# Causal Inference via Propensity Score Regression and Length-Biased Sampling

Ashkan Ertefaie

Doctor of Philosophy

Department of Mathematics and Statistics

McGill University

Montreal,Quebec

2011-05-21

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctorate of Philosophy

# DEDICATION

To my parents, Maryam and Parviz, and my brother, Aria.

## ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance and help of my supervisors, Dr. Masoud Asgharian and Dr. David A. Stephens. I would like to offer my sincere gratitude to them who supported me with their patience and knowledge from the preliminary to the concluding level of this thesis. They were always accessible and willing to help me with my research, and their insights made research life smooth and rewarding for me.

I thank Mr. Schulich for his generosity in funding the Schulich Scholarship. I am much honoured to be recipient of this award during the last two years of my studies.

For her understanding and assistance, a special thanks as well to my girlfriend, Sophie Mongrain, who inspired my efforts and her love and dedication has taken the load off my shoulder. She has been by my side every step along these taxing years. I also thank her for helping me with the French translation of the abstract.

Finally, I offer my regards and blessings to all of those who supported me in any respect during the completion of my studies.

## STATEMENT OF ORIGINALITY

Chapter 2 of this thesis is partially borrowed from a published joint work with Professor David A. Stephens (Ertefaie & Stephens (2010)). In this joint work, I did most of the computations and co-wrote the manuscript. Using simulation studies and real data analysis, we showed that propensity score regression produces a causal effect estimator which has a smaller variance than the one obtained by inverse probability of treatment weighting and augmented inverse probability weighed complete case methods.

# ABSTRACT

Confounder adjustment is the key in the estimation of exposure effect in observational studies. Two well known causal adjustment techniques are the propensity score and the inverse probability of treatment weighting. We have compared the asymptotic properties of these two estimators and showed that the former method results in a more efficient estimator. Since ignoring important confounders result in a biased estimator, it seems beneficial to adjust for all the covariates. This, however, may result in an inflation of the variance of the estimated parameters and induce bias as well. We present a penalization technique based on the joint likelihood of the treatment and response variables to select the key covariates that need to be included in the treatment assignment model. Besides the bias induced by the non-randomization, we discuss another source of bias induced by having a non-representative sample of the target population. In particular, we study the effect of length-biased sampling in the estimation of the treatment effect. We introduced a weighted and a double robust estimating equations to adjust for the biased sampling and the non-randomization in the generalized accelerated failure time model setting. Large sample properties of the estimators are established. We conduct an extensive simulation studies to study the small sample properties of the estimators. In each Chapter, we apply our proposed technique on real data sets and compare the result with those obtained by other methods.

# ABRÉGÉ

L'ajustement du facteur de confusion est la clé dans l'estimation de l'effet de traitement dans les études observationelles. Deux techniques bien connus dajustement causal sont le score de propension et la probabilité de traitement inverse pondéré. Nous avons comparé les propriétés asymptotiques de ces deux estimateurs et avons démontré que la première méthode est un estimateur plus efficace. Étant donné que d'ignorer des facteurs de confusion importants ne fait que biaiser lestimateur, il semble bénéfique de tenir compte de tous les co-variables. Cependant, ceci peut entrainer une inflation de la variance des paramètres estimés et provoquer des biais également. Par conséquent, nous présentons une pénalisation technique basée conjointement sur la probabilité du traitement et sur les variables de la réponse pour sélectionner la clé co-variables qui doit être inclus dans le modèle du traitement attribué. Outre le biais introduit par la non-randomisation, nous discutons d'une autre source de biais introduit par un échantillon non représentatif de la population cible. Plus précisément, nous étudions l'effet de la longueur du biais de léchantillon dans l'estimation de la résultante du traitement. Nous avons introduit une pondération et une solide équation d'estimation double pour ajuster l'échantillonnage biaisé et la non-randomisation dans la généralisation du modéle à temps accéléré échec réglage. Puis, les propriétés des estimateurs du vaste échantillon sont établies. Nous menons une étude étendue pour examiner la simulation des propriétés des estimateurs du petit échantillon. Dans chaque chapitre, nous appliquons notre propre technique sur de véritables ensembles de données et comparons les résultats avec ceux obtenus par d'autres méthodes.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1
# Introduction

The most reliable statistical inference can be extracted from the data collected through a randomized experimental design. In the statistical theory of experimental design, dividing the experimental subjects to the treatment and control groups randomly is called randomization. Randomized experiments are designed to have balance between the treatment and control groups; that is, the covariates' distributions, measured or unmeasured, are similar in the treatment and control groups. For many years, most researchers have agreed that a randomized experimental design is the best method for drawing inference about parameters of interest. However, ethical standards can be violated using a randomized experiment. When the randomization is not possible, the experimenter can use a non-randomized experiment, a fact often practiced in observational studies. During the last two decades, observational studies have been subject to scrutiny to make the statistical inference as precise as possible. Estimation of the causal effect of treatment or exposure from observational data is prone to bias due to confounding of the treatment effect. Typically, in non-randomized experimental design, the treatment assignment mechanism is outside the control of the investigator. As a result, there is a potential for bias in the estimation of the treatment effect. This bias can be corrected by causal adjustment techniques under reasonable assumptions. Covariate subclassification is one of the methods that adjusts for the confounding by categorizing the population into several groups such that inside each group a coin is tossed to determine who receive the treatment. In other words, the subclassification technique divides an observational study into several randomized experiments. However, when we are dealing with a large number of covariates, covariate subclassification is cumbersome and often impractical. Rosenbaum & Rubin (1983) introduce a method to adjust for the difference between covariates' distributions in the treatment and control groups based on the correctly specified treatment assignment mechanism called the *propensity score*, a scalar function of the covariates. Specifically, let $D_i$ denote the treatment arm indicator for treated, $D_i = 1$, and control, $D_i = 0$ and $\mathbf{X}$ denote the $p$-dimensional vector of covariates, then the propensity score for binary treatment is given by

$$\pi(x) = \Pr(D = 1|x).$$

It has been shown that subclassifying just based on the fitted propensity score values guaranties the balancing property on the entire collection of observed covariates. There are some other causal estimation methods which are all adjusting for the potential presence of confounding variables in the collection of covariates $X$ but in different ways. However, the precise implementation details differ. Two methods of causal adjustment, *inverse probability of treatment weighting* (IPTW) and *propensity score* (PS) methods are commonly used. The two methods are constructed in a similar fashion; a model for treatment received is proposed and fitted, and then a regression model for the conditional expectation of the response variable is fitted either using weighting (IPTW) or matching/conditioning (PS). The IPTW method does not impose any modelling assumption on the response mean model to estimate the causal effect consistently. In this method, the fitted propensity score is used to weigh each observation and create a pseudo-population in which treatment assignment and covariates are independent (treatment assignment is randomized). Therefore, the crude difference of the treated and control outcomes in the pseudo-population will be a consistent estimate of the causal effect. Another causal adjustment method is *Double Robust* (DR) estimator which result in a consistent estimate if either the treatment assignment or the conditional mean response models is correctly specified. In this thesis, propensity score conditioning or propensity score regression (PSR) is often used. In the PSR approach, we include a fitted propensity score as a covariate in the response mean model. As such, the whole vector of covariates will be replaced by a scalar propensity score value.

As with all models for observational data, causal models require certain modelling assumptions to be appropriately specified (Robins (1997)). Specifically, throughout this thesis, we make the *stable unit treatment value assumption* (Rubin (2005)), which states that a subject's outcome is not influenced by other subjects' treatment allocation. We further assume *weak unconfoundedness*: for all treatment $d \in D$, the potential outcome $Y_i(d)$ and the treatment received $D$ are conditionally independent given the covariates $X$, $Y(d) \perp D|X$. It follows that $D$ and $Y(d)$ are conditionally independent given the propensity score, $Y(d) \perp D|\pi$

The theoretical properties of the PSR and IPTW adjustment procedures have been studied, but rarely directly compared. Hirano et al. (2003) shows that an estimator based on weighting by the reciprocal of the estimated propensity score is asymptotically equivalent to an efficient estimator that directly controls for all pretreatment variables, such as the estimator of Hahn (1998). On the other hand, Robins et al. (1992) show that the least-squares estimator based on regressing on the correctly specified propensity score can have variance no less than the semiparametric efficiency bound, but possibly larger. In Robins semiparametric setting the exposure can not interact with the vector of covariates and the model has to be an additive model.

In Chapter 2, we introduce a new semiparametric approach which generalizes the Robins setting in different ways. Our proposed model handles the interaction between the exposure and the covariates and does not need to be an additive model. Similar to the Robins estimator, our estimator can be used for either binary or continuous dose. We introduce an efficient influence function corresponding to the proposed semiparametric model using results from semiparametric estimation, and obtain a new efficiency bound based on the projection of estimating function onto the nuisance parameter space that is attained by a propensity score regression estimator. We study the performance - specifically, the bias, variance and MSE - of the proposed estimator for establishing the magnitude of a *direct* effect of treatment, that is, the *unconfounded* and *unmediated* effect on expected response. We show that our procedure produces an estimator with lower variance than IPTW and the DR methods. Our theoretical results are verified by extensive simulation studies. In simulation, we find that propensity score regression method seems to give estimators with smaller variance and lower mean square error. We also study the performance of the PSR and IPTW in longitudinal setting. Note that our focus is on direct effects, as this is the only setting in which IPTW and PSR can be readily compared, although IPTW adjustments also play a role in the estimation of other causal effects, such as total effect.

All of the results presented so far in the literature are based on the correct specification of the propensity score. Any model misspecification can result in an inconsistent estimator. Therefore, before going into any causal effect estimation procedure, we need to make sure that the key covariates have all been selected to model the propensity score. The conservative modelling strategy is to keep all the covariates in the propensity score model. However, adding covariates unrelated to the treatment can decrease the efficiency of covariates related to the treatment and therefore it can destroy the balancing property of the propensity score especially when the dimension of the unrelated covariates is high compared to the related ones. A simulation study (Brookhart et al. (2006a)) conjectures that variables unrelated to the treatment but related to the outcome should be always included in the propensity score model. The inclusion of these variables will decrease the variance of an estimated exposure effect without increasing bias. Their finding led us to find an optimal model selection strategy for the construction of propensity score models. Obviously, variable selection techniques based on prediction of the treatment will miss variables related only to the outcome and could miss important confounders that have a weak relation to the exposure but a strong relation to the outcome. Therefore, those variable selection techniques can result in efficiency loss in causal effect estimation. On the other hand, response covariate selection based strategies can be dominated by those covariates strongly related to the outcome and it may ignore confounders. The challenge here is that ignoring confounders which are strongly related to the treatment (outcome) and weakly related to the outcome (treatment) can induce bias.

In Chapter 3, we propose a novel penalized likelihood method to address the variable selection in the context of causal inference. The proposed variable selection method is based on penalizing the joint conditional likelihood of the outcome and treatment given the covariates. As such, each covariate has two chances to be kept in the model, either through the response or the treatment assignment model. We modify the tuning parameter of the penalty function such that it imposes heavier penalty on those covariates which are just related to the exposure. We show that under certain conditions, our proposed penalized likelihood satisfies the oracle properties, i.e. probability of choosing variables just related to the treatment tends to zero as the sample size goes to infinity, which is consistent with Brookhart et al. (2006a), and the estimators are asymptotically normally distributed. We, therefore, avoid variance inflation due to adding unrelated covariates to the outcome model through the fitted propensity score.

In addition to the bias induced by non-randomization, there is often another source of bias induced by having a non-representative sample from the target population. This issue is the subject of Chapter 4. Our study in Chapter 4 is motivated by a data set from the Canadian Study of Health and Aging (CSHA). CSHA is aiming to describe the epidemiology of dementia across Canada. Samples are randomly taken from either community or institution and the question of interest is to estimate the institualization effect on the survival time with dementia. Since in this particular study the chance of being in the sample is proportional to the survival time, the sample is not a representative sample of the target population. In the survival literature, this phenomenon is called biased-sampling. Biased sampling often exercised in observational studies on disease duration when recruiting incident cases is infeasible, often due to logistic constraints. Based on the accelerated failure time models, we introduce a weighted and a double robust (DR) estimating equations to estimate the causal effect consistently. Although the proposed estimating equations are estimating the same quantity, modelling assumptions are not the same. To obtain a consistent estimator using the weighted estimating equation (WEE) method, the propensity score model has to be correctly specified while the DR method results in a consistent estimator if either the propensity score or the failure time model is correctly specified. Moreover, in general, the DR estimating equation results in a more efficient estimator compared to the WEE. Our analysis reveals that estimating the effect of being institutionalized without considering these two sources of bias can change the treatment/grouping effect from being helpful to harmful. We establish the large sample properties of the estimators and study small sample behaviour of the estimators using simulations.

The remainder of this thesis is structured as follows: Chapter 2 introduces our semiparametric setting and the corresponding efficient influence function to estimate the causal effect. Theoretical results are followed by an extensive simulation studies and real data analysis to compare the performance of IPTW and PSR for binary and continuous treatments. Chapter 3 demonstrates our new model selection method in causal inference which is based on penalized likelihood regression methods. Chapter 4 introduces the concept of double bias which takes into account the biased-sampling as well as the non-randomization to estimate the treatment (grouping) effect. I conclude the thesis with the discussion of my results in Chapter 5.

# CHAPTER 2
# The Semiparametric Efficiency of Propensity Score Adjustment in the Estimation of Average Treatment Effects

## Chapter Summary

We consider the estimation of the total average effect of a dichotomous or continuously-valued treatment or exposure on outcome in the presence of measured confounding. For binary treatments, it is typical to adjust for differences between control and treatment groups using a scalar balancing quantity, the propensity score, which removes the bias induced by differences between these two groups of units. We examine optimality properties of propensity score-based adjustments. We utilize a semiparametric setting which does not impose any restriction on the functional form of the association between the response and the covariates. We assume a parametric model for the propensity score, and regard the parameters in this model as nuisance parameters. Using results from semiparametric inference, we construct an efficient influence function and estimator to estimate the total causal effect as the residual from projecting the score function of the parameters of interest onto the nuisance tangent space. We derive the semiparametric variance bound and demonstrate that it is lower than others previously obtained for propensity score methods. We illustrate how the bound for a competing approach, augmented inverse probability of treatment weighting, can be no lower than the propensity score bound. We then extend the result to continuous-valued treatments. All results are verified in simulation.

