matching variable S_w has a compact and convex support $\mathcal{S}$, with a continuous density bounded and bounded away from zero: there exist constants C_L and C_U such that $C_L \leq f_{w'}(S_w)/f_w(S_w) \leq C_U$ for any $S_w \in \mathcal{S}$; (ii) $\mu(w|S_w)$ and $\sigma^2(w|S_w)$ satisfy the Lipschitz continuity condition; and (iii) there exists $\delta > 0$ such that $\mathbb{E}\{|Y(w)|^{2+\delta}|S\}$ is uniformly bounded for any $S_w \in \mathcal{S}$.*

Assumption 20 *(i) Under $\mathbb{P}^{\theta^*}$, $\mathcal{U}_n(\theta^*) \to \mathcal{N}(0, \Sigma_U)$ in distribution, as $n \to \infty$ where $\Sigma_U = \mathbb{E}\left\{U(W, X, Y; \theta^*) U(W, X, Y; \theta^*)^{\mathrm{T}}\right\}$; (ii) $\Gamma_\theta = \mathbb{E}\left\{\partial U(W, X, Y; \theta)/\partial\theta^{\mathrm{T}}\right\}$ is non-singular around θ^*; and (iii) for any vector of constant $h, \exp\left\{n^{1/2} h^{\mathrm{T}} \Gamma_{\theta^*} \Sigma_U^{-1} \mathcal{U}_n(\theta^*)\right\}$ is uniformly integrable.*

Assumption 21 *There exists a neighborhood of θ^*, such that for any θ in this region, the following conditions hold: for $w = 1, ..., T$, (i) the matching variable $S_w(\theta)$ has a compact and convex support $\mathcal{S}$, with a continuous density bounded and bounded away from zero; (ii) $\mu\{w|S_w(\theta)\}$ and $\sigma^2\{w|S_w(\theta)\}$ satisfy the Lipschitz continuity condition; and (iii) there exists $\delta > 0$ such that $\mathbb{E}\left\{|Y(w)|^{2+\delta} \mid S_w(\theta)\right\}$ is uniformly bounded for any $S_w(\theta) \in \mathcal{S}$.*

Assumption 22 *For all bounded continuous functions $h(W, X, Y)$, the conditional expectation $\mathbb{E}_{\theta_n}\{h(W, X, Y)\}$ converges in distribution to $\mathbb{E}\{h(W, X, Y)\}$, where $\mathbb{E}_{\theta_n}(\cdot)$ is the expectation taken with respect to P^{θ_n}.*

B.1 Proof of Theorem 3

Following Yang and Zhang (2023), Abadie and Imbens (2011) and Abadie and Imbens (2012), under mild regularity conditions on the nonparametric estimation, we have $\hat{B}_n = B_n + o_P(1)$. Then, $\hat{\tau}_{\mathrm{dsm}}$ has the following asymptotic linear form:

$$
\begin{aligned}
& n^{1/2}\left\{\hat{\tau}_{\mathrm{dsm}}(w, w'; \theta^*) - \tau(w, w')\right\} && \text{(B.1)}\\
=& n^{-1/2} \sum_{i=1}^{n} \left\{\mu(w', S_{w',i}) - \mu(w, S_{w,i}) - \tau\right\} && \text{(B.2)}\\
& + n^{-1/2} \sum_{i=1}^{n} D_i(w') \left(1 + M^{-1} K_{S_{w',i}}\right) \left\{Y_i - \mu(w', S_{w',i})\right\} && \text{(B.3)}\\
& - n^{-1/2} \sum_{i=1}^{n} D_i(w) \left(1 + M^{-1} K_{S_{w,i}}\right) \left\{Y_i - \mu(w, S_{w,i})\right\} + o_P(1). && \text{(B.4)}
\end{aligned}
$$

If either the generalized propensity score or generalized prognostic score is correctly specified, by Lemma 6, we have $\mathbb{E}\{\mu(w', S_{w',i}) - \mu(w, S_{w,i})\} = \tau(w, w')$ and therefore $n^{1/2}\{\hat{\tau}_{\text{dsm}}(w, w'; \theta^*) - \tau(w, w')\}$ converges to zero. Ignoring the term $o_P(1)$, let the remaining terms be

$$
\begin{aligned}
T_{1n} &= n^{-1/2} \sum_{i=1}^{n} \{\mu(w', S_{w',i}) - \mu(w, S_{w,i}) - \tau\} \\
T_{2n} &= n^{-1/2} \sum_{i=1}^{n} D_i(w') \left(1 + M^{-1} K_{S_{w',i}}\right) \{Y_i - \mu(w', S_{w',i})\}, \\
T_{3n} &= -n^{-1/2} \sum_{i=1}^{n} D_i(w) \left(1 + M^{-1} K_{S_{w,i}}\right) \{Y_i - \mu(w, S_{w,i})\}.
\end{aligned}
$$

We first show their covariances are zero:

$$
\begin{aligned}
&\operatorname{cov}(T_{1n}, T_{2n}) \\
=&n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}[\mu(w', S_{w',i}) - \mu(w, S_{w,i}) - \tau, \\
&D_j(w') \left(1 + M^{-1} K_{S_{w',j}}\right) \{Y_j - \mu(w', S_{w',j})\}] \\
=&n^{-1} \sum_{i=1}^{n} \operatorname{cov}\left[\mu(w', S_{w',i}) - \mu(w, S_{w,i}) - \tau, D_i(w') \left(1 + M^{-1} K_{S_{w',i}}\right) \{Y_i - \mu(w', S_{w',i})\}\right] \\
=&n^{-1} \sum_{i=1}^{n} \operatorname{cov}(\mathbb{E}\{\mu(w', S_{w',i}) - \mu(w, S_{w,i}) - \tau \mid S_{w',i}, S_{w,i}\}, \\
&\mathbb{E}\left[D_i(w') \left(1 + M^{-1} K_{S_{w',i}}\right) \{Y_i - \mu(w', S_{w',i})\} \mid S_{w',i}, S_{w,i}\right]) \\
&+ n^{-1} \sum_{i=1}^{n} \mathbb{E}[\operatorname{cov}\{\mu(w', S_{w',i}) - \mu(w, S_{w,i}) - \tau, \\
&D_i(w') \left(1 + M^{-1} K_{S_{w',i}}\right) \{Y_i - \mu(w', S_{w',i})\} \mid S_{w',i}, S_{w,i}\}] \\
=&0.
\end{aligned}
$$

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

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

The first term becomes $\mathbb{E}\left[\{\mu(w'|S_{w'})-\mu(w|S_w)-\tau(w,w')\}^2\right]$. Following Abadie and Imbens (2006) and Yang et al. (2016), the second and third term, as $n\to\infty$, becomes

$$
\begin{aligned}
&\text{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}D_i(w')\left(1+M^{-1}K_{S_{w',i}}\right)^2\mathbb{V}\left(Y_i\mid S_{w',i}\right)\right]\\
&+\text{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}D_i(w)\left(1+M^{-1}K_{S_{w},i}\right)^2\mathbb{V}\left(Y_i\mid S_{w,i}\right)\right]\\
=&\mathbb{E}\left(\sigma^2(w'|S_{w'})\left[\frac{1}{p(w'\mid X;\alpha^*)}+\frac{1}{2M}\left\{\frac{1}{p(w'\mid X;\alpha^*)}-p(w'\mid X;\alpha^*)\right\}\right]\right)\\
&+\mathbb{E}\left(\sigma^2(w|S_{w})\left[\frac{1}{p(w\mid X;\alpha^*)}+\frac{1}{2M}\left\{\frac{1}{p(w\mid X;\alpha^*)}-p(w\mid X;\alpha^*)\right\}\right]\right).
\end{aligned}
$$

B.2 Le Cam's Third Lemma

Consider two sequences of probability measures $\left(\mathbb{Q}^{(n)}\right)_{n=1}^{\infty}$ and $\left(\mathbb{P}^{(n)}\right)_{n=1}^{\infty}$. 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

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

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

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

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

B.3 Proof of Theorem 4

To derive the large sample distribution of $\hat{\tau}_{\text{dsm}}(w, w'; \hat{\theta})$, we require some additional assumptions (e.g. Assumptions 20-22). Following arguments by Yang and Zhang (2023), Andreou and Werker (2012) and Abadie and Imbens (2016), we let $\mathbb{P}$ be the distribution of $\{(W_i, X_i, Y_i) : i = 1, \ldots, n\}$. Consider $\mathbb{P} = \mathbb{P}^*$ to be indexed by $\theta^* = (\alpha^{*\mathrm{T}}, \beta_w^{*\mathrm{T}}, \beta_{w'}^{*\mathrm{T}})^{\mathrm{T}}$, which satisfies

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

Under Assumption 20, we invoke results for Z-estimation van der Vaart (2000) and get:

$$n^{1/2}\left(\hat{\theta} - \theta^*\right) = -\Gamma_{\theta^*}^{-1}\mathcal{U}_n\left(\theta^*\right) + o_P(1) \to \mathcal{N}\left(0, \Sigma_{\theta^*}\right),$$

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

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 $\theta^*, \theta_n = \theta^* + n^{-1/2}h$, with a density

$$\frac{\exp\left\{n^{1/2}\left(\theta_n - \theta^*\right)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n\left(\theta^*\right) - 2^{-1}n\left(\theta_n - \theta^*\right)^{\mathrm{T}}\Sigma_{\theta^*}^{-1}\left(\theta_n - \theta^*\right)\right\}}{\mathbb{E}\left[\exp\left\{n^{1/2}\left(\theta_n - \theta^*\right)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n\left(\theta^*\right) - 2^{-1}n\left(\theta_n - \theta^*\right)^{\mathrm{T}}\Sigma_{\theta^*}^{-1}\left(\theta_n - \theta^*\right)\right\}\right]}. \tag{B.5}$$

By Assumption 21 (iii), $\exp\left\{n^{1/2}\left(\theta_n - \theta^*\right)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n\left(\theta^*\right)\right\}$ is uniformly integrable, and thus model (B.5) is uniformly locally asymptotically normal. Because under $\mathbb{P}^{\theta^*}$, $\mathcal{U}_n\left(\theta^*\right) \to \mathcal{N}\left(0, \Sigma_U\right)$ in distribution, the normalizing constant in the denominator converges to one as $n \to \infty$. The Fisher information under the parametric model (B.5) is $n\Sigma_{\theta^*}^{-1}$. Therefore, $\hat{\theta}$ is efficient under model (B.5).

Now consider (W_i, X_i, Y_i), for $i = 1, \ldots, n$, with the local shift $\mathbb{P}^{\theta_n}$ (Bickel et al.; 1993). Under model (B.5), the likelihood ratio under $\mathbb{P}^{\theta_n}$ is

$$\begin{aligned} \log\left(\mathrm{d}\mathbb{P}^*/\mathrm{d}\mathbb{P}^{\theta_n}\right) &= -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n\left(\theta^*\right) + \frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h + o_p(1) \\ &= -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n\left(\theta_n\right) - \frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h + o_p(1), \end{aligned} \tag{B.6}$$

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

$$n^{1/2}\left(\hat{\theta} - \theta_n\right) = \Gamma_{\theta^*}^{-1}\mathcal{U}_n(\theta_n) + o_P(1). \tag{B.7}$$

Given the Assumption 22, we derive the results in Theorem 4 in two steps.