Establishing the causal effect of exposure or treatment $D$ in observational studies of response $Y$ is complicated because of the potential presence of confounding variables in the collection of covariates $X$. We consider the following simple setting identical to Hahn (1998). Let $D_i$ denote the treatment arm indicator for treated, $D_i = 1$, and control, $D_i = 0$. For a given subject, let $Y(1)$ denote outcome if treated, and $Y(0)$ outcome if untreated. Then the causal effect of treatment is $Y(1) - Y(0)$. However, in most cases, just one of the outcomes is observed for each subject. The parameters that have received lots of attention in the causal literature is the *average treatment effect* (ATE) $\mu = \mathbb{E}[Y_i(1) - Y_i(0)]$. One of the methods which yields an unbiased estimator for the causal effect is based on the *propensity score.* Rosenbaum & Rubin (1983) define the propensity score for binary treatment as

$$\pi(x) = \Pr(D = 1|x),$$

where $\pi(x)$ is a known function of covariates. For more details see Rubin (2008) and Rosenbaum (2010). Denote the corresponding random variable by $\pi = \Pr(D = 1|X)$, whose distribution depends on the distribution of $X$ and the precise model used to represent the treatment allocation model. In the usual experimental setting, we will have access to data $\{y_i, d_i, x_i, i = 1, \ldots, n\}$ that we may transform to data $\{y_i, d_i, \pi_i, i = 1, \ldots, n\}$, where $\pi_i \equiv \pi(x_i)$.

## 2.1 Introduction to Semiparametric Theory

Consider the statistical model where $V_1, ..., V_n$ are iid random vectors and the density of a single $V$ is assumed to belong to the class $\{p_V(v; \eta \in \Omega)\}$ with respect to some dominating measure $\upsilon_V$. The parameter $\eta$ can be partitioned in $(\beta, \alpha)$, where $\beta^{r_1 \times 1}$ is the parameter of interest and $\alpha$, the finite or infinite dimensional nuisance parameter. In this chapter, we only deal with the finite dimensional nuisance parameter, say of dimension $r_2$. Therefore, our parameter space has $r = r_1 + r_2$ dimension. In the causal inference setting, $\alpha$ corresponds to the parameters of the parametric propensity score model and $\beta$ are the parameters in the response model. Most of the estimators for $\beta$ are asymptotically linear; that is, there are a mean zero $r$-dimensional measurable function $\varphi(V)$, such that,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(V_i) + o_p(1), \tag{2.1}$$

where the last term in the above equation goes to zero in probability as $n \to \infty$ and $\mathbb{E}(\varphi\varphi') < \infty$.

In this thesis, we restrict ourselves to regular estimators; that is, estimators for which the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$ does not depend on the local data generating process. It has been shown by Hajek (1970) that most of the efficient estimators are asymptotically linear; hence, it is reasonable to restrict our setting to the regular and asymptotically linear (RAL) estimators.

We define the score vector, $S_\eta(v, \eta_0)$, for a single observation $V$ in a parametric model as follows

$$S_\eta(v, \eta_0) = \frac{\partial \log p_V(v, \eta)}{\partial \eta} |_{\eta = \eta_0},$$

it can be partitioned according to $\beta$ and $\alpha$, $S_\eta(v, \eta_0) = (S_\beta(v, \beta_0), S_\alpha(v, \alpha_0))$.

We define $\mu(\eta)$ as a smooth $q$-dimensional function of the $r$-dimensional parameter $\eta$. Tsiatis (2006) shows that if there exists an influence function for a regular asymptotically linear estimator $\hat{\mu}(\eta)$ such that $\mathbb{E}(\varphi\varphi') < \infty$, it will imply that

$$\mathbb{E}\{\varphi(V)S'_\eta(v,\eta_0)\} = \frac{\partial\mu(\eta)}{\partial\eta}. \tag{2.2}$$

Newey (1990) refers to (2.2) as an indication of the differentiability of $\mu(\eta)$, a $q$-dimensional function of parameters $\eta$. We will define $\mu(\eta)$ as an average causal effect and show that it satisfies (2.2).

### 2.1.1 Efficient Influence Function

The efficient influence function $\varphi_{\text{eff}}(V)$ is the influence function with smallest variance matrix; that is, for any influence function $\varphi(V)$, $\varphi(V) \neq \varphi_{\text{eff}}(V)$, $var\{\varphi_{\text{eff}}(V) - \varphi(V)\}$ is negative definite. In order to derive the efficient influence function, we need to define the nuisance tangent space. As a special case where $\eta$ can be partitioned as $(\beta, \alpha)$, using (2.2) it can be easily shown that $\mathbb{E}\{\varphi(V)S'_\alpha(v,\eta_0)\} = 0$; in other words, $\varphi(V)$ is an element of the Hilbert space orthogonal to the nuisance tangent space; that is, the linear subspace generated by the nuisance score vector, namely

$$\Lambda = \{B^{r_1\times r_2}S'_\alpha(v,\eta_0) \text{ for all } B^{r_1\times r_2}\}.$$

The efficient influence function can be derived as residual of projecting any arbitrary influence function $\varphi(V)$ onto the space orthogonal to the tangent space $(\mathcal{T})$, namely

$$\varphi_{\text{eff}}(V) = \varphi(V) - \prod(\varphi(V)|\mathcal{T}^\perp) = \prod(\varphi(V)|\mathcal{T})$$

where

$$\mathcal{T} = \{B^{r_1\times r_2}S'_\eta(v,\eta_0) \text{ for all } B^{r_1\times r_2}\}.$$

Therefore, $\varphi_{\text{eff}}(V)$ is an element of the tangent space and hence can be expressed as $\varphi_{\text{eff}}(V) = B^{q\times r}S'_{\alpha}(v,\eta_0)$. On the other hand it has to satisfy,

$$\mathbb{E}\{\varphi_{\text{eff}}(V)S'_{\eta}(v,\eta_0)\} = \frac{\partial\mu(\eta)}{\partial\eta},$$

thus

$$\varphi_{\text{eff}}(V) = \frac{\partial\mu(\eta)}{\partial\eta}I^{-1}(\eta_0)S'_{\eta}(v,\eta_0),$$

where $I(\eta) = \mathbb{E}\{S_{\eta}(v,\eta_0)S'_{\eta}(v,\eta_0)\}$ (see Tsiatis (2006)). When the parameter $\eta$ can be partitioned as $(\beta,\alpha)$, where $\beta$ is the parameter of interest and $\alpha$ is the nuisance parameter, then the efficient influence function can be written as

$$\varphi_{\text{eff}}(V) = \{\mathbb{E}(S_{\text{eff}}S'_{\text{eff}})\}^{-1}S_{\text{eff}}(V,\eta_0)$$

where

$$S_{\text{eff}}(V,\eta_0) = S_{\beta}(V,\eta_0) - \prod(S_{\beta}(V,\eta_0)|\Lambda)$$

and

$$\prod(S_{\beta}(V,\eta_0)|\Lambda) = \mathbb{E}(S_{\beta}S'_{\alpha})\{\mathbb{E}(S_{\alpha}S'_{\alpha})\}^{-1}S_{\alpha}(V,\eta_0).$$

The semiparametric efficiency bound obtained by $\varphi_{\text{eff}}(V)$ is given by

$$\mathbb{V} = \{\mathbb{E}(S_{\text{eff}}S'_{\text{eff}})\}^{-1}.$$

The variance bound can be rewritten in terms of the Fisher information elements as follows,

$$\mathbb{V} = \{I_{\beta\beta} - I_{\beta\alpha}I^{-1}_{\alpha\alpha}I'_{\beta\alpha}\}^{-1}$$

where

$$I = \begin{bmatrix} I_{\beta\beta} & I_{\beta\alpha} \\ I'_{\beta\alpha} & I_{\alpha\alpha} \end{bmatrix}.$$

In the rest of this chapter, we use the same argument to derive our efficient influence function corresponding to propensity score regression adjustment method. For further detail on semiparametric theory see Tsiatis (2006), Chamberlain (1992) and Newey (1994) and Newey (1990).

### 2.1.2 Balancing via the Propensity Score

Rosenbaum & Rubin (1983) demonstrate that $\pi$ is the coarsest function of covariates that has the *balancing* property; conditional on the propensity score, treatment assignment is independent of covariates, $D \perp X|\pi$, and $\pi$ is a function of any other balancing score. The ATE can be computed

$$\mu = \mathbb{E}[Y(1)-Y(0)] = \mathbb{E}_X[\mathbb{E}[Y(1)|X]-\mathbb{E}[Y(0)|X]] = \mathbb{E}_\pi[\mathbb{E}[Y(1)|\pi]-\mathbb{E}[Y(0)|\pi]] \quad (2.3)$$

where $\mathbb{E}_\pi$ denotes expectation with respect to the distribution of $\pi$ in the entire population. That is, under *strong ignorability* ($\{Y(\delta)\}_{\delta\in\mathcal{D}} \perp D \mid X$), or *weak unconfoundedness* ($Y(\delta) \perp D \mid X$, $\forall\, \delta \in \mathcal{D}$, see Imbens (2000)), subjects with the same value of propensity score but different treatments can be considered as a controls for each other, in the sense that the (conditional) expected difference in their responses equals the average treatment effect for that value of $\pi$.

In the classical method of *propensity score matching*, subjects in the study having the same propensity score are, by the balancing property, suitable for direct comparison in terms of response. In equation (2.3), the internal expectation can be estimated for any fixed $\pi = t$ from sample data by

$$\widehat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(\{1\},\{t\})}(D_i, \pi_i)Y_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(\{0\},\{t\})}(D_i, \pi_i)Y_i = \widehat{\mu}_1(t) - \widehat{\mu}_0(t)$$

say, where $\widehat{\mu}_j(t)$ is the estimator of the group-specific conditional mean at $\pi = t$, $\widehat{\mu}_j(t)$, and $\mathbb{I}_{(\{a\},\{b\})}(d, \pi) = 1$ if $d = a$ and $\pi = b$ and zero otherwise. The resulting estimator of $\mu$ is the average of $\widehat{\mu}(\pi)$ over the distribution of $\pi$; with $K$ strata of propensity score matched individuals at matching scores $t_1, \ldots, t_K$, the estimator is

$$\widehat{\mu} = \sum_{k=1}^{K} (\widehat{\mu}_1(t) - \widehat{\mu}_0(t))\widehat{f}_\pi(t_k) \tag{2.4}$$

where $\widehat{f}_\pi$ is the estimated distribution of $\pi$; in the case $f_\pi$ is known, it can replace $\widehat{f}_\pi$ in equation (2.4). We presume for the moment that $f_\pi$ is known.

Let $\beta$ denote a generic parameter utilized to parameterize the (conditional) distribution of response $Y$, either in the control or treatment groups. The average treatment effect, $\mu = \mu(\beta)$ is defined as

$$\begin{aligned}\mu(\beta) &= \int\int y f_1(y|\mathbf{x}, \beta) f(\mathbf{x})\, dy\, d\mathbf{x} - \int\int y f_0(y|\mathbf{x}, \beta) f(\mathbf{x})\, dy\, d\mathbf{x} \\ &= \int\int y f_1(y|\pi, \beta) f_\pi(\pi)\, dy\, d\pi - \int\int y f_0(y|\pi, \beta) f_\pi(\pi)\, dy\, d\pi \qquad (2.5)\end{aligned}$$

where $f_0$ and $f_1$ are the conditional densities of response in the untreated and treated groups respectively.

In this Chapter, we treat the parameters of the propensity score as the nuisance parameters, and using results from semiparametric inference, obtain the efficiency bound based on the projection of estimating function onto the nuisance parameter space that is defined by a propensity score-based estimator. We show that variance bound obtained from this method is smaller than those already presented in the literature for other propensity score-based methods.

### 2.1.3 Assumptions

As with all models for observational data, causal models require certain modelling assumptions to be appropriately specified (Robins (1997)). Throughout this paper, we make the standard assumptions. Specifically, we make the *stable unit treatment value assumption* (Rubin (2005)), which states that a subject's outcome is not influenced by other subjects' treatment allocation. We also assume *weak unconfoundedness* given the covariates $X$: it follows from this assumption that $D$ and $Y(d)$ are also conditionally independent given the propensity score $\pi$. Finally, for the binary treatment case, we make the *experimental treatment assignment* assumption that $0 < \pi(X) < 1$.

## 2.2 Variance Bounds for Propensity Score Methods

We focus on consistent estimators of $\mu(\beta)$ and their variance; in particular, as we discussed above, we consider *regular and asymptotically linear* (RAL) estimators that take the form

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Y_i)$$

which, under regularity conditions, are consistent and asymptotically normally distributed. Properties - in particular, the asymptotic variance - of RAL estimators are characterized by *influence function*, $\varphi(.)$. We work in a semiparametric context - parameterizing the conditional mean of $Y$, but leaving the conditional distribution specified nonparametrically - and consider the so-called *semiparametric variance* or *efficiency* bounds for such estimators. To develop the semiparametric theory, and to define the semiparametric variance bound, we typically start with a finite dimensional model, the *parametric submodel*, which contains the density that generates the data, and every density in this set belongs to the semiparametric model. See, for example, Newey (1990) for discussion. We begin by considering previous attempts to compute the variance bound for semiparametric estimation of a treatment effect.

### 2.2.1 The Variance Bound in Semiparametric Regression

In pioneering work, Robins et al. (1992) consider the following model

$$Y_i = \beta D_i + h(X_i) + \epsilon_i \qquad \mathbb{E}[\epsilon_i | D_i, X_i] = 0, \tag{2.6}$$

to examine the effect of dose $D$ on $Y$, where $h(X_i)$ is an unknown real-valued function of the confounders. They show that least-squares estimators of $\beta$ based on models for $h(X_i)$ will always be *at least as efficient* as any estimator of $\beta$ based on models for the propensity score. In other words, the least-squares estimator of the parameter $\beta$ based on regressing on the correctly specified propensity score can have variance no smaller than the semiparametric efficiency bound.

Robins et al. consider a parametric submodel $h(X; \alpha)$, such that $h(X; \alpha_0) = h_0(X)$ for some $\alpha_0$, and develop a two-stage estimator that achieves the semiparametric variance bound corresponding to the semiparametric model. The consistency of this two-stage estimator is based on the relatively restrictive assumption that the exposure does not interact with the vector of covariates.

The semiparametric variance bound for estimators of $\beta$ was shown by Chamberlain (1987) to equal

$$\left\{ \mathbb{E}_{D,X}\left[D^2/\sigma^2_{D,X}\right] - \mathbb{E}_X\left[\frac{\mathbb{E}_D\left[D/\sigma^2_{D,X}|X\right]^2}{\mathbb{E}_D\left[1/\sigma^2_{D,X}\right]}\right]\right\}^{-1}$$

where $\sigma^2_{D,X} = \text{Var}[Y|D,X]$, which under homoscedasticity, with $\sigma^2 = \text{Var}[Y|D,X]$ independent of $D$ and $X$, reduces to

$$\left\{\sigma^{-2}\mathbb{E}_X\left[\mathbb{E}_{D|X}\left[D^2|X\right]\right] - \sigma^{-2}\mathbb{E}_X\left[\mathbb{E}_D\left[D|X\right]^2\right]\right\}^{-1} = \frac{\sigma^2}{\mathbb{E}_X[\text{Var}[D|X]]}$$

See Robins et al. (1992) for further discussion and illustrations.

### 2.2.2 Previous Bounds for Propensity Score Models

In considering causal adjustment methods based on the propensity score, Hahn (1998) and Heckman et al. (1998) show that matching based on the *known* propensity score (that is, where the treatment assignment model is presumed known, and no parameters are estimated) can result in efficiency loss compared to adjusting for all pre-treatment variables. Hirano et al. (2003) introduce the estimator based on the *estimated* propensity score, and demonstrate that it is fully efficient for estimation of average treatment effects. They show that their semiparametric estimator achieves the semiparametric efficiency bound obtained in Hahn (1998).

The likelihood utilized by Hahn (1998) is constructed by considering the original data, that is, the random sample $\{y_i, d_i, x_i,\ i = 1, \ldots, n\}$. We study this likelihood here: based on a specific parametric submodel, the likelihood is

$$\mathcal{L}_H(\eta;\mathbf{y},\mathbf{d},\mathbf{x}) = \prod_{i=1}^{n} \{f_1(y_i|x_i,\beta)p(d|x_i,\alpha)f(x_i|\alpha)\}^{d_i} \\ \{f_0(y_i|x_i,\beta)(1-p(d|x_i,\alpha))f(x_i|\alpha)\}^{1-d_i}. \qquad (2.7)$$

where $\alpha$ is an $r_2$-dimensional vector containing parameters that appear in the model for $D|X$ and $X$ respectively. Thus $\eta = (\beta, \alpha)$ is an $r$-dimensional vector, $r = r_1 + r_2$. Using this parametric submodel, Hahn (1998) deduces the variance bound on estimators of the ATE $\mu$ to be

$$\mathbb{E}_X\left[\frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1-\pi(X)} + (\mu(X)-\mu)^2\right] \qquad (2.8)$$

where the expectation is over the distribution of $X$, for $j = 0, 1$ $\sigma_j^2(X) = \mathbb{V}\text{ar}[Y(j)|X]$, and

$$\mu(X) = \mathbb{E}[Y(1) - Y(0)|X].$$

In the homoscedastic case, where $\sigma_0^2 = \sigma_1^2 = \sigma^2$ say, the bound reduces to

$$\mathbb{E}_X\left[\frac{\sigma^2}{\pi(X)(1-\pi(X))} + (\mu(X)-\mu)^2\right]$$

which can be written as the sum

$$\mathbb{E}_\pi\left[\frac{\sigma^2}{\pi(1-\pi)}\right] + \mathbb{E}_X\left[(\mu(X)-\mu)^2\right]$$

where both terms are non-negative.

## 2.3 The Efficiency of the Propensity Score Approach

### 2.3.1 The proposed semiparametric setting

The exposure effect can be estimated via different semiparametric models and the optimum choice of the semiparametric model depends on the existing information about the different components of the model. For example, suppose we do not know the functional form of the association between the covariates, confounders, with the response but we know that the true conditional mean model is additive with respect to the exposure and the unknown function. Then Robins et al. (1992) show that the optimum choice of the semiparametric model is given by equation (2.6). This additive semiparametric model was introduced originally by Engle et al. (1986) and consists of a parametric component $\beta$ and a nonparametric component $h(X)$. As indicated above, we can assume a parametric submodel $h(X;\alpha)$, such that $h(X;\alpha_0) = h_0(X)$ for some $\alpha_0$ where $h_0(X)$ is the true function. Note that the propensity score is a reasonable candidate for parametric submodel for the $h(.)$. Hahn (1998) proposes another semiparametric estimator which does not depend on any modelling assumption for the response, constructing a complete data likelihood using nonparametric imputation, and shows that this estimator achieves the semiparametric bound which was introduced in Begun et al. (1983) and Bickel et al. (1993a).

We consider a generalized version of the semiparametric model used by Robins et al. (1992) by assuming the following possibly non-linear, non-additive model

$$Y_i = \mu(D_i, h(X_i,\alpha), \beta) + \epsilon_i, \qquad (2.9)$$

where $\mu(.)$ is an unknown function of the exposure and the known parametric model $h(X_i;\alpha)$ and the disturbance is a zero mean random variable. This model does not impose any restriction, for example, on having interaction between the exposure and the covariates. In general, the function $\mu(.)$ is unknown and investigators need to perform some model-checking to examine the adequacy of the different possible candidates. Little & An (2004) and Robinson (1988) incorporate nonparametric regression to model the conditional mean response given the fitted propensity score. Huber et al. (1981) introduce a regression model in which the number of the confounders can increase by the sample size.

To obtain a consistent estimator of the exposure effect, we assume a parametric submodel $\mu(D_i, h(X_i;\alpha), \beta)$; depending on being treated or untreated, we write $\mu_1(h(X_i,\alpha),\beta)$ or $\mu_0(h(X_i;\alpha),\beta)$, respectively. We replace the $h(X_i;\alpha)$ by $\pi(X_i;\alpha)$ and construct our proposed semiparametric bound based on the known parametric model $\pi(.)$. In the case of binary exposure, the propensity score can be fitted using a logistic regression,

$$\Pr(D = 1|X) = \frac{\exp\{\alpha' X\}}{1 + \exp\{\alpha' X\}}.$$

We will show that when assuming additivity as in equation (2.6), our established variance bound is equivalent by the one obtained to Robins et al. (1992) and Chamberlain (1992), and lower than the one obtained by Hahn (1998).

### 2.3.2 Efficient Semiparametric Inference for Propensity Score Models

In order to guarantee that the semiparametric efficiency/variance bound is well-defined, we need to check the differentiability of the parameter of interest in our parametric submodel (see, for example Tsiatis (2006, Chap. 4)). This assumption is usually assumed to hold implicitly; in Appendix 5.1, we verify this assumption for the average treatment effect.

Consider an alternative likelihood formulation to equation (3.2), where the $x_i$ are replaced by $\pi_i$, so that the data are $\{d_i, y_i, \pi_i, i = 1, \ldots, n\}$. Consider the following parametric likelihood for these (transformed) data:

$$\mathcal{L}_{PS}(\eta; \mathbf{y}, \mathbf{d}, \pi) = \prod_{i=1}^{n} \{f_1(y_i|\pi_i, \beta) p(d_i|\pi, \alpha) f_\pi(\pi_i|\alpha)\}^{d_i} \tag{2.10}$$

$$\{f_0(y_i|\pi_i, \beta) p(d_i|\pi, \alpha) f_\pi(\pi_i|\alpha)\}^{(1-d_i)}. \tag{2.11}$$

In this equation, suppose that the form of $\pi \equiv \pi(X)$ is known up to a parametric model whose parameters are to be estimated from the data, that is, we assume that $\pi$ and $f_\pi$ are functions of a finite dimensional parameter vector, and collate those parameters in $\alpha$.

We study the additive error case. Let $\epsilon_{ij} = y_i - \mu_j(\pi_i, \beta)$ for $j = 0, 1$ and each $i$, where

$$\mu_j(\pi, \beta) = \mathbb{E}[Y(j)|\pi, \beta]$$

is a scalar function of $\beta$, $\pi$ and possibly other covariates. We can rewrite $\mathcal{L}_{PS}$ in terms of $\epsilon$ and the propensity score, $\pi$, as follows:

$$\prod_{i=1}^{n} \{f_1(\epsilon_{i1}|\beta) p(d_i|\pi_i, \alpha) f_\pi(\pi_i|\alpha)\}^{d_i} \{f_0(\epsilon_{i0}|\beta) p(d_i|\pi_i, \alpha) f_\pi(\pi_i|\alpha)\}^{(1-d_i)}$$

$$= \prod_{i=1}^{n} \{f_1(y_i - \mu_1(\pi_i, \beta)|\beta) p(d_i|\pi_i, \alpha) f_\pi(\pi_i|\alpha)\}^{d_i}$$

$$\{f_0(y_i - \mu_0(\pi_i, \beta)|\beta) p(d_i|\pi_i, \alpha) f_\pi(\pi_i|\alpha)\}^{(1-d_i)}$$

In this formulation, $\beta$ are the parameters of interest, and $\alpha$ are the nuisance parameters that parameterize $f_0, f_1$ and $f_\pi$.

Using subscripts to denote partial derivatives with respect to $\beta$ and $\alpha$ (evaluated at the true parameters), and dropping the dependence on $i$ for convenience, the score functions corresponding to this likelihood are

$$S_\beta = dS_{1\beta}(\epsilon_1|\pi,\beta)\mu_{1\beta}(\pi,\beta) + (1-d)S_{0\beta}(\epsilon_0|\pi,\beta)\mu_{0\beta}(\pi,\beta)$$

$$S_\alpha = dK_1(\epsilon_1) + (1-d)K_0(\epsilon_0) + A(d,\pi)K_2(d|\pi) + K_3(\widehat{\pi})$$

where

$$A(d,\pi) = \frac{p(d=1|\pi)(d-p(d=1|\pi))}{p(d=1|\pi)p(d=0|\pi)} = \frac{\pi(d-\pi)}{\pi(1-\pi)},$$

and

$$K_j(\epsilon_j) = \frac{f_{j\eta}(\epsilon_j|\alpha)}{f(\epsilon_j|\alpha)} \qquad K_2(d|\pi) = p_\alpha(d|\pi,\alpha) \qquad K_3(\pi) = \frac{f_\alpha(\pi|\alpha)}{f(\pi|\alpha)}$$

are all $k \times 1$ vector functions of a scalar argument, and where the double subscripts denote partial derivatives evaluated at the true parameter for each $j$.

### 2.3.3 The efficient score function and influence function

In order to construct the efficient score in this nuisance parameter setting, by standard geometric arguments (see Tsiatis (2006)), consider the projection of the score function for parameters of interest onto the nuisance parameter tangent space, that is, the efficient score function lies perpendicular to the nuisance tangent space. In the model described above, the nuisance tangent space, denoted $\mathcal{T}$, is constructed by inspecting $S_\alpha$, that is

$$\mathcal{T} = \{dK_1(\epsilon_1) + (1-d)K_0(\epsilon_0) + a(\pi)K_2(d|\pi) + K_3(\pi)\} \tag{2.12}$$

that is, the sum of four orthogonal components, where $a(\pi)$ is any square integrable measurable function of $\pi$. By ancillarity, the projection of $S_{j\beta}$ on $K_2(d|\pi)$, $K_3(\pi)$ and $\epsilon_{j'}$ for $j \neq j'$ is zero. The efficient score function, which is the residual after projecting the score function onto the nuisance tangent space for $\beta$, is given by

$$S_{\text{eff}} = S_{1\beta} - \mathbb{E}[S_{1\beta}|\epsilon_1] + S_{0\beta} - \mathbb{E}[S_{0\beta}|\epsilon_0] \tag{2.13}$$

From Newey (1990) and Tsiatis (2006),the space orthogonal to the nuisance tangent space is $\{g(X)\epsilon\}$ where $g(X)$ is a $(k \times 1)$ vector of arbitrary functions of $X$ and $\epsilon = d\epsilon_1 + (1-d)\epsilon_0$. Since the $\mathcal{T}$ is a linear space, the projection (in the Hilbert space with the covariance inner product) of $S_\beta$ onto $\mathcal{T}$ exists, and the unique $g(X)$ satisfies $\mathbb{E}[(S_\beta - g(X)\epsilon)\epsilon|X] = 0$ (see Newey (1990)), which yields

$$g(X) = \frac{\mathbb{E}[S_\beta \epsilon|X]}{\mathbb{E}[\epsilon^2|X]}$$

(by assumption, the conditional expectation $\mathbb{E}[\epsilon^2|X]$ is strictly positive). In the Appendix 5.2, we show that

$$\mathbb{E}[S_\beta \epsilon|\pi] = \frac{\partial \mu^*(\pi,\beta)}{\partial \beta} = \mu^*_\beta(\pi,\beta) \qquad \text{where} \qquad \mu^*(\pi,\beta) = d\mu_1(\pi,\beta) + (1-d)\mu_0(\pi,\beta).$$

Therefore, substituting into equation (2.13) yields

$$S_{\text{eff}} = \frac{\mu^*_\beta(\pi,\beta)}{\mathbb{E}[\epsilon^2|\pi]}\epsilon = \frac{d\epsilon_1\mu_{1\beta}(\pi,\beta) + (1-d)\epsilon_0\mu_{0\beta}(\pi,\beta)}{\pi\sigma_1^2(\pi) + (1-\pi)\sigma_0^2(\pi)}.$$

Following standard results (Tsiatis, 2006)), the efficient influence function for $\mu(\beta)$ is

$$\varphi_{\text{eff}} = \mu'_\beta(\beta)\mathbb{E}[S_{\text{eff}}S'_{\text{eff}}]^{-1}S_{\text{eff}}$$

We can expand this expression as follows

$$\varphi_{\text{eff}} = \mu'_\beta(\beta)BV^{-1}\{d\mu_{1\beta}(\pi,\beta)(y-\mu_1(\pi,\beta)) + (1-d)\mu_{0\beta}(\pi,\beta)(y-\mu_0(\pi,\beta))\} \tag{2.14}$$

where

$$V = \mathbb{E}[\epsilon^2|\pi] = \pi\sigma_1^2(\pi)+(1-\pi)\sigma_0^2(\pi) \quad W = W(\pi) = \mu^*_\beta(\pi,\beta) \quad B = \mathbb{E}[W(\pi)V^{-1}W(\pi)']^{-1}.$$

Again by standard results from semiparametric inference, the semiparametric variance bound is given by

$$\begin{aligned}
\mathbb{V}_{PS} &= \mu'_\beta(\beta)\mathbb{E}[S_{\text{eff}}S'_{\text{eff}}]^{-1}\mu_\beta(\beta) \\
&= \mu'_\beta(\beta)\mathbb{E}[W(\pi)V^{-1}W(\pi)']^{-1}\mu_\beta(\beta) \\
&= \mu'_\beta(\beta)\mathbb{E}\left[\frac{\pi\sigma_1^2(\pi)\mu_{1\beta}(\pi,\beta)\mu_{1\beta}(\pi,\beta)' + (1-\pi)\sigma_0^2(\pi)\mu_{0\beta}(\pi,\beta)\mu_{0\beta}(\pi,\beta)'}{[\pi\sigma_1^2(\pi)+(1-\pi)\sigma_0^2(\pi)]^2}\right]^{-1}\mu_\beta(\beta)
\end{aligned} \tag{2.15}$$