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

$$\tau(w, w'; \theta_n) = \mathbb{E}\left[\mu\left\{w', S_{w'}(\theta_n)\right\} - \mu\left\{w, S_w(\theta_n)\right\}\right].$$

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

$$\begin{aligned}
&\begin{pmatrix} n^{1/2}\left\{\hat{\tau}_{\mathrm{dsm}}(w, w'; \theta_n) - \tau(w, w'; \theta_n)\right\} \\ n^{1/2}\left(\hat{\theta} - \theta_n\right) \\ \log\left(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n}\right) \end{pmatrix} \\
&\to \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} & -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} & -h \\ -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\gamma_1 & -h^{\mathrm{T}} & h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h \end{pmatrix}\right\}.
\end{aligned} \tag{B.8}$$

in distribution, as $n \to \infty$. We then express $\tau(w, w'; \theta_n) = \tau(w, w'; \theta^*) + \gamma_2^{\mathrm{T}}\left(n^{-1/2}h\right) + o\left(n^{-1/2}\right)$, where

$$\gamma_2 = \left.\frac{\partial\tau(w, w'; \theta)}{\partial\theta}\right|_{\theta=\theta^*} = \mathbb{E}\left[\left.\frac{\partial\mu\left\{w', S_{w'}(\theta_n)\right\} - \mu\left\{w, S_w(\theta_n)\right\}}{\partial\theta}\right|_{\theta=\theta^*}\right].$$

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

$$\begin{aligned}
&\begin{pmatrix} n^{1/2}\left\{\hat{\tau}_{\mathrm{dsm}}(w, w'; \theta_n) - \tau(w, w')\right\} \\ n^{1/2}\left(\hat{\theta} - \theta_n\right) \end{pmatrix} \\
&\to \mathcal{N}\left\{\begin{pmatrix} -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h - \gamma_2^{\mathrm{T}}h \\ -h \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} \end{pmatrix}\right\}.
\end{aligned}$$

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

$$\begin{aligned} &\left(\begin{array}{c} n^{1/2} \left\{ \hat{\tau}_{\text{dsm}} \left(w, w'; \theta^* + n^{-1/2}h \right) - \tau(w, w') \right\} \\ n^{1/2} \left(\hat{\theta} - \theta^* \right) \end{array} \right) \\ &\to \mathcal{N} \left\{ \left(\begin{array}{c} -\gamma_1^{\mathrm{T}} \Sigma_U^{-1} \Gamma_{\theta^*} h - \gamma_2^{\mathrm{T}} h \\ 0 \end{array} \right), \left(\begin{array}{cc} V_\tau & \gamma_1^{\mathrm{T}} \Gamma_{\theta^*}^{-1} \\ \Gamma_{\theta^*}^{-1} \gamma_1 & \Sigma_{\theta^*} \end{array} \right) \right\}. \end{aligned} \tag{B.9}$$

in distribution, as $n \to \infty$. In the second step, we provide a heuristic derivation for (B.9) to obtain the large sample distribution of $\hat{\tau}_{\text{dsm}}(w, w'; \hat{\theta})$. If the Normal distribution were exact, then

$$\begin{aligned} &n^{1/2} \left\{ \hat{\tau}_{\text{dsm}} \left(w, w'; \theta^* + n^{-1/2}h \right) - \tau(w, w') \right\} \mid n^{1/2} \left(\hat{\theta} - \theta^* \right) = h \\ \sim &\mathcal{N} \left(-\gamma_2^{\mathrm{T}} h, V_\tau - \gamma_1^{\mathrm{T}} \Sigma_U^{-1} \gamma_1 \right). \end{aligned} \tag{B.10}$$

Given that $n^{1/2} \left(\hat{\theta} - \theta^* \right) = h$, we have $\theta^* + n^{-1/2}h = \hat{\theta}$, and hence $\hat{\tau}_{\text{dsm}} \left(w, w'; \theta^* + n^{-1/2}h \right) = \hat{\tau}_{\text{dsm}}(w, w'; \hat{\theta})$. Marginalizing (B.10) over the asymptotic distribution of $n^{1/2} \left(\hat{\theta} - \theta^* \right)$, we derive $V_{\tau,\text{adj}} = V_\tau - \gamma_1^{\mathrm{T}} \Sigma_U^{-1} \gamma_1 + \gamma_2^{\mathrm{T}} \Sigma_{\theta^*} \gamma_2$. The formal technique to derive $V_{\tau,\text{adj}}$ can be found 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 (B.10) in the first step of the proof. Asymptotic normality of $n^{1/2} \left\{ \hat{\tau}_{\text{dsm}} \left(w, w'; \theta_n \right) - \tau \left(w, w'; \theta_n \right) \right\}$ under $\mathbb{P}^{\theta_n}$ follows from Theorem 1 and the uniform local asymptotic normality of model (B.5). Asymptotic joint normality of $\log \left(\mathrm{d}\mathbb{P}^{\theta^*} / \mathrm{d}\mathbb{P}^{\theta_n} \right)$ and $n^{1/2} \left(\hat{\theta} - \theta_n \right)$ follows from (B.6) and (B.7). Also, $n^{1/2} \left\{ \hat{\tau}_{\text{dsm}} \left(w, w'; \theta_n \right) - \tau \left(w, w'; \theta_n \right) \right\} = D_n \left(\theta_n \right) + o_P(1)$, where

$$D_n \left(\theta_n \right) = n^{-1/2} \sum_{i=1}^{n} \left\{ \mu \left(w', S_{w',i} \right) - \mu \left(w, S_{w,i} \right) - \tau \right\} \tag{B.11}$$

$$+ n^{-1/2} \sum_{i=1}^{n} D_i(w') \left(1 + M^{-1} K_{S_{w',i}} \right) \left\{ Y_i - \mu \left(w', S_{w',i} \right) \right\} \tag{B.12}$$

$$- n^{-1/2} \sum_{i=1}^{n} D_i(w) \left(1 + M^{-1} K_{S_{w,i}} \right) \left\{ Y_i - \mu \left(w, S_{w,i} \right) \right\} + o_P(1). \tag{B.13}$$

It remains to show that under $\mathbb{P}^{\theta_n}$,

$$\begin{pmatrix} D_n(\theta_n) \\ \mathcal{U}_n(\theta_n) \end{pmatrix} \to \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}} \\ \gamma_1 & \Sigma_U \end{pmatrix} \right\} \tag{B.14}$$

in distribution as $n \to \infty$. To prove this, we consider the linear combination

$$T_n = c_0 D_n(\theta_n) + c^{\mathrm{T}} \mathcal{U}_n(\theta_n). \tag{B.15}$$

We can apply the martingale central limit theorem to derive the large sample distribution of T_n (Billingsley 1995). The details of constructing the martingales can be found in Yang and Zhang (2023), and the argument is very similar for the multiple treatment case so we omit it here. It follows that under $\mathbb{P}^{\theta_n}$, we have T_n is asymptotically normal. Under Assumption B.10, we derive the expression for its asymptotic variance and specify γ_1.

B.4 Additional Simulation Results

In this section, we provide some additional simulation results under different sample size N_w and under the case of $\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$.

We follow the setting of design (P1) in Section 5. In Figure B.1, to save space, we compare the performance of four estimators: `gpsm`, `m.prog`, `m.ps`, and `aipw` as sample size increases. We only show the results of estimating $\mathbb{E}\{Y(2) - Y(1)\}$ under four model specification scenarios (S1)–(S4) across three different sample sizes $N_w = 100, 200, 300$ for all w, as the results for $\mathbb{E}\{Y(3) - Y(1)\}$ and $\mathbb{E}\{Y(3) - Y(2)\}$ follow a similar pattern. The `naive` and `m.x` estimators are biased and the `aipw` estimator is doubly robust compared to the `ipw` estimator. This indicates that the `m.ds` and `aipw` are doubly robust, and as sample size increases, the variances of the estimators decreases. Figure B.2 illustrates the double robustness of the double score matching estimator when there is sample size imbalance among the treatment levels, i.e. $N_1 = 100, N_2 = 200$ and $N_3 = 300$. Lastly, we consider the case where $\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$. Let the outcome be $Y_i(w) = Z_i^{\mathrm{T}} \beta_{w,0} + \epsilon_i$, with $\beta_{1,0}^{\mathrm{T}} = (0, 1, 1, 1, 1, 1, 1)/2, \beta_{2,0}^{\mathrm{T}} = (0.5, 1, 1, 1, 1, 0, 0), \beta_{3,0}^{\mathrm{T}} = (0.5, 0, 0, 1, 1, 1, 1)$, with $\epsilon_i \sim \mathcal{N}(0, 1)$. The sample sizes are again $N_w = 300$ for all w. The results are given in Figure B.3.

We repeat the experiment over the same scenarios described above for design (P2), where the generalized propensity scores are more extreme. Under (P2), when sample sizes

across treatment levels are unequal, Figure B.4 shows that the proposed DSM estimator is still doubly robust, and is also more efficient than `ipw` and `aipw`. When the mean potential outcomes are nonzero (the outcome is generated using the same linear model as above but with coefficients $\beta_{1,0}^{\mathrm{T}} = (0, 1, 1, 1, 1)/2, \beta_{2,0}^{\mathrm{T}} = (0.5, 1, 1, 1, 1), \beta_{3,0}^{\mathrm{T}} = (0.5, 0, 0, 1, 1)$), Figure B.5 illustrates the double robustness and superior efficiency of the proposed estimator.

The coverage rates and the average of the estimated standard errors of the double score matching estimator are given in Table B.1 under design (P1) and Table B.2 under design (P2). For both designs, as the sample size for each treatment level decreases, the standard errors increase. However, the wild bootstrap procedure tends to overestimate the variance of the double score matching estimator when the sample size is small.

Figure B.1: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effect $E\{Y(2) - Y(1)\}$ under four scenarios under (P1) for the generalized propensity score (GPS) and generalized prognostic score (GPGS) for different sample sizes $N_w = 100, 200, 300$ for $w = 1, 2, 3$.

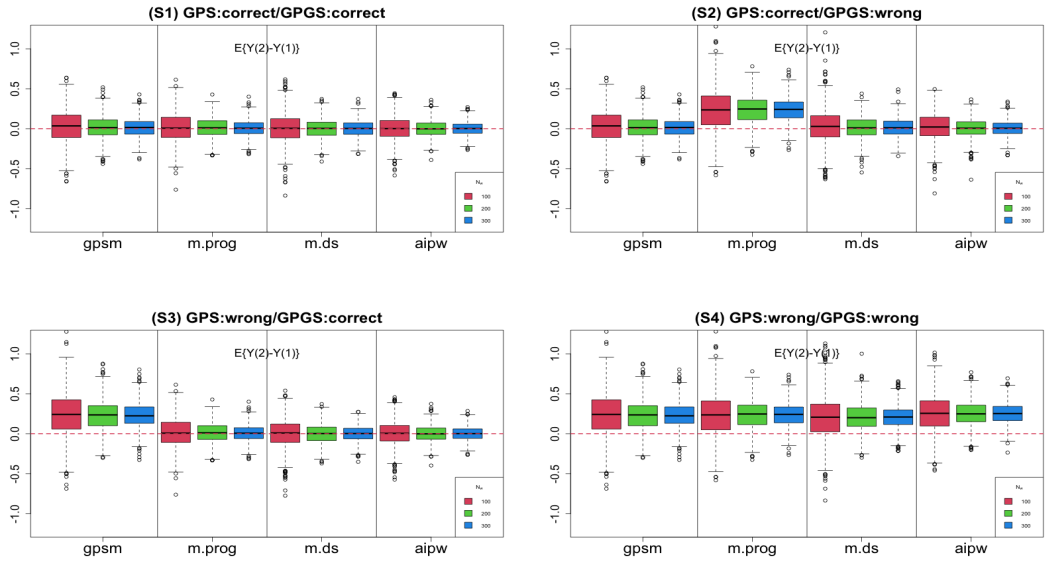


Figure B.2: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P1) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = 100, N_2 = 200$ and $N_3 = 300$.