Inspection of (2.14) yields

$$\begin{aligned}
\varphi_{\text{eff}} &= \mu'_\beta(\beta)\mathbb{E}[S_{\text{eff}}S'_{\text{eff}}]^{-1}S_{\text{eff}} \\
&= \mu'_\beta(\beta)I^{-1}\frac{\mu^*_\beta(\pi,\beta)\epsilon}{\mathbb{E}[\epsilon^2|\pi]} \\
&= \frac{I^{-1}}{\mathbb{E}[\epsilon^2|\pi]}(d\mu_{1\beta}(\pi,\beta)(y-\mu_1(\pi,\beta))+(1-d)\mu_{0\beta}(\pi,\beta)(y-\mu_0(\pi,\beta))) \qquad (2.16)
\end{aligned}$$

where

$$I = \mathbb{E}\left[\frac{\pi\sigma_1^2(\pi)\mu_{1\beta}(\pi,\beta)\mu_{1\beta}(\pi,\beta)'+(1-\pi)\sigma_0^2(\pi)\mu_{0\beta}(\pi,\beta)\mu_{0\beta}(\pi,\beta)'}{[\pi\sigma_1^2(\pi)+(1-\pi)\sigma_0^2(\pi)]^2}\right].$$

We refer to the estimator obtained using this influence function as a *Propensity Score Regression* (PSR) estimator of the causal effect. The next theorem summarizes the asymptotic behaviour of the proposed semiparametric estimator; it gathers together the results from the previous two subsections:

**Theorem 2.1** *The propensity score regression estimator of the causal effect $\mu(\beta)$, estimated using the efficient influence function (2.16) has the following asymptotic properties*

$$\sqrt{n}(\mu(\hat{\beta})-\mu(\beta)) \sim \mathcal{N}(0,\mathbb{V}_{PS})$$

*where $\mathbb{V}_{PS}$ is given in (2.15).*

**Proof** The parameter $\mu(\beta)$ is a differentiable parameter in the sense of Newey (1990); the estimator is RAL with influence function given by (2.14). Therefore, by results from Newey (1990) (see also Tsiatis (2006)), the estimator is asymptotically normal with asymptotic variance $\mathbb{V}_{PS}$.

### 2.3.4 Example: Propensity Score Regression

Here, we assume that the semiparametric mean model has an additive form similar to (2.6) and show that by imposing the additivity our bound will be equivalent to Chamberlain (1992). Suppose the true conditional mean model is $\mathbb{E}[Y|D, X] = \beta_1 d + \beta_2 X$, where $X$ is a continuous covariate, and assume that $\sigma_1^2(\pi) = \sigma_0^2(\pi) = \sigma^2$. Using the propensity score regression approach, we have

$$\mu_1(\pi, \beta) = \beta_1 + \beta_2\pi \qquad\qquad \mu_0(\pi, \beta) = \beta_2\pi$$

and $\mu(\beta) = \beta_1$. Further, we have

$$\mu^\star(\pi, \beta) = d\mu_1(\pi, \beta) + (1 - d)\mu_0(\pi, \beta) = d\beta_1 + \pi\beta_2$$

so that

$$\mu^\star_{\beta_1}(\pi, \beta) = d \qquad \mu^\star_{\beta_2}(\pi, \beta) = \pi \quad W(\pi) = \begin{bmatrix} d \\ \pi \end{bmatrix}$$

and $\mathbb{E}[WV^{-1}W']^{-1} = \sigma^2\mathbb{E}[WW']^{-1}$. Now,

$$[WW'] = \begin{bmatrix} D^2 & D\pi \\ D\pi & \pi^2 \end{bmatrix}$$

so conditional on $\pi$, taking conditional expectations,

$$\mathbb{E}_{D|\pi}[WW'] = \begin{bmatrix} \pi & \pi^2 \\ \pi^2 & \pi^2 \end{bmatrix}$$

and taking expectations wrt $\pi$, setting $E = \mathbb{E}[\pi], E_2 = \mathbb{E}[\pi^2]$.

$$\mathbb{E}[WW'] = \begin{bmatrix} E & E_2 \\ E_2 & E_2 \end{bmatrix} \qquad \Longrightarrow \qquad \mathbb{E}[WW']^{-1} = \frac{1}{EE_2 - E_2^2} \begin{bmatrix} E_2 & -E_2 \\ -E_2 & E \end{bmatrix}$$

and therefore the variance bound for $\theta_1$ is the upper left entry in this matrix, that is

$$\frac{\sigma^2}{(\mathbb{E}[\pi] - \mathbb{E}[\pi^2])} = \frac{\sigma^2}{\mathbb{E}[\pi(1-\pi)]}$$

Under the linearity and homoscedasticity assumptions the variance bound of propensity score regression is $\sigma^2/\mathbb{E}[\pi(1-\pi)]$. This is identical to the semiparametric variance bound in Chamberlain (1992), and the form

$$\frac{\sigma^2}{\mathbb{E}_X[\text{Var}[D|X]]}$$

given by Robins et al. (1992). In Appendix 5.3, we demonstrate that this results also holds for score *stratification*, another common propensity score adjustment method.

### 2.4 The Semiparametric Bound for Inverse Probability Weighting

We now demonstrate that propensity score regression and stratification are superior, in terms of asymptotic variance, compared to another common causal adjustment procedure based on the propensity score construction.

### 2.4.1 Inverse Probability of Treatment Weighting

Inverse Probability of Treatment Weighting (IPTW) is a widely used approach to causal adjustment. For estimating the ATE of a binary treatment, the estimator

$$\widehat{\mu}_{IPTW} = \frac{\left(\sum_{i=1}^{n} \frac{D_i Y_i}{\pi_i}\right)}{\left(\sum_{i=1}^{n} \frac{D_i}{\pi_i}\right)} - \frac{\left(\sum_{i=1}^{n} \frac{(1-D_i) Y_i}{1-\pi_i}\right)}{\left(\sum_{i=1}^{n} \frac{1-D_i}{1-\pi_i}\right)} = \widehat{\mu}_{IPTW}(1) - \widehat{\mu}_{IPTW}(0) \qquad (2.17)$$

is commonly used. It is the difference of weighted means in the treated and non treated subjects, with weights proportional to $\Pr(D_i = d_i | X_i)$ for the observed $d_i$.

In the case of *coarsened* data, the class of weighted (full data) influence functions represented by (2.17) is

$$\mathcal{G} = \left\{ \frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)} - \mu \right\}$$

where $\mu$ is the average treatment effect. The weighted full data influence function includes fully observed individuals, and its efficiency can be improved by contributing individuals with some missing data in an augmentation term. As shown by Tsiatis (2006), this efficient influence function is equal to

$$\mathcal{G}_{\text{eff}} = \left\{ \frac{DY}{\pi(X)} - \frac{(D-\pi(X))\mu(1,X)}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)} - \frac{(D-\pi(X))\mu(0,X)}{1-\pi(X)} - \mu \right\}$$

or equivalently

$$\mathcal{G}_{\text{eff}} = \left\{ \frac{D(Y-\mu(1,X))}{\pi(X)} - \frac{(1-D)(Y-\mu(0,X))}{1-\pi(X)} + (\mu(1,X) - \mu(0,X)) - \mu \right\}$$

where $\mu(j, X) = \mathbb{E}[Y|D = j, X]$ for $j = 0, 1$. The estimator corresponding to this influence function is called an *Augmented Inverse Probability Weighed Complete Case* (AIPWCC) estimator. The asymptotic variance of this estimator is the variance of the efficient influence function. The asymptotic variance of the AIPWCC estimator is given by

$$\mathbb{E}[\varphi_{\text{eff}}^2] = \mathbb{E}\left\{\frac{(Y - \mu(1, X))^2}{\pi(X)} + \frac{(Y - \mu(0, X))^2}{1 - \pi(X)} + (\mu(1, X) - \mu(0, X) - \mu)^2\right\}$$

which is equivalent to the variance bound introduced by Hirano et al. (2003) and can be estimated using a sandwich estimator given in Tsiatis (2006, Chap. 9). The AIPWCC estimator is referred to as a *doubly robust* (DR) estimator in the sense that it is consistent if either the treatment mechanism or the response mean models are correctly specified. This estimator depends on the unknown response mean model, $\mathbb{E}[Y|D = 1, X]$, which can be estimated by positing a parametric model. If the posited mean model is the correct model, the AIPWCC results in an efficient estimator; it will lead to a "locally efficient" estimator if the posited model is misspecified.

Now, we want to show that the IPTW estimator (2.17) is the efficient estimator in the nonparametric setting: that is, when there is no restriction on the class of density functions. Let $(Y_i, D_i, X_i)$ for $i = 1, ..., n$ be independent and identically distributed random vectors with a joint density $f(Y_i, D_i, X_i)$. By Tsiatis (2006, Thm. 4.4), the tangent space for nonparametric model is the entire Hilbert space. The space orthogonal to the tangent space, the nuisance tangent space, is empty. Therefore, for this model, the class of influence functions consists of just one element and this influence function must be the efficient influence function. Let function $\xi_{IP}(D_i, Y_i, X_i)$ be defined by

$$\xi_{IP}(d, y, x) = \frac{d}{t}(y - \mu_1(x)) - \frac{1-d}{1-t}(y - \mu_0(x)) \tag{2.18}$$

and consider the following score function corresponding to any parametric submodel $f(Y_i, D_i, X_i; \beta)$

$$S_\beta(d, y, t|\beta) = dS_{1\beta}(y|x, \beta) + (1 - d)S_{0\beta}(y|x, \beta). \tag{2.19}$$

It follows that

$$\mathbb{E}[\xi_{IP}(D, Y, X)s(D, Y, X|\beta)] = \frac{\partial\mu(\beta)}{\partial\beta} \tag{2.20}$$

and hence $\mu(\beta)$ is a differentiable parameter. Since $\xi_{IP}(D_i, Y_i, X_i)$ satisfies (2.20), following Newey (1990), it is an influence function. By uniqueness of the influence function for nonparametric models, $\xi_{IP}(D_i, Y_i, X_i)$ must be the efficient influence function. Therefore, the semiparametric efficiency bound is

$$\begin{aligned}\mathbb{E}[\xi^2_{IP}(D, Y, X)] &= \mathbb{E}\left[\mathbb{E}\left\{\frac{D}{\pi(X)^2}(Y - \mu_1(X))^2 + \frac{1-D}{(1-\pi(X))^2}(Y - \mu_0(X))^2|X\right\}\right] \\ &= \mathbb{E}\left[\frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1-\pi(X)}\right] \end{aligned} \tag{2.21}$$

which is the same as Hahn (1998) when the covariate distribution is known. The variance bound introduced by Hirano et al. (2003) is equivalent to (2.21) if treatment does not interact with covariates and the response mean model is linear.

The propensity score variance bound in (2.15) is lower than the bound in equation (2.21). From equation (2.8), in the case of homoscedasticity, we have

$$\mathbb{E}\left[\frac{\sigma^2}{\pi(X)(1-\pi(X))}+(\mu(X)-\mu)^2\right] \geq \sigma^2\mathbb{E}_\pi\left[\frac{1}{\pi(1-\pi)}\right] \geq \frac{\sigma^2}{\mathbb{E}_\pi[\pi(1-\pi)]}$$

by Jensen's inequality. Because it can be easily shown by the law of total variance that if all confounders are integrated out first, and then the variance bound computed, the resulting bound will be bigger than that obtained by first finding the conditional variance bound, and then integrating out the confounders. The Hahn (1998) variance bound is valid for inverse probability of treatment weighting, because in this method outcomes are regressed on the received treatment using weighted regression therefore there is no need to integrate out the confounders before or after obtaining the variance bound.

## 2.5 Continuous Treatments

### 2.5.1 Generalized Propensity Score

In this section we define the Generalized Propensity Score which is the generalization of the classical binary treatment propensity score. We first examine the single interval case. When treatment is a continuous random variable, it is possible to construct a balancing score using an approach based on the *Generalized Propensity Score* (GPS). Following Imbens (2000) and Hirano & Imbens (2004), we define the (observed) GPS, $\pi(d, x)$ for dose $d$ and covariate $x$ by

$$\pi(d, x) = f_{D|X}(d|x) \tag{2.22}$$

that is, the conditional mass/density function for $D$ given $X = x$ evaluated at $D = d$. Additionally $\pi(d, X)$ and $\pi(D, X)$ are corresponding random quantities. It has been shown by Hirano & Imbens (2004) that GPS random quantity $\pi(d, X)$ acts as a balancing score, in that $D$ and $X$ are conditionally independent given $\pi(d, X)$. Secondly, for any $d$, the allocation of the treatment dose is conditionally independent of the potential response, given the propensity score, $Y(d) \perp D \mid \pi(d, X)$, that is, we have unconfoundedness of $Y(d)$ and $D$ given $\pi(d, X)$. Therefore $\pi(d, X)$ breaks the dependence between $D$ and $X$, and hence the causal effect of $D$ on $X$ can be estimated by conditioning on $\pi(d, X)$ for each $d$ in turn, and then averaging over the distribution of $\pi(d, X)$. The role of the GPS in estimating the *Average Potential Outcome* (APO) is made clear by identity given in Imbens (2000)

$$\mu(d) = \mathbb{E}[Y(d)] = \mathbb{E}_X[\mathbb{E}[Y(d)|\pi(d, X)]] = \mathbb{E}_{\pi(d,X)}[\mathbb{E}[Y(d)|\pi(d, X)]]$$

We used the same algorithm to estimate the APO as Moodie & Stephens (2010) here:

I **Form the GPS Model:** Using the regression approach, construct the propensity model for $D$ given $X$, $\pi(d, x) = f_{D|X}(d|x, \alpha)$. Estimate parameters $\alpha$ using data $\{(d_i, x_i), i = 1, \ldots, n\}$.

II **Compute the Fitted GPS Model:** Compute the estimated GPS,

$$\widehat{\pi}_i = f_{D|X}(d_i|x_i, \widehat{\alpha}).$$

- **III Form the Observable Model:** Using the regression approach, construct a predictive model for the conditional expectation density $f_{Y|D,\pi(d,X)}(y|d,\pi(d,x),\beta)$. Estimate parameters $\beta$ using data $\{(y_i, d_i, \widehat{\pi}_i), i = 1, \ldots, n\}$.
- **IV Estimate the APO:** Estimate the APO for each $d$ by

$$\widehat{\mu}(d) = \widehat{\mathbb{E}}[Y(d)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y|D,\pi(d,X)}[Y_i(d)|d, \widehat{\pi}_i, \widehat{\beta}]$$

then $\widehat{\mu}(d)$ is the GPS-adjusted estimated dose-response function.