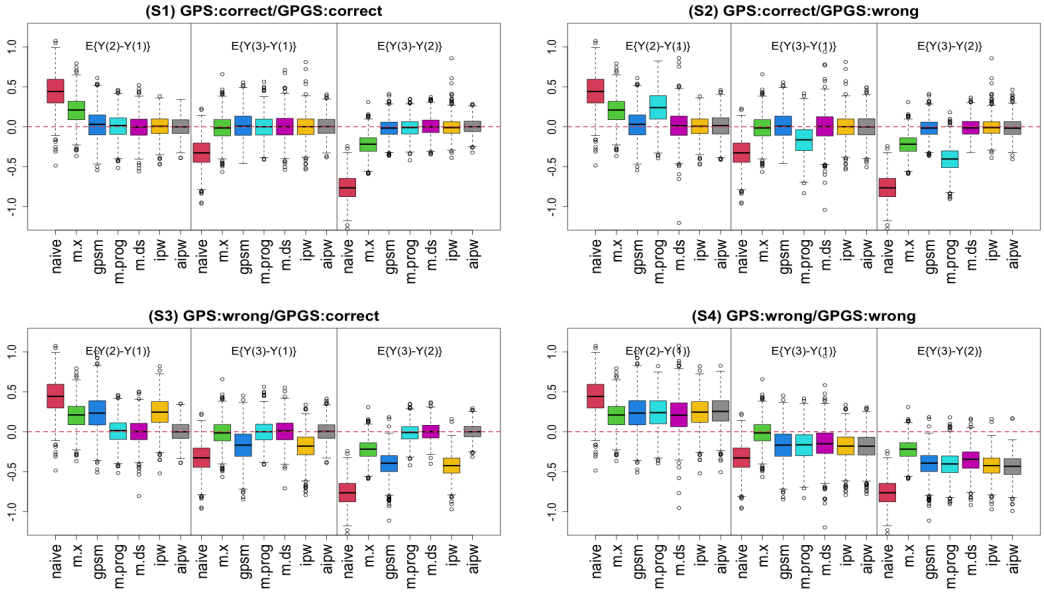


Figure B.3: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P1) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = N_2 = N_3 = 300$, when $\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$.

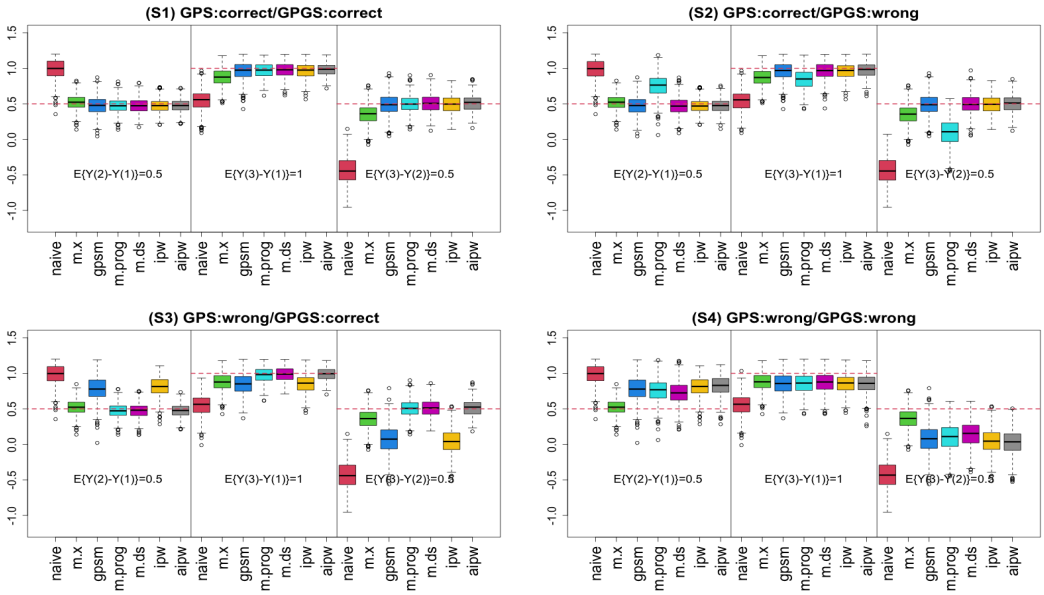


Table B.1: Simulation results based on 1000 Monte Carlo simulated datasets for the coverage rates (CR) and standard errors (SE) for the proposed double score matching estimators of the average treatment effects under four scenarios with (P1) for the generalized propensity score (GPS) and prognostic score (GPGS) models.

	$\mathbb{E}\{Y(2)-Y(1)\}$		$\mathbb{E}\{Y(3)-Y(1)\}$		$\mathbb{E}\{Y(3)-Y(2)\}$	
	CR	SE	CR	SE	CR	SE
			$N_w = 300$ for all w			
(S1) GPS:correct/GPGS:correct	94.3	0.10	95.4	0.12	95.8	0.11
(S2) GPS:correct/GPGS:wrong	94.8	0.12	97.0	0.13	96.8	0.12
(S3) GPS:wrong/GPGS:correct	95.2	0.10	95.2	0.11	94.6	0.11
(S4) GPS:wrong/GPGS:wrong	71.2	0.14	82.8	0.14	30.6	0.15
			$N_w = 200$ for all w			
(S1) GPS:correct/GPGS:correct	94.8	0.13	96.4	0.16	96.7	0.16
(S2) GPS:correct/GPGS:wrong	96.7	0.16	97.8	0.18	96.7	0.16
(S3) GPS:wrong/GPGS:correct	96.4	0.13	96.2	0.15	96.6	0.14
(S4) GPS:wrong/GPGS:wrong	80.0	0.18	88.4	0.19	50.6	0.19
			$N_w = 100$ for all w			
(S1) GPS:correct/GPGS:correct	97.3	0.22	98.9	0.33	98.7	0.31
(S2) GPS:correct/GPGS:wrong	98.0	0.26	98.5	0.34	98.4	0.31
(S3) GPS:wrong/GPGS:correct	96.9	0.20	98.4	0.30	97.4	0.28
(S4) GPS:wrong/GPGS:wrong	89.8	0.28	94.1	0.34	77.7	0.33
			$N_1 = 100, N_2 = 200, N_3 = 300$			
(S1) GPS:correct/GPGS:correct	96.1	0.19	97.3	0.19	95.0	0.12
(S2) GPS:correct/GPGS:wrong	97.4	0.23	97.4	0.22	97.5	0.14
(S3) GPS:wrong/GPGS:correct	95.8	0.18	96.4	0.18	95.7	0.12
(S4) GPS:wrong/GPGS:wrong	87.4	0.24	91.9	0.23	43.7	0.17

Figure B.4: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P2) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = 100, N_2 = 200$ and $N_3 = 300$.

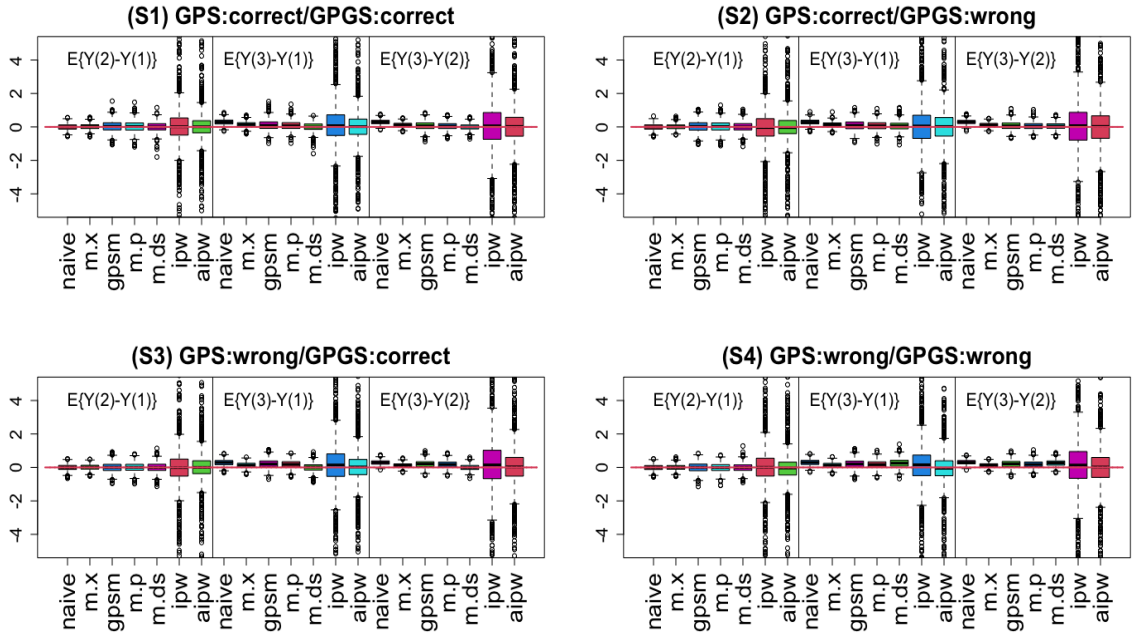


Figure B.5: Simulation result based on 1000 Monte Carlo simulated datasets for the average treatment effects under four scenarios with (P2) for the generalized propensity score (GPS) and prognostic score (GPGS) models for $N_1 = N_2 = N_3 = 300$, when $\mathbb{E}\{Y(1)\} \neq \mathbb{E}\{Y(2)\} \neq \mathbb{E}\{Y(3)\}$.

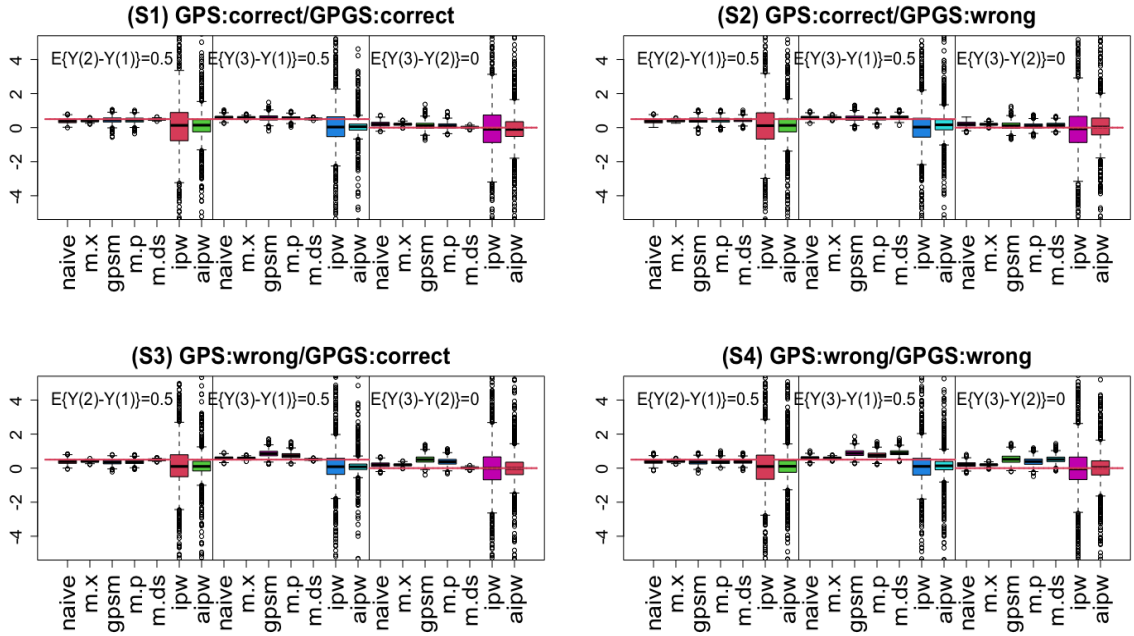


Table B.2: Simulation results based on 1000 Monte Carlo simulated datasets for the coverage rates (CR) and standard errors (SE) for the proposed double score matching estimators of the average treatment effects under four scenarios with (P2) for the generalized propensity score (GPS) and prognostic score (GPGS) models.