An alternative approach proposed by Hirano & Imbens (2004) suggests that the APO may be approximated by estimating the dose-response effect within strata defined by the linear predictor of the treatment density function, and then combining these estimates to form a single, weighted average. This approach is straightforward to implement and often provides an estimate of the dose-response relationship that has little or no residual bias, although it may be less efficient than the regression approach described above.

We are interested in deriving a semiparametric variance bound for average potential outcome estimated using the GPS. The joint likelihood function of $(y_i, d_i, \pi)$ is given by

$$\prod_{i=1}^{n} \{f_{d_i}(y_i|\pi,\beta) f(d|\pi,\alpha) f(\pi_i|\alpha)\}^{\mathbb{I}_d(d_i)}$$

Using subscripts to denote partial derivatives with respect to $\theta$ and $\eta$, the score functions corresponding to this likelihood are

$$S_\beta = S_{d\beta}(y|\pi,\beta)\mu_{d\beta}(\pi,\beta) \qquad\qquad S_\alpha = K_1(\epsilon_d) + K_2(d|\pi) + K_3(\pi)$$

where

$$K_1(\epsilon_d) = \frac{f_{d\alpha}(y - \mu_d(\pi, \beta)|\alpha)}{f_d(y - \mu_d(\pi, \beta)|\alpha)} \qquad K_2(d|\pi) = \frac{f_\alpha(d|\pi, \alpha)}{f(d|\pi, \alpha)} \qquad K_3(\pi) = \frac{f_\alpha(\pi|\alpha)}{f(\pi|\alpha)}$$

and $\epsilon_d = y_i - \mu_d(\pi, \beta)$. By ancillarity, the projection of $S_\beta$ on $K_2(d|\pi)$, $K_3(\pi)$ and $\epsilon_{j'}$ for $j \neq j'$ is zero. Thus, the efficient score function is given by

$$S_{\text{eff}} = S_\beta - \mathbb{E}[S_\beta|\epsilon_d]$$

By simple adaptation to the binary treatment case, we can show that

$$S_{\text{eff}} = \frac{\epsilon_d \mu_{d\beta}(\pi, \beta)}{\sigma_d^2(\pi)}. \tag{2.23}$$

Thus, the semiparametric variance bound in continuous treatment case is given by

$$\mathbb{V}_{GPS} = \mu_\beta(d, \beta)' \mathbb{E}\left[\frac{\mu_{d\beta}(\pi, \beta)\mu_{d\beta}(\pi, \beta)'}{\sigma^2(\pi)}\right]^{-1} \mu_\theta(d, \beta), \tag{2.24}$$

where $\mu(d, \beta)$ is a parametrized APO.

### 2.5.2 Inverse Probability Treatment of Weighting

For a continuous treatment one can still obtain the unbiased estimates of the causal parameter via IPTW by fitting the regression model for $D$ given $X = x$ to obtain stabilized weights

$$w_i^s \propto \frac{f(d_i)}{f(d_i|x_i)}$$

where, for doses on $\mathbb{R}$, $f(d_i|x_i)$ can be estimated by using a linear regression model to yield

$$\widehat{f}(d|x) = (2\pi\widehat{\sigma}^2)^{-1/2} \exp(-(d - (\widehat{\alpha}_0 + \widehat{\alpha}_1 x))^2/(2\widehat{\sigma}^2)).$$

To estimate the numerator $f(d)$, one might specify normal density with the average of observed $D$ and empirical variance as a mean and variance of the density.

Analogous to the established semiparametric bound for nonparametric models in the binary treatment case, the variance bound for continuous treatment case can be obtained as follows. Let functions $\epsilon_d(D_i, Y_i, X)$ and $S_\beta(d, y, X|\beta)$ be defined by

$$\epsilon_d(d, y, t) = \frac{1}{t}(y - \mu_d(X)) \qquad S_\beta(d, y, t|\beta) = S_{d\beta}(y|X, \beta).$$

It can be shown that

$$\mathbb{E}[\epsilon_d(D, Y, X) s(D, Y, X|\beta)] = \frac{\partial \mu(\beta)}{\partial \beta}$$

and hence $\mu(\theta)$ is a differentiable parameter. Therefore the efficient score function is given by

$$S_{\text{eff}} = \frac{\mu_{d\beta}\epsilon_d/\pi}{\mathbb{E}[\epsilon^2|X]} \tag{2.25}$$

and the variance bound corresponding to the IPTW estimator is

$$\mathbb{V}^C_{IPTW} = \mu_\beta(d, \beta)' \, \mathbb{E}[S_{\text{eff}} S'_{\text{eff}}]^{-1} \mu_\beta(d, \beta) = \mu_\beta(d, \beta)' \mathbb{E}\left[\frac{\mu_{d\beta}\mu'_{d\beta}}{\pi^2 \sigma^2(X)}\right]^{-1} \mu_\beta(d, \beta), \tag{2.26}$$

where $\mu(d, \beta)$ is a parametrized APO. The following theorem states the asymptotic behaviour of the continuous treatment effect estimators based on the IPTW and the propensity score regression approaches.

**Theorem 2.2** *The causal effect, $\mu(\beta)$, estimated using the efficient score functions (2.23) and (2.25) have the following asymptotic properties*

$$\sqrt{n}(\mu_{GPS}(\hat{\beta}) - \mu(\beta)) \sim \mathcal{N}(0, \mathbb{V}_{GPS})$$
$$\sqrt{n}(\mu^{C}_{IPTW}(\hat{\beta}) - \mu(\beta)) \sim \mathcal{N}(0, \mathbb{V}^{C}_{IPTW})$$

*where $\mathbb{V}_{GPS}$ and $\mathbb{V}^{C}_{IPTW}$ are defined in (2.24) and (2.26), respectively.*

## 2.6 Simulation Studies

### 2.6.1 Simulation study I: Binary Treatment

To illustrate the differences between the variance bounds, we perform two Monte Carlo simulation experiments. The first simulation study examine the performance of the PS and IPTW methods, and compares the standard deviations of the estimators with the bounds developed in this paper under homogeneity and heterogeneity of variance. In this simulation, the structural relationship of interest is defined by

- Homogeneous Case: $Y \sim \mathcal{N}(d + x_1 - 0.5x_2 + dx_1, 5^2)$
- Heterogeneous Case: $Y \sim \mathcal{N}(d + x_1 - 0.5x_2, |\, d + x_1|^2)$

where $X_1$, $X_2$ are normally distributed with mean 2 and variance 1 and $D$ is the treatment arm indicator with the chance of $[1 + \exp(0.5 - x_1 + x_2)]^{-1}$ for being in the treated group, $d = 1$. In this simulation study, we assume that the exposure interact with one of the covariates, $X_1$. Therefore, the estimator introduced in Robins et al. (1992) can not be used. Table 2–1 shows the results of this simulation based on 1000 datasets of sizes 500 and 10,000. The SVB column is defined as the square root of variance bounds for PS, IPTW and DR methods (computed using Monte Carlo). As expected, the variance of the estimator obtained by PS is smaller than

Table 2–1: Binary treatment simulation study under homogeneity and heterogeneity of variance. SVB is the square root of variance bounds related to PS or IPTW methods estimators.

| $\sigma = 5$ | $n = 500$ | | | | $n = 10000$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | s.d. | MSE | SVB | Bias | s.d. | MSE | SVB |
| PS | 0.006 | 0.563 | 0.316 | 0.537 | 0.000 | 0.122 | 0.015 | 0.120 |
| IPTW | 0.016 | 0.681 | 0.464 | 0.632 | 0.004 | 0.148 | 0.022 | 0.140 |
| DR | 0.030 | 0.660 | 0.436 | 0.632 | 0.001 | 0.143 | 0.020 | 0.140 |
| $\sigma = \|d + x_1\|$ | $n = 500$ | | | | $n = 10000$ | | | |
| | Bias | s.d. | MSE | SVB | Bias | s.d. | MSE | SVB |
| PS | 0.001 | 0.546 | 0.298 | 0.268 | 0.004 | 0.123 | 0.015 | 0.056 |
| IPTW | 0.022 | 0.642 | 0.412 | 0.385 | 0.001 | 0.148 | 0.022 | 0.084 |
| DR | 0.002 | 0.633 | 0.401 | 0.385 | 0.007 | 0.143 | 0.021 | 0.084 |

DR. The DR estimator results in an estimator with a smaller variance than IPTW, as it corresponds to the projected IPTW influence function onto the tangent space. In summary, we observe that the variance of the PS estimator is lower than the variance of the DR estimator, which in turn is lower than the IPTW estimator. Moreover, the mean square error (MSE) of the IPTW and DR estimators are always larger than that for the PS.

### 2.6.2 Simulation Study II: Continuous Treatment

We conduct a simulation study to explore the attainability of the introduced variance bound by the estimated APO using GPS and IPTW. We assume $Y \sim N(d+x_1+x_2, 5^2)$, where $X_j$ for $j = 1, 2$ are exponentially distributed with $\mathbb{E}[X_i] = 1$ and, conditionally, $D$ has an exponential distribution with

$$\mathbb{E}[D|X_1, X_2] = \frac{1}{x_1 + x_2}$$

Table 2–2: Continuous treatment simulation study. APO is estimated using GPS method. SVB is the square root of the derived variance bound.

| $\sigma = 5$ | $n = 500$ | | | | $n = 10000$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | s.d. | MSE | SVB | Bias | s.d. | MSE | SVB |
| $d = 0.4$ | 0.091 | 0.251 | 0.071 | 0.224 | 0.046 | 0.057 | 0.005 | 0.050 |
| $d = 0.5$ | 0.137 | 0.246 | 0.079 | 0.227 | 0.080 | 0.060 | 0.010 | 0.051 |
| $d = 0.7$ | 0.073 | 0.239 | 0.062 | 0.227 | 0.060 | 0.053 | 0.006 | 0.051 |
| $d = 1.0$ | 0.001 | 0.230 | 0.053 | 0.222 | 0.001 | 0.052 | 0.003 | 0.050 |
| $d = 2.5$ | 0.060 | 0.307 | 0.098 | 0.262 | 0.048 | 0.062 | 0.006 | 0.050 |
| $d = 3.0$ | 0.073 | 0.359 | 0.134 | 0.309 | 0.054 | 0.068 | 0.008 | 0.056 |
| $d = 4.0$ | 0.135 | 0.488 | 0.256 | 0.426 | 0.080 | 0.085 | 0.014 | 0.073 |
| $d = 5.0$ | 0.252 | 0.644 | 0.478 | 0.554 | 0.129 | 0.109 | 0.028 | 0.093 |

Here we fit the true generalized propensity score model, $f(d|x_1, x_2)$. Table 2–2 summarizes the results for estimated APO using the true GPS for $d \in \{0.4, 0.5, 1, 2.5, 3, 4, 5\}$. Our simulation study reveals that the GPS technique results in a consistent efficient estimator which touches the asymptotic variance bound established in this paper. Table 2–3 reflects the result of APO estimation using the IPTW method. Comparing these two tables shows that even for the continuous treatment, the GPS approach produces the estimator with lower variance and MSE.

## 2.7 Causal Inference in the Longitudinal Setting

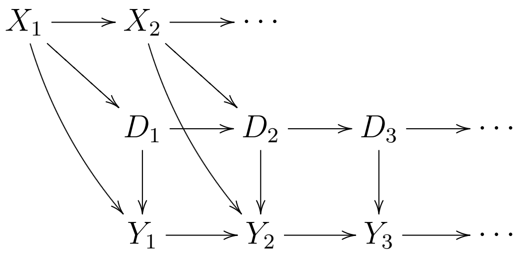
In the previous section, we discussed about asymptotic behaviour of two main methods of causal adjustment, inverse probability of treatment weighting (IPTW) and propensity score (PS) in a single interval setting. In this section, we briefly review this two methods and compare the performance of them using some simulations and real data analysis in longitudinal settings. We start with the binary treatment and

Table 2–3: Continuous treatment simulation study. APO is estimated using IPTW method. SVB is the square root of the derived variance bound.

| $\sigma = 5$ | $n = 500$ | | | | $n = 10000$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | s.d. | MSE | SVB | Bias | s.d. | MSE | SVB |
| $d = 0.4$ | 0.133 | 0.258 | 0.084 | 0.236 | 0.083 | 0.064 | 0.011 | 0.053 |
| $d = 0.5$ | 0.140 | 0.253 | 0.083 | 0.233 | 0.087 | 0.064 | 0.012 | 0.053 |
| $d = 0.7$ | 0.157 | 0.254 | 0.089 | 0.229 | 0.093 | 0.066 | 0.013 | 0.052 |
| $d = 1.0$ | 0.172 | 0.259 | 0.097 | 0.224 | 0.102 | 0.069 | 0.015 | 0.052 |
| $d = 2.5$ | 0.010 | 0.369 | 0.125 | 0.366 | 0.095 | 0.070 | 0.014 | 0.063 |
| $d = 3.0$ | 0.157 | 0.480 | 0.226 | 0.471 | 0.058 | 0.077 | 0.009 | 0.075 |
| $d = 4.0$ | 0.463 | 0.713 | 0.722 | 0.670 | 0.080 | 0.163 | 0.033 | 0.110 |
| $d = 5.0$ | 0.702 | 0.957 | 1.408 | 0.795 | 0.300 | 0.301 | 0.180 | 0.158 |

then extend it to continuous dose in multi-interval setting. We look at the bias, variance and MSE of the estimators of a direct effect of treatment obtained by these two methods.

We consider a longitudinal study with treatment doses $D_{ij}$, responses $Y_{ij}$, and covariates $X_{ij}$ for subjects $i = 1, 2, \ldots, n$ on repeated observation $j = 1, \ldots, K_i$. All variables including dose can be binary, categorical or continuous. A directed acyclic graphic (DAG) representation of the data generating processes we consider is depicted below:



Note that $X_i$ confounds the effect of $D_i$ on $Y_i$. We also admit the possibility that the confounder, treatment and response sequences exhibit autocorrelation. Under an assumption of time-homogeneity, the direct effect of $D$ on $Y$ can be assessed.

We address direct effects of treatment, but this may not reflect the inferential objective of the study in all cases. In longitudinal studies, treatment regimes followed over time may have different effects on *overall* outcome, that is, some response measured only at the end of the study. Marginal structural models (MSMs) are a class of causal models for the estimation from observational data, of the causal effect of a time-dependent exposure in the presence of time-dependent covariates. Typically, MSMs are utilized to estimate the *total* (causal) effect of treatment on an end of study outcome. The parameters of MSMs can be consistently estimated using IPTW, but not by PS methods, as conditioning on the propensity score explicitly blocks (in DAG terms) the path between treatment and subsequent response.

### 2.8 Estimation of Direct and Indirect Effect

When the effect of exposure on outcome is mediated by an intermediate variable, the total exposure effect can be decomposed into direct and indirect effects. Multivariate regression of the outcome on the exposure and the intermediate variable as predictors been often used to estimate the direct effect and the difference between the direct and total exposure effect is called indirect effect. However, Kaufman et al. (2004) show that when the exposure and intermediate variable interact to cause the outcome, subtracting the direct effect from the total effect does not give the indirect effect. In this section, we briefly explain the difference between these two causal effects and the conditions needed to identify them.

The direct effect can be defined as the effect of exposure on the outcome after blocking the exposure effect on an intermediate variable. The indirect effect is a part of exposure effect on the outcome passing through an intermediate variable. Petersen et al. (2006) show that when the exposure and intermediate variable interact to cause the outcome, multivariate regression estimates just a part of direct effect called controlled direct effect. Therefore, they suggested to split the direct effect into controlled and natural direct effect. The exposure effect on an outcome when the intermediate is controlled at a specific level is called controlled direct effect and the natural direct effect is defined as a part of the exposure effect on an outcome while blocking the intermediate effect, but allow the intermediate vary as it would in the absence of exposure. Note that the former can be estimated by blocking all the causal effects on the intermediate, whereas in the later just the effect of exposure on the intermediate is blocked.

The natural and controlled direct effect can be explained through the counterfactual framework. Under the presence of intermediate variable, $Z$, there are two types of counterfactual values: one for outcome and one for intermediate variable denoted by $Y_d$ (an outcome if treatment were set to $d$) and $Z_d$ (an intermediate if treatment were set to $d$), respectively. The controlled direct effect is the difference in the counterfactual outcomes if treated at level $D = d$ versus treated at level $D = d^*$ while the intermediate value is set to $Z = z$. Therefore, the average controlled direct effect is $\mathbb{E}[Y_{dz} - Y_{d^*z}]$ in which $Y_{dz}$ is an counterfactual outcome value when exposure and the intermediate are set to $d$ and $z$, respectively. On the other hand, the natural direct effect can be defined as a difference between a counterfactual outcome if treated at level $D = d$ versus treated at level $D = d^*$ while the intermediate is set to its counterfactual value at $D = d$, $Y_{dz_{d^*}} - Y_{d^*z_{d^*}}$.

VanderWeele (2009) introduces a novel technique to estimate the direct and indirect effects by fitting one marginal structural model for outcome and one for an intermediate variable. As noted by VanderWeele , the consistent estimation of controlled direct effect needs the following no unmeasured confounders assumptions: 1) no unmeasured confounders for exposure-outcome relationship 2) no unmeasured confounders for intermediate-outcome relationship. These two assumptions can be summarized as follows ($X$ and $W$ are the exposure and intermediate confounders, respectively):

$$Y \perp D|X$$

$$Y_{dz} \perp Z|D, X, W$$

where $A \perp B|C$ means $A$ is independent of $B$ given $C$. In other words, we assume that there is no residual confounding of the effect of exposure and the intermediate variable on the outcome after including the measured covariates in the model. Joffe & Colditz (1998) and Kaufman et al. (2005) introduce methods to estimate the control direct effect in the presence of unmeasured confounders. Under these two assumptions, the parameters of a marginal structural model $\mathbb{E}[Y_{dz}] = g(d, z)$ can be consistently estimated using inverse probability of treatment weighting methods and the controlled direct effect is given by $\mathbb{E}[Y_{dz} - Y_{d^*z}] = g(d, z) - g(d^*, z)$.

To estimate the natural direct effect, VanderWeele (2009) suggests to fit an additional marginal structural model on the exposure-intermediate, thus an additional no unmeasured confounders assumption is needed for the exposure and intermediate relationship. More specifically

$$Z_d \perp D|X.$$

We also require

$$Y_{dz} \perp Z_{d^*}|X.$$

The last condition states that there is no intermediate confounder that is affected by the exposure.

Under the above assumptions an unobserved counterfactual outcome can be estimated using the observed outcomes and the average counterfactual outcome is given by

$$\mathbb{E}[Y_{dz}] = \sum_x \mathbb{E}[Y|D = d, Z = z, X = x]p(X = x),$$

and the average natural direct effect is estimated by

$$\mathbb{E}[Y_{dz_{d^*}} - Y_{d^*z_{d^*}}] = \sum_x [g(d, h(d^*, x), x) - g(d^*, h(d^*, x), x)]p(X = x)$$

where $\mathbb{E}[Y_{dz}|X] = g(d, h(d, x), x)$ and $\mathbb{E}[Z_d|X] = h(d, x)$. Note that, in contrast to controlled direct effect, for natural direct effect the counterfactual outcome is estimated using the conditional marginal structural model.

The natural indirect effect compares the counterfactual outcomes when the exposure level is set to $D = d$ and the intermediate is set to what would it had been if the exposure is set to $d$ versus the intermediate value that would had been observed if the exposure were set to $d^*$, $Y_{dZ_d} - Y_{dZ_{d^*}}$. VanderWeele (2009) decomposes the total direct effect to natural direct and indirect effect even if the outcome model involves the interaction between exposure and intermediate variable and nonlinear models. Therefore, the total exposure effect of $d$ versus $d^*$ is given by

$$Y_d - Y_{d^*} = (Y_{dZ_d} - Y_{dZ_{d^*}}) + (Y_{dZ_{d^*}} - Y_{d^*Z_{d^*}})$$

where the first part is the natural indirect effect and the second part is the natural direct effect.

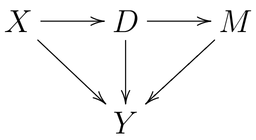
Pioneers work on the identifiability of direct and indirect effect impose some stronger assumptions than those presented in VanderWeele (2009). For instance, Robins & Greenland (1992) impose the assumption of no interaction between exposure and intermediate which can be very unrealistic in some settings. In fact, if there is no interaction between exposure and intermediate variable the controlled and natural direct effect will be equivalent because $Y_{dz} - Y_{d^*z}$ will be free of $z$. For further detail on the direct and indirect causal effects see Robins et al. (2010), Hafeman & VanderWeele (2010), Robins & Richardson (2010), VanderWeele (2010) and Van Der Laan & Petersen (2004).

### 2.9 Simulation Study: Binary treatment with a mediating variable

Another case where PS performs better than IPTW in terms of MSE is in the presence of mediating variable, and it also appears that PS method is more successful in removing bias. In this section we report results of a small simulation study with a mediating variable. We have one time independent covariate $X_i$, one posttreatment intermediate variable $M_i$ that may serve as a mediator for the treatment outcome, a treatment indicator $D_i$ and response $Y_i$. We use the following densities:

$$\begin{aligned} X_i &\sim \mathcal{N}(2,10) \\ M_i &\sim \mathcal{N}(d_i,5) \\ D_i &\sim Bernoulli(p(x_i)) \qquad p(x_i) = \frac{1}{1+\text{expit}\{-2-0.2x_i\}} \\ Y_i &\sim \mathcal{N}(-d_i+x_i+m_i,5) \end{aligned}$$

for $i = 1, ..., n$. Note that $M$ can be written as $d_i + \mathcal{N}(0,5)$ so the true ATE is zero. A DAG representation of the data generating processes is as follows:



We generated 1000 data sets of size 300 and 5000. We have used the following models for response variable using PS and IPTW and propensity score:

$$\begin{aligned} \text{logit}\{\pi_i\} &= \text{logit}\{p(D_i=1|X)\} = \alpha_0 + \alpha_1 x_i \\ y_i &= \beta_0 + \beta_1 d_i + \beta_2 \pi_i + \epsilon_i \end{aligned} \tag{PS}$$

$$\begin{aligned} w_i^{-1} &= \frac{\exp\{D_i(\alpha_0+\alpha_1 X_i)\}}{1+\exp\{(\alpha_0+\alpha_1 X_i)\}} \\ y_i &= \beta_0' + \beta_1' d_i + \beta_2' m_i + \epsilon_i \end{aligned} \tag{IPTW}$$

We also utilize the truncated version of IPTW, IPTW.t, in which we have retained only those observations with either $0.05 \leq \pi_i \leq 0.95$, where $\pi_i = w_i^{-1}$. Although, in general, truncation may result in a biased estimate, it reduces the variance by

Table 2–4: ATE estimates based on IPTW, IPTW.t and PS for causal parameter $\beta^* = 1$.

| | $n = 300$ | | | $n = 5000$ | | |
|---|---|---|---|---|---|---|
| | ATE | s.d. | MSE | ATE | s.d. | MSE |
| IPTW | 1.729 | 4.083 | 19.657 | 0.207 | 2.017 | 4.113 |
| IPTW.t | 0.174 | 1.907 | 3.667 | 0.015 | 0.435 | 0.190 |
| PS | 0.088 | 1.402 | 1.973 | 0.001 | 0.332 | 0.110 |

deleting those observation with a weight close to the boundaries zero or one. Using the idea in VanderWeele (2009) we fitted another model considering $M$ as a response variable to deal with counterfactuals in $M$,

$$\widehat{m}_i = \lambda_0 + \lambda_1 d_i$$

so that the total causal effect using IPTW is $\beta_1' + \beta_2' \lambda_1$. Table 2–4 shows the estimated ATE based on IPTW, IPTW.t and PS. Under the assumption of correct model specification for the probability of treatment, IPTW has larger bias and standard deviation compared to PS for $n = 300$ and $n = 5000$. In the presence of the mediator, the PS method is more successful in removing the bias and also has smaller variance than the IPTW methods.

## 2.10 Causal Adjustment for Repeated Measures Data

### 2.10.1 The Multivariate GPS (MGPS)

In the case of dose response estimation from repeated measures or multi-interval data because of correlation structure in the data the potential patterns of time-varying confounding are more complex that can be dealt with using a univariate GPS approach. The GPS approach introduced in this section is suitable for the analysis of repeated measures response data with interval-dependent dosing. We denote $Y_{ij}$ as a response of $i$th unit, $i = 1, ..., n$ in interval $j$, $j = 1, ..., n_i$; dose and covariate variables are similarly subscripted. Furthermore, sequential weak unconfoundedness can be defined as

$$Y_{ij}(d) \perp D_{ij} | X_{i1}, ..., X_{ij}.$$

That is, at each interval, assignment to dose $D_{ij}$ is weakly unconfounded with the response during interval $j$ given covariates, previous response, and dose values measured up to the start of the $j$th interval. Moodie & Stephens (2010) shows that if we define $\bar{X}_{ij} = (X_{1j}, ..., X_{ij})$ as a history of covariates, response and previous doses and let $\pi_{ij}(d, \bar{X}_{ij})$ be the multivariate GPS then, for every dose $d$,

$$Y_{ij}(d) \perp D_{ij} | \pi_{ij}(d, \bar{X}_{ij})$$

that is, for $d \in D$, current potential response $Y_{ij}(d)$ is conditionally independent of the distribution of dose received $D_{ij}$ given the MGPS $\pi_{ij}$ , for all $i$ and $j$. In the same paper, it has also been shown that the APO obtained by averaging $\mathbb{E}[Y_{ij}(d) | \pi(d, \bar{X}_{ij})]$ over the distribution of the covariates $\bar{X}_{ij}$, is an unbiased estimator of the dose response function $\mu(d) = \mathbb{E}[Y_{ij}(d)]$. Note that a univariate GPS analysis that does not construct $\pi$ by conditioning on $\bar{X}_{ij} = \bar{x}_{ij}$ for each $j$ does not necessarily achieve bias removal.

We have carried out extensive testing of the MGPS approach and performed comparisons with non-causal and standard GPS (MGPS) methods. Our examples demonstrate the importance of the use of the multivariate extension of the GPS. We have also compared the performance of the GPS and IPTW estimators for estimating the direct treatment effect. Our simulation studies show that the MGPS introduced by Moodie & Stephens (2010) outperforms the IPTW in terms of both bias and variance reduction. We have discussed some of the limitations of the MGPS in the conclusion section of this Chapter.

### 2.10.2 The IPTW Estimator for Repeated Measure Data

To implement IPTW in the repeated measures setting, the following model is fitted for response variable to estimate the total treatment effect,

$$\mathbb{E}[Y_{ij}|D_{ij} = d_{ij}] = \beta_0 + \beta_1 d_{ij}$$

with the stabilized weights

$$w_{ij}^{s} = \frac{p(D_{ij} = d_{ij}|D_{i(j-1)} = d_{i(j-1)})}{p(D_j = d_{ij}|D_{i(j-1)} = d_{i(j-1)}, \overline{X}_k = x_{ij})}$$

where $\overline{D}_{-1} = 0$. Thus the outcome at interval $j$ is weighted with the inverse probability of treatment at that interval, modelled as a function of previous covariates, responses and doses. This is the natural extension of IPTW to the time-homogeneous repeated measures setting.

Note that the difference between the fitted models in this section and MSM approach used in Robins & Brumback (2000) is that here we do not have a single response at the end of follow up, but several responses and weights corresponding to each interval. In other word, our weighted model produces the pseudo-intervals based on observed treatment doses at each time point rather than the pseudo-population through received treatment doses path up to end of follow-up.

Table 2–5: ATE estimates based on IPTW, IPTW.t and PS for causal parameter $\beta^* = 1$

| | $n = 300$ | | | $n = 5000$ | | |
|---|---|---|---|---|---|---|
| | ATE | s.d. | MSE | ATE | s.d. | MSE |
| IPTW | 0.536 | 1.511 | 2.499 | 0.949 | 0.589 | 0.349 |
| IPTW.t | 0.946 | 0.952 | 0.910 | 1.004 | 0.235 | 0.055 |
| PS | 0.948 | 0.910 | 0.627 | 0.997 | 0.193 | 0.037 |

## 2.11 Longitudinal Setting: Simulation Studies and Examples

### 2.11.1 Binary treatment

In this section we report results of a small longitudinal simulation study carried out to evaluate the performance of the IPTW and PS explained. We have one time dependent covariate $X_{ij}$, treatment indicator $D_{ij}$ and response $Y_{ij}$ with the following densities:

$$
\begin{aligned}
&X_{ij} \sim \mathcal{N}(1, 2)\\
&D_{ij} \sim \mathit{Bernoulli}(\text{expit}\{\mathbb{I}\{j = 1\}(2 - x_{ij}) + \mathbb{I}\{j > 1\}(2 - 0.2Y_{i(j-1)} - x_{ij})\})\\
&Y_{ij} \sim \mathcal{N}(D_{ij} + 2X_{ij}, 5)
\end{aligned}
$$

for $i = 1, ..., n$ and $j = 1, ..., 5$, where $\text{expit}\{x\} = \exp(x)/(1 + \exp(x))$ and $\mathbb{I}\{.\}$ is an indicator function. We generated 1000 data sets of size 300 and 5000. Table 2–5 shows the estimated ATE based on IPTW and PS.