	$\mathbb{E}\{Y(2)-Y(1)\}$		$\mathbb{E}\{Y(3)-Y(1)\}$		$\mathbb{E}\{Y(3)-Y(2)\}$	
	CR	SE	CR	SE	CR	SE
			$N_w = 300$ for all w			
(S1) GPS:correct/GPGS:correct	95.4	0.16	94.9	0.16	95.2	0.16
(S2) GPS:correct/GPGS:wrong	95.3	0.17	93.8	0.18	94.9	0.18
(S3) GPS:wrong/GPGS:correct	94.4	0.15	93.2	0.15	93.2	0.15
(S4) GPS:wrong/GPGS:wrong	94.2	0.15	75.5	0.18	77.4	0.17
			$N_w = 200$ for all w			
(S1) GPS:correct/GPGS:correct	96.2	0.21	93.9	0.21	94.2	0.21
(S2) GPS:correct/GPGS:wrong	96.0	0.22	93.9	0.22	96.2	0.22
(S3) GPS:wrong/GPGS:correct	95.9	0.19	94.9	0.19	94.9	0.19
(S4) GPS:wrong/GPGS:wrong	95.2	0.19	84.6	0.21	82.9	0.22
			$N_w = 100$ for all w			
(S1) GPS:correct/GPGS:correct	97.6	0.40	96.4	0.35	97.1	0.35
(S2) GPS:correct/GPGS:wrong	97.7	0.40	96.0	0.36	95.7	0.36
(S3) GPS:wrong/GPGS:correct	97.8	0.35	96.6	0.32	97.0	0.32
(S4) GPS:wrong/GPGS:wrong	96.2	0.32	90.7	0.33	91.1	0.33
			$N_1 = 100, N_2 = 200, N_3 = 300$			
(S1) GPS:correct/GPGS:correct	96.6	0.33	96.6	0.30	95.4	0.19
(S2) GPS:correct/GPGS:wrong	93.7	0.29	95.1	0.27	92.8	0.20
(S3) GPS:wrong/GPGS:correct	96.6	0.31	96.5	0.28	95.7	0.18
(S4) GPS:wrong/GPGS:wrong	94.8	0.27	81.7	0.25	71.6	0.19

APPENDIX

C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

For a vector $\boldsymbol{v}$, we will use the shorthand $\boldsymbol{v}^{\otimes 2}$ to denote $\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$. We define $\mathrm{diag}(\boldsymbol{v})$ as the square matrix that has $\boldsymbol{v}$ on the diagonal and zero everywhere else. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we will use $\boldsymbol{A} \otimes \boldsymbol{B}$ to denote the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$.

C.1 Regularity Conditions and Lemmas

In this section, we provide the regularity conditions and lemmas. Recall that $\boldsymbol{\mu}_H(\omega, \boldsymbol{X}) = E\{\boldsymbol{H}(\omega)|\boldsymbol{X}\}$ and $\boldsymbol{\sigma}_H^2(\omega, \boldsymbol{X}) = \mathrm{var}\{\boldsymbol{H}(\omega)|\boldsymbol{X}\}$. Moreover, denote $S^{(0)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{j=1}^{n}\{1 + k_j(W_j)/M\}\, Y_j(t)\exp(\boldsymbol{A}_{W_j}^{\mathrm{T}}\boldsymbol{\beta})$ and $\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} S^{(0)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{j=1}^{n}\{1 + k_j(W_j)/M\}\, Y_j(t)\exp(\boldsymbol{A}_{W_j}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{A}_{W_j}$, and $\boldsymbol{S}^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}^{2}} S^{(0)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{j=1}^{n}\{1 + k_j(W_j)/M\}\, Y_j(t)\exp(\boldsymbol{A}_{W_j}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{A}_{W_j}^{\otimes 2}$, then $\widehat{\boldsymbol{Q}}(\boldsymbol{\beta}, t) = \boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$.

Assumption 23 *The following regularity conditions hold:*

(i) $(\boldsymbol{X}_i, W_i, U_i, \Delta_i)$ *are i.i.d. draws from the distribution of* $(\boldsymbol{X}_i, W_i, U_i, \Delta_i)$*;*

(ii) *the random variable* $e_\omega(\boldsymbol{X})$ *is continuously distributed, has interval support* $[\underline{p}, \bar{p}]$*, and has a density that is a continuous function on* $[\underline{p}, \bar{p}]$*;*

(iii) $\boldsymbol{\theta}^* \in int(\Theta)$ *with* Θ *compact,* $\boldsymbol{X}$ *has a bounded support, and* $E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ *is non-singular;*

(iv) $e_\omega(\cdot) : \mathbb{R} \to (0, 1)$ *is continuously differentiable with strictly positive and bounded derivative;*

(v) there exists a component of $\boldsymbol{X}$ *that is continuously distributed, has nonzero coefficient in* $\boldsymbol{\theta}^*$ *and admits a continuous density function conditional on the rest of* $\boldsymbol{X}$*;*

(vi) there exists $\varepsilon > 0$ *such that, for all* $\boldsymbol{\theta}$ *with* $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \varepsilon$, $\boldsymbol{\mu}_H \{\omega, e_\omega(\boldsymbol{X}; \boldsymbol{\theta}) = p\}$ *is Lipschitz-continuous in* p, $\sigma_H^2 \{\omega, e_\omega(\boldsymbol{X}; \boldsymbol{\theta}) = p\}$ *is continuous in* p*, and there is* $\delta > 0$ *such that* $E\left[|\boldsymbol{H}(\omega)|^{2+\delta} \mid e_\omega(\boldsymbol{X}; \boldsymbol{\theta}) = p\right]$ *is uniformly bounded;*

(vii) $E_{\boldsymbol{\theta}_n}\left[r(Y, W, \boldsymbol{X}) \mid W, e_\omega(\boldsymbol{X}; \boldsymbol{\theta}_n)\right]$ *converges to* $E[r(Y, W, \boldsymbol{X}) \mid W, e_\omega(\boldsymbol{X}; \boldsymbol{\theta}^*)]$ *almost surely, for any* $\mathbb{R}^{k+2}$*-to-*$\mathbb{R}$ *bounded and measurable function,* $r(y, w, \boldsymbol{x})$*, continuous in* $\boldsymbol{x}$*, and any sequence,* $\boldsymbol{\theta}_n \to \boldsymbol{\theta}^*$*;*

(viii) $E\left\{\mathrm{d}M^{(\omega)}(t) \mid p\right\}$ *is Lipschiz continuous in* p*;*

(ix) the time $\tau > 0$ *is such that* $\int_0^\tau \lambda_0(t)dt < \infty$*;*

(x) there exists a neighbourhood $\mathcal{B}$ *of* $\boldsymbol{\beta}^*$ *and scalar, vector and matrix functions* $s^{(0)}$,$s^{(1)}$ *and* $s^{(2)}$ *defined on* $\mathcal{B} \times [0, 1]$ *such that for* $j = 0, 1, 2$*, we have* $\sup_{t \in [0,\tau], \boldsymbol{\beta} \in \mathcal{B}} \left\| S^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}, t) \right\| \to_p 0$*;*

(xi) the functions $s^{(0)}(\cdot, t)$,$s^{(1)}(\cdot, t)$ *and* $s^{(2)}(\cdot, t)$ *are bounded and* $s^{(0)}(\cdot, t)$ *is bounded away from 0 on* $\mathcal{B} \times [0, \tau]$*;*

(xii) $\forall \epsilon > 0$*, there exists* $\delta > 0$ *such that* $||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| < \delta$ *implies* $|s^{(l)}(\boldsymbol{\beta}, t) - s^{(l)}(\boldsymbol{\beta}^*, t)| < \epsilon$*,* $\forall t \in [0, \tau]$*,* $l = 0, 1, 2$*;*

(xiii) $\boldsymbol{A}(\boldsymbol{\beta}^*)$ *is positive definite;*

The regularity conditions (i)-(vii) are almost taken directly from Abadie and Imbens (2016) and Yang et al. (2016), except we make modifications for survival outcomes and multiple treatment levels. The regularity conditions (viii)-(xiii) are standard in the survival analysis

literature and are often assumed for technical convenience; see, for instance Andersen and Gill (1982) and Fleming and Harrington (2011).

Lemma 8 below appears as Lemma S.11 in Abadie and Imbens (2016), which is useful in the proofs of our results.

Lemma 8 *Suppose that* (W_1, X_1),...,(W_n, X_n) *are independent and identically distributed, where* X *is a scalar continuous variable with a bounded support. Suppose also that* $\sigma_H^2(\omega, x)$ *is uniformly bounded over the support for* X*. Let* $n_\omega = \sum_{i=1}^n I(W_i = \omega)$ *be the number of individuals received treatment* ω*, and* $p^* = \mathrm{pr}(W = 1) > 0$ *and* $f_\omega(X)$ *be the density function of the scalar continuous variable* X *when* $W = \omega$*.Then, for a given* ω*,*

$$\begin{aligned}&\frac{1}{n_\omega}\sum_{i=1}^{n} I(W_i = \omega)\sigma_H^2(\omega, X_i)k_i(\omega)\\ &\quad\to ME\left\{\sigma_H^2(\omega, X)\left(\frac{p^*}{1-p^*}\right)^{1-2\omega}\frac{f_{1-\omega}(X)}{f_\omega(X)} \mid W_i = \omega\right\},\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n_\omega}\sum_{i=1}^{n} I(W_i = \omega)\sigma_H^2(\omega, X_i)k_i(\omega)^2\\ &\quad\to ME\left\{\sigma_H^2(\omega, X)\left(\frac{p^*}{1-p^*}\right)^{1-2\omega}\frac{f_{1-\omega}(X)}{f_\omega(X)} \mid W_i = \omega\right\}\\ &\quad+\frac{M(2M+1)}{2}E\left\{\sigma_H^2(\omega, X)\left[\left(\frac{p^*}{1-p^*}\right)^{1-2\omega}\frac{f_{1-\omega}(X)}{f_\omega(X)}\right]^2 \mid W_i = \omega\right\},\end{aligned}$$

in probability, as $n \to \infty$*.*

C.2 Proof of Asymptotic Unbiasedness of the Partial Score Function

This section includes three parts that follow the similar logic of proof. The first and the second parts provide some results useful for later sections. The proof for the asymptotic unbiasedness of $n^{-1}\boldsymbol{S}_n(\boldsymbol{\beta}^*)$ is located in the third part.