Under the assumption of correct model specification for weights and propensity score, IPTW has larger bias and standard deviation compared to PS for $n = 300$ and $n = 5000$. Although weight truncation helps the IPTW method to reduce the MSE, it still has a slightly larger MSE than PS method. As we expected both methods are successful in removing the bias in the large sample size, $n = 5000$.

### 2.11.2 Simulation: Nonlinear, nonadditive treatment effect

Here, we use the same simulation study as in Moodie & Stephens (2010) which is the extended version of Hirano & Imbens (2004) to a two interval setting.

***Data Generation:*** Suppose that at first and second interval, have

$$Y_1(d)|X_{11}, X_{12} \sim \mathcal{N}(d + (X_{11} + X_{12})\exp(-d(X_{11} + X_{12})), 1)$$
$$Y_2(d)|X_{21}, X_{12} \sim \mathcal{N}(d + (X_{21} + X_{12})\exp(-d(X_{21} + X_{12})), 1)$$

The marginal distribution of each of $X_{11}$, $X_{12}$, and $X_{21}$ are all unit exponential and the marginal mean of the response in both intervals is identical. Let $D_1 \sim Exp(X_{11} + X_{12})$, $D_2 \sim Exp(X_{21} + X_{12})$. The APO at dose $d$, $\mu(d)$, can be obtained by integrating out the covariates analytically, yielding

$$\mu(d) = d + \frac{2}{(1+d)^3}.$$

In this section we want to compare the performance of estimators of APO based on IPTW, GPS and MGPS. As suggested by Robins & Brumback (2000), stabilized weights are estimated using a normal density for the IPTW analysis, and weighted splines has been used to fit the model for responses on dose. In GPS analysis, a multivariate GPS analysis, involves the GPS vector $\pi^M = (\pi_1, \pi_2)$:

$$\pi_1 = (X_{11} + X_{12})\exp(-d(X_{11} + X_{12}))$$
$$\pi_2 = (X_{21} + X_{12})\exp(-d(X_{21} + X_{12}))$$

Table 2–6: Pointwise bias estimates for causal curve based on IPTW using splines and GPS

| | IPTW | | | MGPS | | |
|---|---|---|---|---|---|---|
| | $\mu(d) - \widehat{\mu}(d)$ | Var | MSE | $\mu(d) - \widehat{\mu}(d)$ | Var | MSE |
| $d = 0.05$ | -0.595 | 0.063 | 0.417 | 0.001 | 0.043 | 0.043 |
| $d = 0.10$ | -0.519 | 0.030 | 0.300 | 0.001 | 0.030 | 0.030 |
| $d = 0.20$ | -0.350 | 0.019 | 0.141 | 0.000 | 0.016 | 0.016 |
| $d = 0.55$ | -0.037 | 0.018 | 0.020 | 0.000 | 0.005 | 0.005 |
| $d = 0.65$ | -0.016 | 0.023 | 0.024 | 0.000 | 0.005 | 0.005 |
| $d = 1.00$ | -0.012 | 0.030 | 0.030 | 0.000 | 0.006 | 0.006 |
| $d = 2.50$ | -0.053 | 0.102 | 0.105 | 0.001 | 0.006 | 0.006 |
| $d = 5.50$ | -0.023 | 0.317 | 0.317 | 0.000 | 0.009 | 0.009 |

where consists of correctly specified models. A univariate GPS analysis might fail to include information from the previous interval and hence the GPS used would be $\pi^U = (\pi_1, \pi_2^*)$ where $\pi_1$ is as before, but $\pi_2^* = X_{21} \exp(-dX_{21})$.

We generated 1000 data sets of size 250. The estimated APO using MGPS are exactly correct, while the UGPS and IPTW analysis are clearly biased. The general shape of the UGPS and IPTW APO are correct, however these estimators do not catch the curve (see Figure 2–1). Table 2–6 shows the bias, variance and MSE's of the estimated APO using IPTW and MGPS. The bias and MSE obtained by MGPS are significantly smaller.

As pointed out by Hirano et al. (2003), the efficiency of the GPS estimator can be improved by using the estimated GPS. In this simulation, the GPS can be estimated using a Gamma generalized linear model, for example.

### 2.11.3 Example: The MSCM Study

Alexander & Markowitz (1986) studied the relationship between maternal employment and paediatric health care utilization. The investigation was motivated by the major social and demographic changes that have occurred in the US since 1950. The Mothers' Stress and Children's Morbidity Study (MSCM) enrolled 167 preschool children between the ages of 18 months and 5 years that attended an inner-city paediatric clinic. Each individual provided information regarding their family and work outside the home. Daily measures of maternal stress and child illness were recorded during 4 weeks of follow-up. We use these data to determine casual effect of stress on child illness. We used logistic regression to fit the model for weights and propensity score over each interval with employment ($e$), married ($m$), previous stress ($s$) and previous illness ($i$) as covariates, as follows:

$$\text{logit}\{p(s_{i1} = 1)\} = \alpha_0 + \alpha_1 e + \alpha_2 m$$
$$\text{logit}\{p(s_{it} = 1)\} = \alpha'_0 + \alpha'_1 s_{i(t-1)} + \alpha'_2 i_{i(t-1)} + \alpha'_3 e + \alpha'_4 m$$

for $t = 1$ and $t > 1$ respectively. Since our response, illness, is a dichotomous random variable we fitted the following logistic models for IPTW and PS methods:

$$\text{logit}\{p(i_{it} = 1)\} = \gamma_0 + \gamma_1 s_{it} \tag{IPTW}$$
$$\text{logit}\{p(i_{it} = 1)\} = \gamma'_0 + \gamma'_1 s_{it} + \gamma'_2 \pi_i(x) \tag{PS}$$

where $\pi(x)$ is the propensity score. In order to see the effect of sample size in our estimators, we estimate the ATE for different sample sizes by randomly deleting individuals. Results are presented in Table 2–7.

As the sample size increases estimators become more similar, and for each sample size the IPTW standard errors are slightly smaller. Since there is a large overlap between estimated parameter confidence intervals using IPTW and PS, neither is

Table 2–7: Parameter estimates based on IPTW and MGPS for MSCM study

| | $\widehat{\gamma}_1$ | s.e.($\widehat{\gamma}_1$) | $\widehat{\theta}_1$ | s.e.($\widehat{\theta}_1$) |
|---|---|---|---|---|
| $n = 50$ | 0.708 | 0.144 | 0.622 | 0.205 |
| $n = 70$ | 0.644 | 0.129 | 0.617 | 0.189 |
| $n = 100$ | 0.520 | 0.109 | 0.576 | 0.150 |
| $n = 167$ | 0.547 | 0.083 | 0.537 | 0.115 |

preferable to the other one in this example. We have also checked the truncated weights, IPTW.t, estimators, but the results are omitted because they were fairly similar.

In the next example we have a longitudinal data set with continuous response and treatment dose and we will compare the performance of univariate GPS, multivariate GPS and IPTW approaches.

### 2.11.4 Example: MOTAS Amblyopia Study

Amblyopia is the most common childhood vision disorder, and is characterized by reduced visual function in one eye. A standard treatment for the condition is occlusion therapy (patching) of the properly functioning eye. Until recently, the apparent beneficial effect of occlusion therapy had not been well quantified, partly due to difficulty in the accurate measurement of the occlusion dose. The Monitored Occlusion Treatment of Amblyopia Study (MOTAS) (Stewart et al. (2004)) was the first clinical study aimed at quantifying the dose response relationship of occlusion, facilitated by the use of an electronic occlusion dose monitor, consisting of an eye patch with two electrodes attached to its undersurface connected to a battery-powered data-logger powered by battery from which patch use was read by clinicians at follow-up visits.

The MOTAS design and a full description of the study base have been published previously (Stewart et al. (2002), Stewart et al. (2004)). At study entry, all children who required spectacles entered the refractive adaptation phase; the remainder entered the occlusion phase directly. Children still considered amblyopic after refractive adaption began occlusion and were prescribed six hours of occlusion daily. Visual acuity was measured on the logarithm of Minimum Angle of Resolution (logMAR) scale; improvement is indicated by a decrease in logMAR. Visual function and monitored occlusion dose were recorded at approximately two-week intervals until acuity ceased to improve, at which point children exited the study and returned to usual care. A total of 116 children were enrolled in MOTAS; we analyze data of the 68 who took part in the occlusion phase (whether they participated in the refractive adaption phase of the study or not) who, although prescribed six hours of occlusion daily, received varying occlusion doses because of incomplete concordance. Our notation is as follows: for child $i$, the response, $Y_{ij}$ , is the change in visual acuity during interval $j$, and $D_{ij}$ is the random occlusion dose (in hours) received in interval $j$. Intervals are approximately two weeks in length, thus a child who concorded perfectly with prescribed treatment would be have a dose of 84 hours in an interval (that is, six hours daily for 14 days). However, children typically did not follow the prescribed occlusion dose, and both higher and lower than prescribed doses were observed.

In the study, dose is a continuous variable, but 60 out of 404 (about 15%) of intervals in the occlusion phase had a zero dose. In order to acknowledge the mixture nature of the dose distribution in the GPS or IPTW, we assume that

$$D_{ij} \overset{\mathcal{L}}{=} \psi(\bar{x}_{ij},\gamma)\mathbb{I}\{d=0\} + (1-\psi(\bar{x}_{ij},\gamma))\mathbb{I}\{d\neq 0\}D^{+}_{ij}$$

where $D^+_{ij}$ is strictly positive random variable and $0 < \psi(\bar{x}_{ij}, \gamma) < 1$ is a mixing weight which can be estimated using logistic model on binary ($D_{ij} = 0/D_{ij} > 0$) dose data.

Following the fitted model by Moodie & Stephens (2010), we included the visual acuity at start of interval, age, sex, interval number, length of interval (in days), and amblyopic type (anisometropic, strabismic, mixed) as a covariate in the GPS or IPTW model and if we add the previous dose to these covariates MGPS can be fitted. These covariates were used to predict both the probability of having any occlusion at all ($D/D > 0$) in a logistic model and the probability of receiving a particular dose (greater than zero) of occlusion in a Gamma model. The UGPS used is

$$\widehat{\pi}(d, x_{ij}) = \widehat{\psi}(x_{ij}, \widehat{\gamma})\mathbb{I}\{d = 0\} + (1 - \widehat{\psi}(x_{ij}, \widehat{\gamma}))\mathbb{I}\{d \neq 0\}f(d|x_{ij}, \widehat{\phi}, \widehat{\alpha})$$

where $f(d|x_{ij}, \widehat{\phi}, \widehat{\alpha})$ is a Gamma density with shape $\phi$ and scale determined by $\alpha$. We used the same model to assign the weights for each individual in IPTW method. The fitted model model for MGPS is identical with $x_{ij}$ replaced by $\bar{x}_{ij}$ which includes the previous dose.

$$\begin{aligned}\widehat{\pi}(d, x_{ij}, d_{i(j-1)}) &= \widehat{\psi}(x_{ij}, d_{i(j-1)}, \widehat{\gamma})\mathbb{I}\{d = 0\} \\ &\quad +(1 - \widehat{\psi}(x_{ij}, d_{i(j-1)}, \widehat{\gamma}))\mathbb{I}\{d \neq 0\}f(d|x_{ij}, d_{i(j-1)}, \widehat{\phi}, \widehat{\alpha})\end{aligned}$$

As response in the MOTAS is the vector of changes in visual acuity, there is little observed serial correlation in the data. The observable model for change in visual

Table 2–8: Estimated parameters in APO models based on UGPS and MGPS, estimated variances are in bracket

| | UGPS | MGPS |
|---|---|---|
| $\beta_1$ | -0.107(0.031) | -0.135(0.046) |
| $\beta_2$ | 9.00e-6(1.84e-4) | 1.740e-4(2.28e-4) |
| $\beta_3$ | 2.917(1.580) | 5.668(2.264) |
| $\beta_4$ | 0.080(0.047) | 0.069(0.066) |

acuity, $Y$, in the GPS method is modelled via the expectation

$$\mathbb{E}_{Y|D,\pi}[Y|D=d,\pi,\beta] = \beta_0 + \mathbb{I}\{\pi < 0.05\}(\beta_1 + \beta_2 d + \beta_3 \pi + \beta_4 d.\pi)$$

and in order to decrease the bias in IPTW estimator, we have used the semiparametric regression using weighted splines to fit the model for $Y$ on $D$. A plot of the dose-response curve is presented in Figure 2–2. The MGPS, univariate GPS and IPTW APO's are plotted for comparison with 95% confidence interval based on MGPS. As Figure 2–2 shows, there is no significant difference between the estimated APO using either IPTW or GPS method. Numerical values of the estimated parameters using least square estimates, $\beta_1, ..., \beta_4$, are presented in Table 2–8 for UGPS and MGPS.

The plot indicates that the direct effect of dose on visual acuity, when confounding between dose and the responses is adjusted for using the GPS approach, is appreciable; the average potential effect on change in visual acuity measurement $Y_{ij}$ is significantly negative (corresponding to vision improvement) over the entire range of positive doses considered.