For $\omega = 0, 1, ..., J$, define $\mathrm{d}M^{(\omega)}(t) = \mathrm{d}N^{(\omega)}(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*)Y^{(\omega)}(t)$. From the standard theory for the counting process, $\mathrm{d}M^{(\omega)}(t)$ is a martingale process with respect

to the population and its baseline hazard is $\Lambda_0(t)$. Next we will prove that

$$I(W_i = \omega)\{1 + k_i(\omega)/M\}\left\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*)Y_i(t)\right\}$$

is a martingale for the imputed pseudo-population which means that the imputed pseudo-population has similar covariates distribution with the target population. First, we show that for $\omega = 0, 1, ..., J$,

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n} I(W_i = \omega)\{1 + k_i(\omega)/M\}\left\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*)Y_i(t)\right\} \\ &\to E\left\{\mathrm{d}N^{(\omega)}(t) - \mathrm{d}\Lambda_0(t)\exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*)Y^{(\omega)}(t)\right\} = E\left\{\mathrm{d}M^{(\omega)}(t)\right\}, \end{aligned} \tag{C.1}$$

as $n \to \infty$. We show (C.1) for $\omega = 1$. The proof for $\omega \neq 1$ is similar and therefore omitted. We express (C.1) for $\omega = 1$ as

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n} I\left(W_i = 1\right)\left\{1 + k_i(1)/M\right\}\left\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_1^*)Y_i(t)\right\} - E\left\{\mathrm{d}M^{(1)}(t)\right\} \\ =&n^{-1}\sum_{i=1}^{n} I\left(W_i = 1\right)\left\{1 + k_i(1)/M\right\}\left\{\mathrm{d}N_i^{(1)}(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_1^*)Y_i^{(1)}(t)\right\} - E\left\{\mathrm{d}M^{(1)}(t)\right\} \\ =&n^{-1}\sum_{i=1}^{n} I\left(W_i = 1\right)\left\{1 + k_i(1)/M\right\}\mathrm{d}M_i^{(1)}(t) - E\left\{\mathrm{d}M^{(1)}(t)\right\} \\ =&n^{-1}\sum_{i=1}^{n} I\left(W_i = 1\right)\left\{1 + k_i(1)/M\right\}\left[\mathrm{d}M_i^{(1)}(t) - E\left\{\mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_i)\right\}\right] \\ &+ n^{-1}\sum_{i=1}^{n}\left(1 - I\left(W_i = 1\right)\right)M^{-1} \\ &\qquad \sum_{j \in \mathcal{J}_M\{1, e(\boldsymbol{X}_i)\}}\left[E\left\{\mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_j)\right\} - E\left\{\mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_i)\right\}\right] \\ &+ n^{-1}\sum_{i=1}^{n} E\left\{\mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_i)\right\} - E\left\{\mathrm{d}M^{(1)}(t)\right\} \\ =&T_1 + T_2 + T_3, \end{aligned}$$

where the second line follows by the consistency assumption, and

$$
\begin{aligned}
T_1 &= n^{-1} \sum_{i=1}^{n} I\left(W_i = 1\right) \{1 + k_i(1)/M\} \left[\mathrm{d}M_i^{(1)}(t) - E\left\{ \mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_i) \right\} \right], \\
T_2 &= n^{-1} \sum_{i=1}^{n} \left(1 - I\left(W_i = 1\right)\right) M^{-1} \\
&\qquad \sum_{j \in \mathcal{J}_M\{1, e(\boldsymbol{X}_i)\}} \left[E\left\{ \mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_j) \right\} - E\left\{ \mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_i) \right\} \right] \\
T_3 &= n^{-1} \sum_{i=1}^{n} E\left\{ \mathrm{d}M^{(1)}(t) \mid e_1(\boldsymbol{X}_i) \right\} - E\left\{ \mathrm{d}M^{(1)}(t) \right\}.
\end{aligned}
\tag{C.2}
$$

Under Assumption 23 (i) - (vi), Abadie and Imbens (2006) showed that $k_i(1)^{\delta}$ is bounded almost surely for any $\delta > 0$, and the discrepancy due to matching is $O_p(n^{-1})$ for a scalar $e_1(\boldsymbol{X})$. It follows that T_1 and T_2 are consistent for zero. Lastly, by the strong law of large numbers, T_3 is consistent for zero. Therefore, (C.1) follows.

Note that by some algebra we have

$$
\sum_{i=1}^{n} I(W_i = \omega)\{1 + k_i(\omega)/M\} \mathrm{d}\Lambda_0(t) \exp(\boldsymbol{A}_{\omega}^{\mathrm{T}} \boldsymbol{\beta}^*) Y_i(t) = 0
$$

so that $\boldsymbol{S}_n(\boldsymbol{\beta}^*)$ can be equivalently expressed as

$$
\boldsymbol{S}_n(\boldsymbol{\beta}^*) = \sum_{\omega=0}^{J} \sum_{i=1}^{n} \int_0^{\tau} I(W_i = \omega)\{1 + k_i(\omega)/M\} \left\{ \boldsymbol{A}_{\omega} - \widehat{\boldsymbol{Q}}(\boldsymbol{\beta}^*, t) \right\} \mathrm{d}M_i^{(\omega)}(t).
$$

Since $n^{-1} \sum_{i=1}^{n} I(W_i = \omega)(1 + k_i(\omega)/M) \left\{ \boldsymbol{A}_{\omega} - \widehat{\boldsymbol{Q}}(\boldsymbol{\beta}^*, t) \right\} \mathrm{d}M_i^{(\omega)}(t)$ is bounded, by dominated convergence theorem,

$$
n^{-1} \boldsymbol{S}_n(\boldsymbol{\beta}^*) = n^{-1} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \int_0^{\tau} I(W_i = \omega)\{1 + k_i(\omega)/M\} \left\{ \boldsymbol{A}_{\omega} - \widehat{\boldsymbol{Q}}(\boldsymbol{\beta}^*, t) \right\} \mathrm{d}M_i^{(\omega)}(t) \tag{C.3}
$$

$$
\to \sum_{\omega=0}^{J} \int_0^{\tau} \left\{ \boldsymbol{A}_{\omega} - \widehat{\boldsymbol{Q}}(\boldsymbol{\beta}^*, t) \right\} E\left\{ \mathrm{d}M^{(\omega)}(t) \right\} = \boldsymbol{0}. \tag{C.4}
$$

C.3 Proof of Theorem 5

Taylor expansion of $\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ around $\boldsymbol{\beta}^*$ leads to

$$\mathbf{0} = \boldsymbol{S}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right) + \frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \boldsymbol{S}_n(\tilde{\boldsymbol{\beta}}) \left\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\},$$

where $\tilde{\boldsymbol{\beta}}$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$. Then

$$-\frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \frac{\boldsymbol{S}_n(\tilde{\boldsymbol{\beta}})}{n} n^{1/2} \left\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\} = n^{-1/2} \boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right).$$

Below, we derive the asymptotic distribution of $n^{-1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right)$.

Theorem 8 *Suppose Assumptions 10 and 11 and Assumption 23 hold. Then,*

$$n^{-1/2}\boldsymbol{S}_n(\boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, \boldsymbol{V}_s\} \tag{C.5}$$

as $n \to \infty$, where

$$\begin{aligned} \boldsymbol{V}_s =& E\left(\left[\sum_{\omega=0}^{J} \boldsymbol{\mu}_H\left\{\omega, e_\omega(\boldsymbol{X})\right\}\right]^{\otimes 2}\right) && \text{(C.6)} \\ &+ \sum_{\omega=0}^{J} E\left[\mathrm{cov}_H\left\{\omega, e_\omega\left(\boldsymbol{X}\right)\right\}\left\{\frac{2M+1}{2Me_\omega(\boldsymbol{X})} - \frac{e_\omega(\boldsymbol{X})}{2M}\right\}\right]. && \text{(C.7)} \end{aligned}$$

We will show that $n^{-1/2}\boldsymbol{S}_n(\boldsymbol{\beta}^*)$ can be expressed as a sum of n independent and identically distributed random vectors plus a term that converges in probability to a zero vector. We can write

$$\begin{aligned} & n^{-1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right) \\ &= n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \int_0^\infty I\left(W_i = \omega\right)\left(1 + k_i(\omega)/M\right)\left\{\boldsymbol{A}_\omega - \hat{\boldsymbol{Q}}\left(\boldsymbol{\beta}^*, t\right)\right\} \mathrm{d}M_i(t) \\ &= n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \int_0^\infty I\left(W_i = \omega\right)\left(1 + k_i(\omega)/M\right)\left\{\boldsymbol{A}_\omega - \boldsymbol{Q}\left(\boldsymbol{\beta}^*, t\right)\right\} \mathrm{d}M_i(t) \\ &\quad + n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \int_0^\infty I\left(W_i = \omega\right)\left(1 + k_i(\omega)/M\right)\left\{\boldsymbol{Q}\left(\boldsymbol{\beta}^*, t\right) - \hat{\boldsymbol{Q}}\left(\boldsymbol{\beta}^*, t\right)\right\} \mathrm{d}M_i(t). \end{aligned}$$

where

$$\boldsymbol{Q}\left(\boldsymbol{\beta}^*, t\right) = \frac{\left(\exp\left(\beta_1^*\right) S^{(1)}(t), \cdots, \exp\left(\beta_J^*\right) S^{(J)}(t)\right)^{\mathrm{T}}}{S^{(0)}(t) + \sum_{\omega=1}^{J} \exp\left(\beta_\omega^*\right) S^{(w)}(t)}$$

with $S^{(w)}(t) = E\{Y^{(w)}(t)\}$. By Assumption 23 (x)-(xii), the second term converges to zero in probability. Recall that we defined $\boldsymbol{H}_i(\omega) = \int_0^\infty \{\boldsymbol{A}_\omega - \boldsymbol{Q}\left(\boldsymbol{\beta}^*, t\right)\} \mathrm{d}M_i(t)$. Therefore, we have

$$n^{-1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right) = n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} I\left(W_i = \omega\right)\left(1 + k_i(\omega)/M\right) \boldsymbol{H}_i(\omega) + o_p(1).$$

Let $\boldsymbol{\mu}_H\left(\omega, e_\omega\left(\boldsymbol{X}_i\right)\right) = E\left\{H(\omega) \mid e_\omega\left(\boldsymbol{X}_i\right)\right\}$, and $\mathrm{cov}_H\left(\omega, e_\omega\left(\boldsymbol{X}_i\right)\right) = \mathrm{cov}\left\{H(\omega) \mid e_\omega\left(\boldsymbol{X}_i\right)\right\}$, we have

$$\begin{aligned}
& n^{-1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right) \\
=& n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} I\left(W_i = \omega\right)\left(1 + k_i(\omega)/M\right)\left\{\boldsymbol{H}_i(\omega) - \boldsymbol{\mu}_H\left(\omega, e_\omega\left(\boldsymbol{X}_i\right)\right)\right\} \\
& - n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \left(1 - I\left(W_i = \omega\right)\right) M^{-1} \sum_{j \in \mathcal{J}_M\{\omega, e(\boldsymbol{X}_i)\}} \left\{\boldsymbol{\mu}_H\left(\omega, e_\omega\left(\boldsymbol{X}_j\right)\right) - \boldsymbol{\mu}_H\left(\omega, e_\omega\left(\boldsymbol{X}_i\right)\right)\right\} \\
& + n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \boldsymbol{\mu}_H\left(\omega, e_\omega\left(\boldsymbol{X}_i\right)\right) + o_p(1)
\end{aligned}$$

Similar to the argument for T_2 given in the previous section, the matching discrepancy bias goes to zero, i.e. the second term converges in probability to zero. Therefore,

$$\begin{aligned}
n^{-1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta}^*\right) =& n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} I\left(W_i = \omega\right)\left\{1 + k_i(\omega)/M\right\}\left[\boldsymbol{H}_i(\omega) - \boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i\right)\right\}\right] \\
& + n^{-1/2} \sum_{\omega=0}^{J} \sum_{i=1}^{n} \boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i\right)\right\} + o_p(1) \\
=& \sum_{l=1}^{2n} \xi_{n,l} + o_p(1)
\end{aligned}$$

where

$$\xi_{n,l} = \begin{cases} n^{-1/2} \sum_{\omega=0}^{J} \boldsymbol{\mu}_H \left\{\omega, e_\omega \left(\boldsymbol{X}_l\right)\right\} & \\ & 1 \leq l \leq n \\ n^{-1/2} \sum_{\omega=0}^{J} I\left(W_{l-n} = \omega\right) \left\{1 + k_{l-n}(\omega)/M\right\} \left[\boldsymbol{H}_{l-n}(\omega) - \boldsymbol{\mu}_H \left\{\omega, e_\omega \left(\boldsymbol{X}_{l-n}\right)\right\}\right] & \\ & n+1 \leq l \leq 2n \end{cases}$$

Consider the $\sigma-$fields

$$\mathcal{F}_{n,l} = \begin{cases} \sigma\left\{W_1, \cdots, W_l, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_l\right\}, & 1 \leq l \leq n \\ \sigma\left\{W_1, \cdots, W_n, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{H}_1\left(W_1\right), \cdots, \boldsymbol{H}_{l-n}\left(W_{l-n}\right)\right\}, & n+1 \leq l \leq 2n \end{cases}$$

Then for each $n \geq 1$,

$$\left\{\sum_{j=1}^{l} \xi_{n,j}, \mathcal{F}_{n,l}, 1 \leq l \leq 2n\right\}$$

is a martingale. Based on the central limit theorem for martingale arrays, we have

$$\mathrm{cov}_{s1} = \mathrm{plim} \sum_{l=1}^{n} E\left\{\xi_{n,l}^2 \mid \mathcal{F}_{n,l-1}\right\} = E\left(\left[\sum_{\omega=0}^{J} \boldsymbol{\mu}_H \left\{\omega, e_\omega(\boldsymbol{X})\right\}\right]^{\otimes 2}\right)$$

and based on Lemma 8,

$$\begin{aligned} \mathrm{cov}_{s2} &= \mathrm{plim} \sum_{l=n+1}^{2n} E\left\{\xi_{n,l}^2 \mid \mathcal{F}_{n,l-1}\right\} \\ &= \mathrm{plim}\, n^{-1} \sum_{\omega=0}^{J} \sum_{i=1}^{n} I\left(W_i = \omega\right) \left\{1 + k_i(\omega)/M\right\}^2 \mathrm{cov}_H \left\{\omega, e_\omega\left(\boldsymbol{X}_i\right)\right\} \\ &= \sum_{\omega=0}^{J} E\left[\mathrm{cov}_H \left\{\omega, e_\omega\left(\boldsymbol{X}\right)\right\} \left\{\frac{2M+1}{2M e_\omega(\boldsymbol{X})} - \frac{e_\omega(\boldsymbol{X})}{2M}\right\}\right]. \end{aligned}$$

This concludes the proof of Theorem 8.

Furthermore we have

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{\beta}^*) &= -\operatorname{plim}\frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\frac{\boldsymbol{S}_n(\tilde{\boldsymbol{\beta}})}{n} \\ &= E\left[\int_0^\tau \left\{\operatorname{diag}\left(\boldsymbol{Q}\left(\boldsymbol{\beta}^*,t\right)\right)-\boldsymbol{Q}\left(\boldsymbol{\beta}^*,t\right)^{\otimes 2}\right\}\left\{\sum_{\omega=0}^{J}\mathrm{d}N^{(\omega)}(t)\right\}\right].\end{aligned}$$

By Slutsky's Theorem, the GPSM estimator $\hat{\boldsymbol{\beta}}$ is asymptotic normal with mean 0 and covariance matrix

$$\boldsymbol{V}=\boldsymbol{A}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{V}_s\boldsymbol{A}(\boldsymbol{\beta}^*)^{-1}.$$

This concludes the proof of Theorem 5.

C.4 Proof of Theorem 6

The log-likelihood function of $\boldsymbol{\theta}$ is

$$l(\boldsymbol{\theta}\mid\underline{O})=\sum_{i=1}^{n}\log\mathrm{P}\left(W_i\mid\boldsymbol{X}_i,\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\sum_{\omega=0}^{J}I\left(W_i=\omega\right)\log e_\omega(\boldsymbol{X}_i;\boldsymbol{\theta}),$$

where $\underline{O}$ is the observed data. The $pJ\times 1$ normalized score function is

$$\begin{aligned}\boldsymbol{\Delta}_n(\boldsymbol{\theta}) &= n^{-1/2}\frac{\partial}{\partial\boldsymbol{\theta}}l(\boldsymbol{\theta}\mid\underline{O})=n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\frac{I\left(W_i=\omega\right)}{e_\omega(\boldsymbol{X}_i;\boldsymbol{\theta})}\begin{pmatrix} f\left(\omega;1\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right)\boldsymbol{X}_i \\ \vdots \\ f\left(\omega;J\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right)\boldsymbol{X}_i\end{pmatrix} \\ &= n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\frac{I\left(W_i=\omega\right)}{e_\omega(\boldsymbol{X}_i;\boldsymbol{\theta})}\boldsymbol{f}\left(\omega\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right)\otimes\boldsymbol{X}_i,\end{aligned}$$

where for $\omega\in\{0,...,J\}$ and $\omega'\in\{1,...,J\}$, we define $f\left(\omega;\omega'\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right)\boldsymbol{X}_i=\frac{\partial}{\partial\boldsymbol{\theta}_{w'}}e_\omega(\boldsymbol{X}_i;\boldsymbol{\theta})$, and $\boldsymbol{f}\left(\omega\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right)=\left(f\left(\omega;1\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right),...,f\left(\omega;J\mid\boldsymbol{X}_i;\boldsymbol{\theta}\right)\right)^{\mathrm{T}}$.

Following Abadie and Imbens (2016), we use the local experiment argument. Let $\boldsymbol{\theta}_n=\boldsymbol{\theta}^*+n^{-1/2}\boldsymbol{h}$ where $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$. Define $P^{\boldsymbol{\theta}_n}$ as the data distribution

under $\{e_\omega(\boldsymbol{X};\boldsymbol{\theta}), \omega = 0, \cdots, J\}$. Then under $P^{\boldsymbol{\theta}_n}$, define

$$\begin{aligned}\Gamma_n\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n\right) &= l\left(\boldsymbol{\theta}^* \mid \underline{O}\right) - l\left(\boldsymbol{\theta}_n \mid \underline{O}\right) \\ &= \sum_{i=1}^{n}\sum_{\omega=0}^{J}\left\{\log \mathrm{P}\left(\omega \mid \boldsymbol{X}_i;\boldsymbol{\theta}^*\right) - \log \mathrm{P}\left(\omega \mid \boldsymbol{X}_i;\boldsymbol{\theta}_n\right)\right\}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{D}_n\left(\boldsymbol{\theta}_n\right) =& n^{-1/2}\sum_{\omega=0}^{J}\sum_{i=1}^{n} I\left(W_i = \omega\right)\left\{1 + k_i(\omega)\right\}\left[\boldsymbol{H}_i(\omega) - \boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i;\boldsymbol{\theta}_n\right)\right\}\right] \\ &+ n^{-1/2}\sum_{\omega=0}^{J}\sum_{i=1}^{n}\boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i;\boldsymbol{\theta}_n\right)\right\} + o_p(1)\end{aligned}$$

Note that

$$n^{-1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta}^* \mid \boldsymbol{\theta}_n\right) = \boldsymbol{D}_n\left(\boldsymbol{\theta}_n\right) + o_p(1)$$

Our goal is to show that under $P^{\boldsymbol{\theta}_n}$, when $n \to \infty$, the joint limiting distribution

$$\begin{pmatrix} \boldsymbol{D}_n\left(\boldsymbol{\theta}_n\right) \\ n^{1/2}\left\{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right\} \\ \Gamma_n\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left\{\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ -\boldsymbol{h}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}^*}\boldsymbol{h}/2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_s & \boldsymbol{C}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}^*}^{-1} & -\boldsymbol{C}^{\mathrm{T}}\boldsymbol{h} \\ \boldsymbol{I}_{\boldsymbol{\theta}^*}^{-1}\boldsymbol{C} & \boldsymbol{I}_{\boldsymbol{\theta}^*}^{-1} & -\boldsymbol{h} \\ -\boldsymbol{h}^{\mathrm{T}}\boldsymbol{C} & -\boldsymbol{h}^{\mathrm{T}} & \boldsymbol{h}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}^*}\boldsymbol{h} \end{pmatrix}\right\},$$

where

$$\boldsymbol{C} = \sum_{\omega=0}^{J} E\left[\frac{1}{e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right) \otimes \operatorname{cov}\left\{\boldsymbol{X}, \boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}\right) \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right]$$

is a $pJ \times J$ matrix, and

$$\boldsymbol{I}\left(\boldsymbol{\theta}^*\right) = \sum_{\omega=0}^{J} E\left[\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes E\left\{\boldsymbol{X}^{\otimes 2} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right]$$

is the $pJ \times pJ$ fisher information matrix. Then we can invoke Le Cam's third lemma (Le Cam et al. 2000) to obtain our final result.

Since we can write

$$n^{1/2}\left\{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right\} = n^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}^*}^{-1}\boldsymbol{\Delta}_n\left(\boldsymbol{\theta}_n\right) + o_p(1)$$
$$n^{-1/2}\Gamma\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n\right) = -n^{-1/2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{\Delta}_n\left(\boldsymbol{\theta}_n\right) - \frac{1}{2}n^{-1/2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}^*}\boldsymbol{h} + o_p(1),$$

it suffices to show

$$\begin{pmatrix} \boldsymbol{D}_n\left(\boldsymbol{\theta}_n\right) \\ \boldsymbol{\Delta}_n\left(\boldsymbol{\theta}_n\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left\{\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_s & \boldsymbol{C}^{\mathrm{T}} \\ \boldsymbol{C} & \boldsymbol{I}_{\boldsymbol{\theta}^*} \end{pmatrix}\right\}.$$

Towards this end, we consider the linear combination

$$\begin{aligned}
L_n =& \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{D}_n\left(\boldsymbol{\theta}_n\right) + \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\Delta}_n\left(\boldsymbol{\theta}_n\right) \\
=& n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J} I\left(W_i = \omega\right)\left\{1 + k_i(\omega)/M\right\}\boldsymbol{z}_1^{\mathrm{T}}\left[\boldsymbol{H}_i(\omega) - \boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i; \boldsymbol{\theta}_n\right)\right\}\right] \\
&+ n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i; \boldsymbol{\theta}_n\right)\right\} \\
&+ n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\boldsymbol{z}_2^{\mathrm{T}}\frac{I\left(W_i = \omega\right)}{e_\omega(\boldsymbol{X}_i; \boldsymbol{\theta}_n)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X}_i; \boldsymbol{\theta}_n\right) \otimes \boldsymbol{X}_i \\
=& n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J} I\left(W_i = \omega\right)\left\{1 + k_i(\omega)/M\right\}\boldsymbol{z}_1^{\mathrm{T}}\left\{\boldsymbol{H}_i(\omega) - \boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}_i\right)\right\} \\
&+ n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J} I\left(W_i = \omega\right)\left\{1 + k_i(\omega)/M\right\}\boldsymbol{z}_1^{\mathrm{T}}\left[\boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}_i\right) - \boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i; \boldsymbol{\theta}_n\right)\right\}\right] \\
&+ n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_i; \boldsymbol{\theta}_n\right)\right\} \\
&+ n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\boldsymbol{z}_2^{\mathrm{T}}\frac{I\left(W_i = \omega\right)}{e_\omega(\boldsymbol{X}_i; \boldsymbol{\theta}_n)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X}_i; \boldsymbol{\theta}_n\right) \otimes \left[\boldsymbol{X}_i - E\left\{\boldsymbol{X}_i \mid e_\omega\left(\boldsymbol{X}_i; \boldsymbol{\theta}_n\right)\right\}\right] \\
&+ n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{J}\boldsymbol{z}_2^{\mathrm{T}}\frac{I\left(W_i = \omega\right)}{e_\omega(\boldsymbol{X}_i; \boldsymbol{\theta}_n)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X}_i; \boldsymbol{\theta}_n\right) \otimes E\left\{\boldsymbol{X}_i \mid e_\omega\left(\boldsymbol{X}_i; \boldsymbol{\theta}_n\right)\right\}.
\end{aligned}$$

We write L_n as $L_n = \sum_{l=1}^{3n} \xi_{n,l} + o_p(1)$ where

$$\xi_{n,l}$$
$$= \begin{cases} n^{-1/2} \sum_{\omega=0}^{J} \boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{\mu}_H \left\{\omega, e_\omega\left(\boldsymbol{X}_l; \boldsymbol{\theta}_n\right)\right\} \\ +\boldsymbol{z}_2^{\mathrm{T}} \frac{I(W_l=\omega)}{e_\omega(\boldsymbol{X}_l;\boldsymbol{\theta}_n)} \boldsymbol{f}\left(\omega \mid \boldsymbol{X}_l; \boldsymbol{\theta}_n\right) \otimes E\left\{\boldsymbol{X}_l \mid e_\omega\left(\boldsymbol{X}_l; \boldsymbol{\theta}_n\right)\right\} \\ \qquad\qquad \text{for } \ 1 \leq l \leq n; \\ n^{-1/2} \sum_{\omega=0}^{J} I\left(W_{l-n}=\omega\right) \\ \left(\left\{1+k_{l-n}(\omega)/M\right\} \boldsymbol{z}_1^T\left[\boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}_{l-n}\right)-\boldsymbol{\mu}_H\left\{\omega, e_\omega\left(\boldsymbol{X}_{l-n}; \boldsymbol{\theta}_n\right)\right\}\right]\right. \\ \left.+\boldsymbol{z}_2^{\mathrm{T}} \frac{1}{e_\omega(\boldsymbol{X}_{l-n};\boldsymbol{\theta}_n)} \boldsymbol{f}\left(\omega \mid \boldsymbol{X}_{l-n}; \boldsymbol{\theta}_n\right) \otimes\left[\boldsymbol{X}_{l-n}-E\left\{\boldsymbol{X}_{l-n} \mid e_\omega\left(\boldsymbol{X}_{l-n}; \boldsymbol{\theta}_n\right)\right\}\right]\right) \\ \qquad\qquad \text{for } \ n+1 \leq l \leq 2n; \\ n^{-1/2} \sum_{\omega=0}^{J} I\left(W_{l-2n}=\omega\right)\left\{1+k_{l-2n}(\omega)/M\right\} \boldsymbol{z}_1^{\mathrm{T}}\left\{\boldsymbol{H}_{l-2n}(\omega)-\boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}_{l-2n}\right)\right\} \\ \qquad\qquad \text{for } \ 2n+1 \leq l \leq 3n. \end{cases}$$

Consider the σ-fields

$$\mathcal{F}_{n,l} = \begin{cases} \sigma\left\{W_1, \cdots, W_l, \boldsymbol{X}_1^{\mathrm{T}} \boldsymbol{\theta}_n, \cdots, \boldsymbol{X}_l^{\mathrm{T}} \boldsymbol{\theta}_n\right\} \\ \qquad\qquad \text{for } \ 1 \leq l \leq n; \\ \sigma\left\{W_1, \cdots, W_n, \boldsymbol{X}_n^{\mathrm{T}} \boldsymbol{\theta}_n, \cdots, \boldsymbol{X}_n^{\mathrm{T}} \boldsymbol{\theta}_n, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_{l-n}\right\} \\ \qquad\qquad \text{for } \ n+1 \leq l \leq 2n; \\ \sigma\left\{W_1, \cdots, W_n, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{H}_1\left(W_1\right), \cdots, \boldsymbol{H}_{l-2n}\left(W_{l-2n}\right)\right\} \\ \qquad\qquad \text{for } \ 2n+1 \leq l \leq 3n. \end{cases}$$

Then for each $n \geq 1$,

$$\left\{\sum_{j=1}^{l} \xi_{n,j}, \mathcal{F}_{n,l}, 1 \leq l \leq 2n\right\}$$

is a martingale. By applying the martingale central limit theorem, we obtain

$$L_N \xrightarrow{d} N\left\{0, \sigma_{L,1}^2+\sigma_{L,2}^2+\sigma_{L,3}^2\right\},$$

where

$$\begin{aligned}\sigma_{L,1}^2 =& \operatorname{plim} \sum_{l=1}^{n} E\left\{\xi_{n,l}^2 \mid \mathcal{F}_{n,l-1}\right\} \\ =& \boldsymbol{z}_1^{\mathrm{T}} E\left(\left[\sum_{\omega=0}^{J} \boldsymbol{\mu}_H\left\{\omega, e_\omega(\boldsymbol{X};\boldsymbol{\theta}^*)\right\}\right]^{\otimes 2}\right) \boldsymbol{z}_1 \\ &+ \sum_{\omega=0}^{J} \boldsymbol{z}_2^{\mathrm{T}} E\left[\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)} \boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes E\left\{\boldsymbol{X} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}^{\otimes 2}\right] \boldsymbol{z}_2\end{aligned}$$

$$\begin{aligned}\sigma_{L,2}^2 =& \operatorname{plim} \sum_{l=n+1}^{2n} E\left\{\xi_{n,l}^2 \mid \mathcal{F}_{n,l-1}\right\} \\ =& \sum_{\omega=0}^{J} \boldsymbol{z}_1^{\mathrm{T}} E\left[\left\{\frac{2M+1}{2M e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)} - \frac{e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}{2M}\right\} \operatorname{var}\left\{\boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}\right) \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right] \boldsymbol{z}_1 \\ &+ \sum_{\omega=0}^{J} \boldsymbol{z}_2^{\mathrm{T}} E\left[\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)} \boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes \operatorname{var}\left\{\boldsymbol{X} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right] \boldsymbol{z}_2 \\ &+ 2\sum_{\omega=0}^{J} \boldsymbol{z}_2^{\mathrm{T}} E\left[\frac{1}{e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)} \boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right) \otimes \operatorname{cov}\left\{\boldsymbol{X}, \boldsymbol{\mu}_H\left(\omega, \boldsymbol{X}\right) \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right] \boldsymbol{z}_1\end{aligned}$$

$$\begin{aligned}\sigma_{L,3}^2 =& \operatorname{plim} \sum_{l=2n+1}^{3n} E\left\{\xi_{n,l}^2 \mid \mathcal{F}_{n,l-1}\right\} \\ =& \sum_{\omega=0}^{J} \boldsymbol{z}_1^{\mathrm{T}} E\left[\left\{\frac{2M+1}{2M e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)} - \frac{e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}{2M}\right\} E\left\{\operatorname{cov}_H\left(\omega, \boldsymbol{X}\right) \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right] \boldsymbol{z}_1.\end{aligned}$$

Note that

$$
\begin{aligned}
&\boldsymbol{I}\left(\boldsymbol{\theta}^*\right) \\
=&\sum_{\omega=0}^{J} E\left[\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes E\left\{\boldsymbol{X}^{\otimes 2} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right] \\
=&\sum_{\omega=0}^{J} E\left(\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes \left[E\left\{\boldsymbol{X} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}^{\otimes 2} + \operatorname{var}\left\{\boldsymbol{X} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right]\right) \\
=&\sum_{\omega=0}^{J} E\left[\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes E\left\{\boldsymbol{X} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}^{\otimes 2}\right] \\
&+\sum_{\omega=0}^{J} E\left[\frac{1}{e_\omega^2\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)}\boldsymbol{f}\left(\omega \mid \boldsymbol{X};\boldsymbol{\theta}^*\right)^{\otimes 2} \otimes \operatorname{var}\left\{\boldsymbol{X} \mid e_\omega\left(\boldsymbol{X};\boldsymbol{\theta}^*\right)\right\}\right].
\end{aligned}
$$

Therefore we have

$$\sigma_{L,1}^2 + \sigma_{L,2}^2 + \sigma_{L,3}^2 = \boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{V}_s \boldsymbol{z}_1 + \boldsymbol{z}_2^{\mathrm{T}} I\left(\boldsymbol{\theta}^*\right) \boldsymbol{z}_2 + 2\boldsymbol{z}_2^{\mathrm{T}} \boldsymbol{C} \boldsymbol{z}_1.$$

The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ based on the estimated GPS is

$$n^{1/2}\left\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\} \to \mathcal{N}\left\{\boldsymbol{0}, \boldsymbol{V}_{adj}\right\}$$

where

$$\boldsymbol{V}_{adj} = A\left(\boldsymbol{\beta}^*\right)^{-1} \widetilde{\boldsymbol{V}}_s A\left(\boldsymbol{\beta}^*\right)^{-1}$$

and

$$\widetilde{\boldsymbol{V}}_s = \boldsymbol{V}_s - \boldsymbol{C}^{\mathrm{T}} \boldsymbol{I}\left(\boldsymbol{\theta}^*\right)^{-1} \boldsymbol{C}.$$

This concludes the proof of Theorem 6.

C.5 Additional Simulations

C.5.1 An algorithm that simulates the potential survival times

We describe a way to generate the potential survival times in our simulation studies so that they depend on the covariates while still satisfies the marginal proportional hazard model. Let $\boldsymbol{\eta} = (\eta_1, ..., \eta_5)^{\mathrm{T}}$ be the coefficients of $(X_1, ..., X_5)$. Let $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_5)^{\mathrm{T}}$ be such that each covariate X_k is generated from an exponential distribution with mean

$1/\lambda_k$, and λ_0 be the baseline hazard corresponding to the baseline survival function $S^{(0)}(t) = \exp(-\lambda_0 t)$. Under Algorithm 1 we have

---

Algorithm 1 for generating $T^{(0)}, ..., T^{(J)}$ that satisfies the marginal PH model

---

Step 1. Generate $T^{(0)}$ from $S^{(0)}(t) = \exp(-\lambda_0 t)$, where $\lambda_0 = 1/3$ for $\boldsymbol{\beta}^* = 0$, and $\lambda_0 = 3$ for $\boldsymbol{\beta}^* = (0.1, 0.2, 0.3, 0.4)^{\mathrm{T}}$.

Step 2. Generate u from Unif$[0, 1]$. For each $\omega \in \{1, ..., 4\}$, solve

$$\left\{\prod_{k=1}^{5}\left(1 - \frac{\eta_k t}{\lambda_k}\right)\right\} \exp\left[\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\eta} - \lambda_0 \exp(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*)\}\, t\right] - 1 + u = 0$$

for t, where $\boldsymbol{\eta} = (-9/2, -6/2, -3/2, -3/2, 3/2)^{\mathrm{T}}$ and $\boldsymbol{\lambda} = (2, 2, 2, 2, 2)^{\mathrm{T}}$. Let $T^{(w)}$ be the solution t.

$$S_{T|W,\boldsymbol{X}}(t \mid W = \omega, \boldsymbol{X}) = \left\{\prod_{k=1}^{5}\left(1 - \frac{\eta_k}{\lambda_k}t\right)\right\} \exp\left\{\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\eta} - \lambda_0 \exp\left(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*\right)\right] t\right\}.$$

By the choices of parameters, it is easy to see that $S_{T|W,\boldsymbol{X}}(t = 0 \mid W = \omega, \boldsymbol{X}) = 1$. Also $S_{T|W,\boldsymbol{X}}(t = \tau \mid W = \omega, \boldsymbol{X} = \boldsymbol{x}) = 0$ because $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\eta} - \lambda_0 \exp\left(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*\right) < 0$. Finally, we have that $\mathrm{d}S_{T|W,\boldsymbol{X}}(t \mid W = w, \boldsymbol{X} = \boldsymbol{x})/\mathrm{d}t \leq 0$ because of $\sum_{i=1}^{5} \lambda_i^{-1} \leq \min\left(\lambda_0, \lambda_0 e^{\beta_1^*}, ..., \lambda_0 e^{\beta_J^*}\right)$.

The marginal distribution of $T^{(w)}$ is

$$\begin{aligned}
S^{(\omega)}(t) &= \int S_{T|W,\boldsymbol{X}}(t \mid W = \omega, \boldsymbol{X} = \boldsymbol{x}) \mathrm{d}F(\boldsymbol{x}) \\
&= \int \left\{\prod_{k=1}^{5}\left(1 - \frac{\eta_k}{\lambda_k}t\right)\right\} \exp\left\{\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\eta} - \lambda_0 \exp\left(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}\right)\right] t\right\} \mathrm{d}F(\boldsymbol{x}) \\
&= S^{(0)}(t)^{\exp\left(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*\right)}
\end{aligned}$$

Therefore, the marginal distribution of $T^{(w)}$ satisfies $S^{(\omega)}(t) = S^{(0)}(t)^{\exp\left(\boldsymbol{A}_\omega^{\mathrm{T}}\boldsymbol{\beta}^*\right)}$.

C.5.2 Additional results when the causal effects are present

This section shows the simulation results when causal effects are present, i.e., $\boldsymbol{\beta}^* = (0.1, 0.2, 0.3, 0.4)^{\mathrm{T}}$. Figure C.1 and C.2 show the simulation results when the GPS model is correctly and incorrectly specified, respectively. The performance of competing estimators is similar to the case when there is no treatment effect.

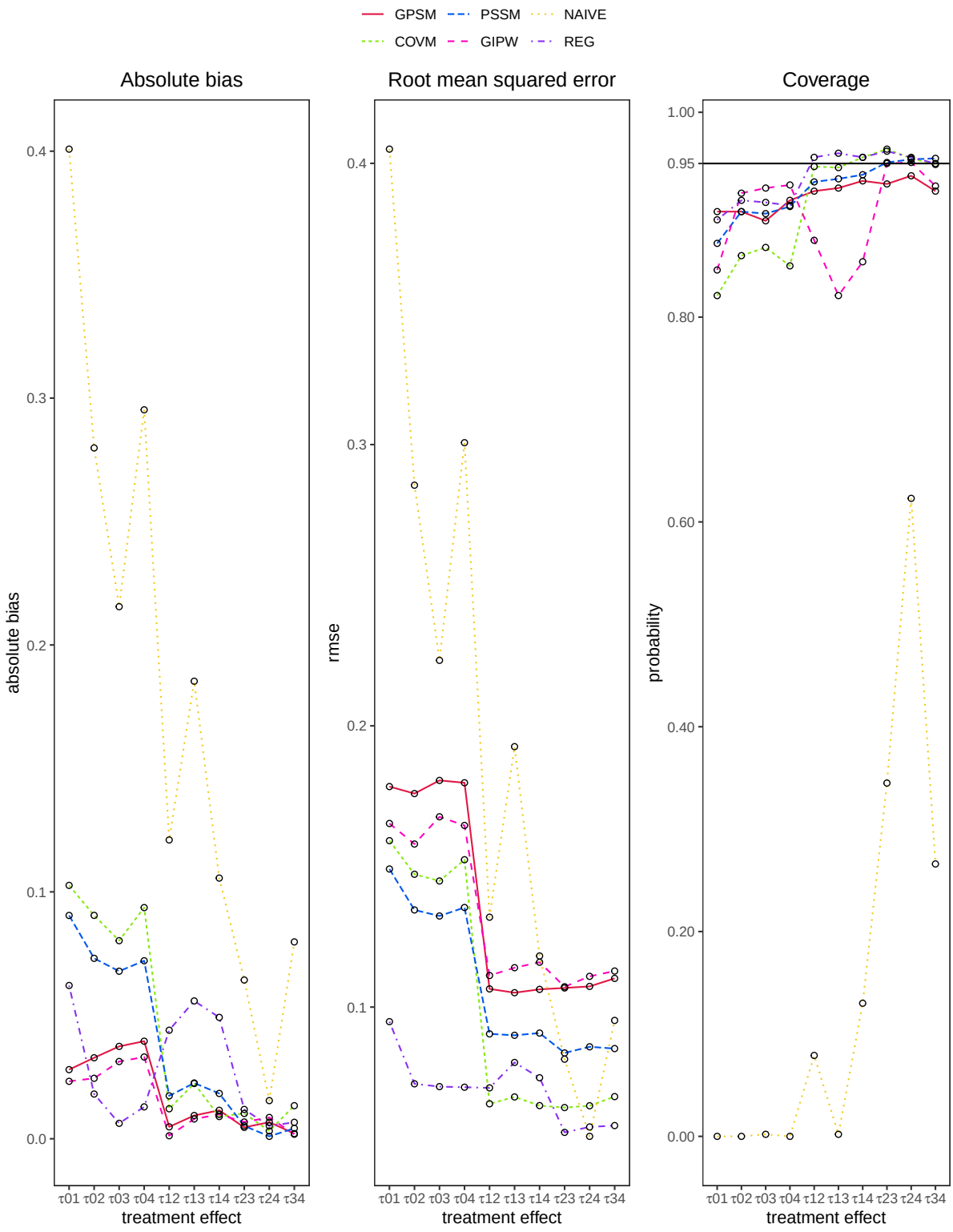


Figure C.1: Simulation results when the GPS model is *correctly* specified and treatment effect is present.

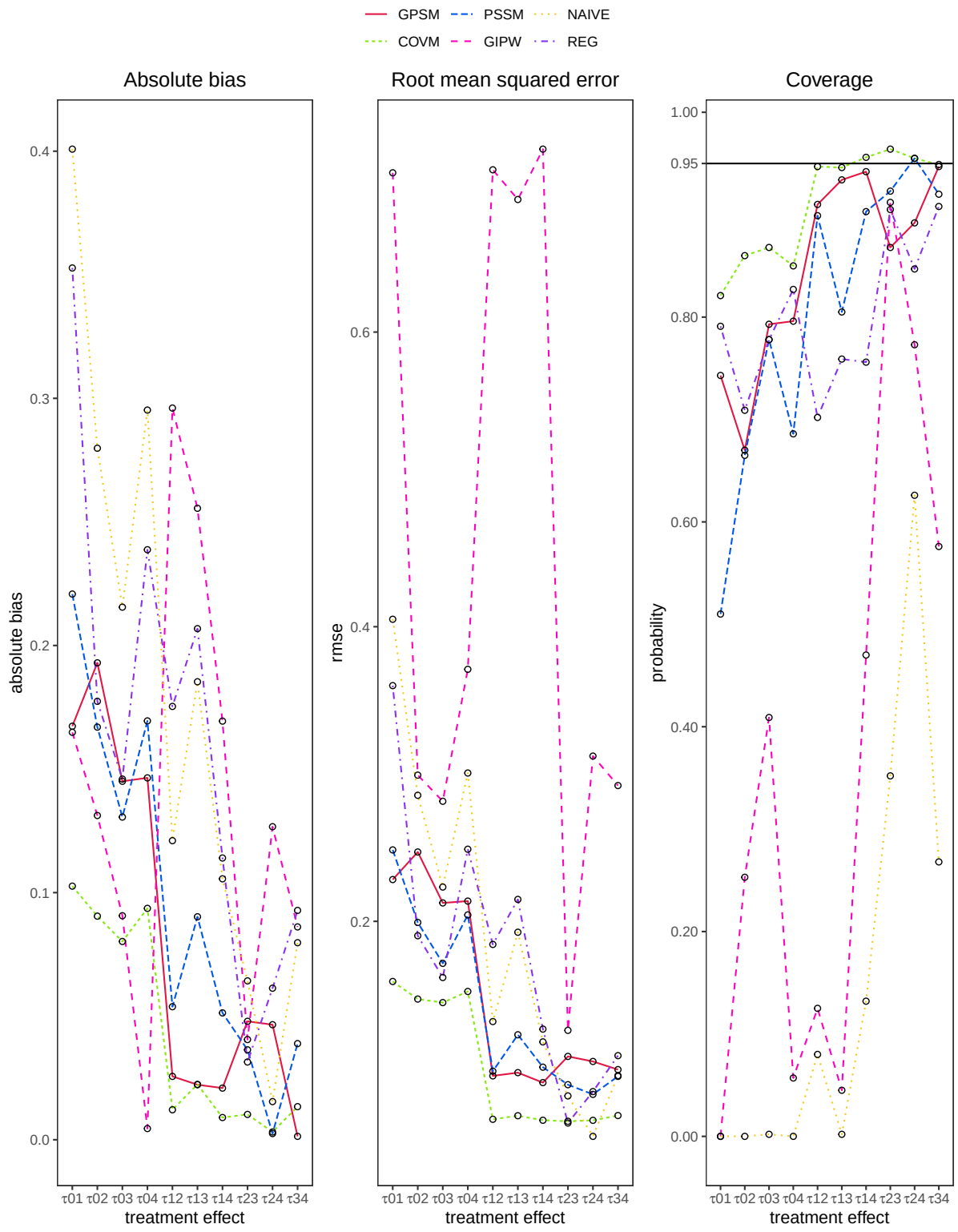


Figure C.2: Simulation results when the GPS model is *correctly* specified and treatment effect is present.