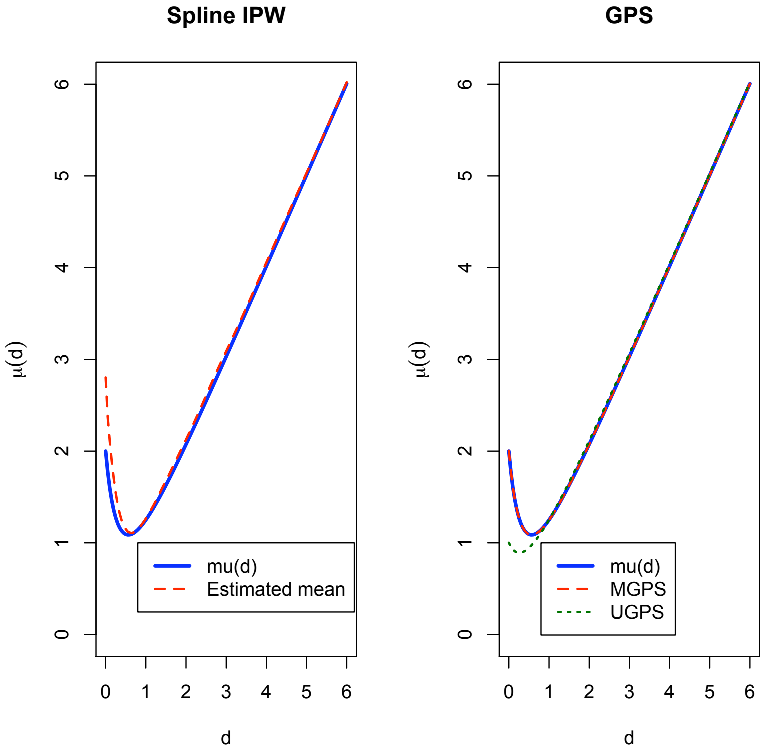


Figure 2–1: Simulated Example: The dose-response APO curves for the IPTW and GPS analyses.

## 2.12 Conclusion

In this Chapter, our primary focus has been on the use of the propensity score in a model based adjustment, where propensity score replaces the whole vector of covariates named as propensity score regression adjustment. We derived the semiparametric variance bound of the estimated causal effect using propensity score regression adjustment, and showed that the obtained bound is equal to the efficiency bound introduced in the literature on semiparametric regression. A parametric model was assumed for the propensity score and the parameters of this model are treated as nuisance parameters. The nuisance tangent space was built based on the parameters in the treatment mechanism model. Using the theory of semiparametric inference, the efficient influence function corresponding to the estimator of interest has been constructed as a residual from projecting any influence function onto the nuisance tangent space. In a simple example, we have shown that the propensity score stratification estimator also attains the semiparametric efficiency bound.

The semiparametric variance bound obtained is lower than the bound for doubly robust estimator. Therefore, the propensity score regression/stratification dominates IPTW and doubly robust estimators in terms of efficiency. The drawback, of course is that both treatment assignment and the mean models have to be correctly specified to result in consistent estimator, which is not the case in doubly robust estimators.

Our studies clearly demonstrate that in a range of simulation studies in single and multiple interval settings, PS methods outperform IPTW in terms of MSE. Therefore, in the context of moderate to high-dimensional covariate/confounder vectors, the scalar propensity score provides a straightforward causal adjustment approach which seems to have superior finite dimensional performance.

We outlined the Generalized Propensity Score, a generalization of the classical binary treatment propensity score, and showed that since the confounding pattern is more complex in longitudinal data, the GPS needs to take into account the correlation between observations. We explained how the GPS can be modified to keep the balancing property in the context of repeated measures data. We compared the performance of the IPTW and GPS approach to estimate the average potential outcome through simulation studies, MSCM and MOTAS data. Our studies reveal that the ATE estimator using propensity score regression adjustment has a smaller variance and is more successful in removing bias than corresponding methods that use weighting, under correct model specification.

As noted by Rubin, before running any regression model, which considers a propensity score as a single covariate, and looking at the coefficients to estimate the causal effect certain criteria have to be checked to make sure that there is enough information in data to be extracted and used to estimate the causal parameter of interest (see Rubin (2004b, 2008)).

One limitation of PS methods at this stage is that they have only been developed for use in the estimation of direct effects, and cannot be used for the estimation of total effects, whereas the marginal structural models approach that utilizes IPTW does allow the estimation of total effects. Hernán et al. (2004) also show that in the presence of time varying confounders if there exist an unmeasured common effect of a confounder and the response, then covariate adjustment techniques such as propensity score regression result in bias treatment effect estimate.

In the next Chapter, we use a slight modification of the propensity score regression as a causal adjustment method which has the double robust property and introduce a novel confounder selection technique using the penalization. We also show that one may gain efficiency by adjusting just for the key confounders selected by our proposed technique.

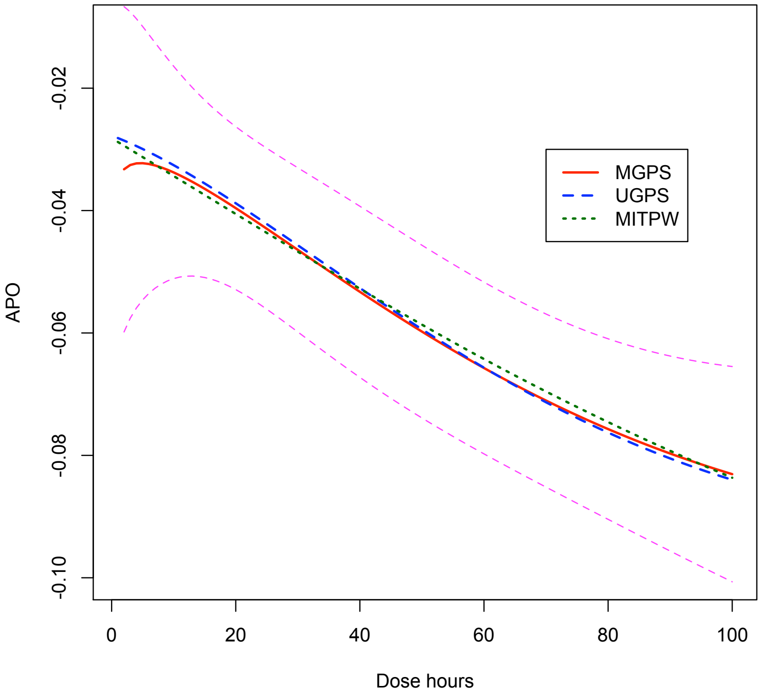


Figure 2–2: MOTAS data: The estimated average potential change in visual acuity (APO) vs dose for multi-interval IPTW (MIPW), UGPS and MGPS. Pointwise 95% confidence interval (light dashed) computed for MGPS.

# CHAPTER 3
# Variable Selection in Causal Inference

## Chapter Summary

We address the common problem faced by practitioners of how to select variables for the propensity score model commonly used to correct for confounding in observational data. We construct a double robust regression procedure which incorporates the treatment assignment model as well as the response mean model for estimating the treatment effect. In an analytic example, we show that selecting important covariates can increase efficiency of estimation of the causal parameter while retaining consistency. We demonstrate that treatment-only variable selection techniques may ignore important covariates which are strongly related to the outcome but not to the treatment. We introduce a novel covariate selection technique based on penalized likelihood which considers the response and treatment assignment models simultaneously. The selected covariates via our proposed method can then be used in other causal adjustment techniques as well as our proposed regression estimator. We derive the asymptotic properties of the estimators, and illustrate their small-sample behaviour using simulation. We apply the proposed method to analyze the National Supported Work Demonstration (NSWD) data.

## 3.1 introduction

In the analysis of observational data, when attempting to establish the magnitude of the effect of treatment (or exposure) in the presence of confounding, the practitioner is faced with certain modelling decisions that facilitate the estimation of the causal effect of treatment. Specifically, in the majority of cases, two statistical models must be proposed;

(i) the *conditional mean model* that models the expected response as a function of predictors, and

(ii) the *treatment allocation model* that describes the mechanism via which treatment is allocated (or, at least, received) by individuals in the study, again as a function of the predictors.

Predictors that appear in the data generating mechanisms for (i) and (ii) are termed confounders, and the omission of confounders from model (ii) is typically regarded as a serious error, as it leads to inconsistent estimators of the treatment effect. Because of this, practitioners usually adopt a conservative approach, and attempt to ensure that they do not omit confounders by fitting a richly parameterized treatment allocation model. The danger with this strategy is that it can lead to non-confounders, that is, predictors that predict treatment allocation, but not response, being included in the treatment allocation model.

The inclusion of such “spurious” variables in model (ii) is usually regarded as harmless. However, it is reasonable to assume that there must be some inferential penalty for failing to fit the correct model, and this penalty usually takes the form of inflation of variance of the estimator. This problem is still present for the conditional mean model, but better understood, and in practice less problematic as practitioners seem to be more concerned with adjustment for confounding, and therefore more likely to introduce the spurious variables in model (ii). Other than limited simulation evidence, little formal guidance as to how the practitioner should act in this setting has been provided by statisticians.

This Chapter addresses these issues. We demonstrate the variance inflation caused by the inclusion of spurious variable, and provide a technique based on penalization that provides automatic consistent variable selection for both conditional mean and treatment allocation models.

A common strategy in causal inference with large number of covariates has so far been through dimension reduction using the propensity score method. We therefore focus on propensity score adjustment, and estimation of the *direct* causal effect (Rubin (2004a)). In many applications, particularly in the clinical sciences, ethical considerations often result in covariate imbalance at the intervention stage. As such, only few covariates usually form the basis for an assignment to the control or treatment group. We encounter the same situation in most observational studies too, i.e. a small number of covariates often suffices to explain the observed imbalance in groups. It therefore seems that a more reasonable line of attack should involve variable selection rather than dimension reduction. However, to minimize the risk of ignoring confounders, it seems beneficial to adjust for all the covariates by fitting a "rich" propensity score model. This, however, may result in an inflation of the variance of the estimated parameters and induce bias as well (Greenland (2008) and Schisterman et al. (2009)). A variable selection approach can in practice lead to significant efficiency gain. Our analytic example in Section 3.3 indicates that this issue is not only confined to applications with a large number of attributes. Even with a small number of variables, removing spurious variables can result in significant efficiency gain. This point has also been conjectured by Brookhart et al. (2006a) and studied by means of simulations.

Confounder selection methods based on either the propensity score or the response model may result in failure to account for important confounders which barely predict the treatment or the response, respectively (Crainiceanu et al. (2008)). Vansteelandt et al. (2010) show that confounder selection procedures based on $AIC$ and $BIC$ can be sub-optimal and introduce a method which targets the treatment effect by minimizing the mean square error of the estimated treatment effect. The latter is closely related to the cross-validation method introduced by Brookhart & Van Der Laan (2006).

In this Chapter, we present a penalization technique based on the joint likelihood of the treatment and response variables to select the key covariates that need to be included in the treatment assignment/PS model for estimation of the causal effect. Section 3.2 contains some preliminaries on penalized estimation technique. We present an analytic example in Section 3.3 that illustrates how variable selection can lead to efficiency gain. We then develop the appropriate methodology and study the theoretical properties of the method in Sections 3.4. The methodology is then tested using simulation in Section 3.5, before being applied to a real data set in Section 3.6. As shown by Robins & Greenland (1992), in the presence of mediation, the treatment and the intermediate variables interact to cause the outcome, and hence the direct and indirect effect may not be identifiable in general (see Petersen et al. (2006), Robins et al. (2010), and Hafeman & VanderWeele (2010)). To avoid such non-identifiability issues, we focus on estimation of the unmediated causal effect.

## 3.2 Preliminaries

In this section we present some preliminaries on different covariate types and penalized estimation that are needed in the sequel.

### 3.2.1 Different Types of Covariates

The strong ignorability condition is key in causal inference; it follows automatically from the correct specification of the treatment allocation model. Therefore, if important confounders are omitted, this assumption may be violated and lead to an inconsistent estimator. On the other hand, adding all the unrelated covariates in the treatment assignment model may inflate the variance of the estimator of the causal effect. Therefore, before proposing a causal effect estimation procedure, the correct covariates must be selected to be entered into the propensity score model.

We assume that there are three types of covariates: (I) those which are just related to the treatment, (II) which are related to the outcome as well as the treatment (confounders), and (III) which are just related to the outcome variable. In standard causal inference problems (without mediation), it is often assumed that all the covariates related to the treatment are also related to the outcome. In this situation, all covariates related to the treatment have to be included in the propensity score model which can be achieved by any treatment allocation-based covariate selection techniques. In the directed acyclic graph (DAG) in Figure 3–1, Type-I and Type-III are covariates either related just to the outcome or just to the treatment.

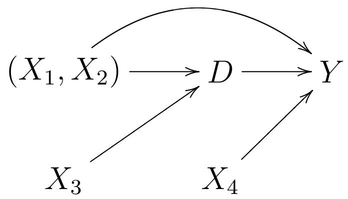


Figure 3–1: Covariate types: Type-I: $X_3$, Type-II: $(X_1, X_2)$ and Type-III: $X_4$.

In this situation, model selection techniques which only consider the treatment mechanism ignore covariates that are related to the outcome but not treatment and hence lead to inefficient estimators. In contrast, our proposed model selection technique involves the treatment and the outcome models at the same time to keep the important covariates in the model.

Brookhart & Van Der Laan (2006) introduced a technique based on a cross-validation criterion to select a model for the treatment mechanism in marginal structural models. Cross-validation has, however, some drawbacks. For example, it can be extremely variable, and is not readily applicable in high-dimensional cases when the number of covariates is larger than the sample size. The simulation study Brookhart et al. (2006a) suggests that variables unrelated to the treatment, but related to the outcome, should always be included in the propensity score model. The inclusion of these variables will decrease the variance of an estimated exposure effect without increasing bias. Our penalized estimation technique confirms the simulation-based results of Brookhart et al. (2006a).

In the section 3.3, we explain through a simple example why including covariates related only to the outcome can improve the efficiency of the estimator without imposing any bias, whereas including covariates merely related to treatment does not introduce bias, but does lead to variance inflation in the estimator of the causal effect.

### 3.2.2 Penalized Estimation

Several methods for choosing the most important covariates in the regression of an outcome $Y$ on covariates $\mathbf{X}$ in terms of response prediction include stepwise and subset selection. When the dimension of $\mathbf{X}$ is large, however, these methods are computationally expensive and unstable. Tibshirani (1996) proposed the LASSO, which shrinks some coefficients and sets the others to zero by adding a penalty function to the sum-of-squares function. The penalized least-squares estimator $\hat{\eta}_{ls}$ is

$$\hat{\eta}_{ls} = \arg\min_{\eta} \left\{ ||\mathbf{y} - \mathbf{X}\eta||^2 + n \sum_{j=1}^{p} p_\lambda(|\eta_j|) \right\},$$

where $\eta$ is a $p$-dimensional regression coefficient and $p_\lambda(.)$ is a penalty function. This is readily generalized to likelihood based models, where the maximum penalized likelihood estimator (MPLE) is given by

$$\hat{\eta}_{ml} = \arg\min_{\eta} \left\{ l_m(\eta) + n \sum_{j=1}^{p} p_\lambda(|\eta_j|) \right\},$$

and $l_m(\eta)$ is the minus log-likelihood. We note that MPL estimators are shrinkage estimators. As such, they have more bias, though less variation. Over the past decade, many other shrinkage methods have been introduced using other penalty functions. Some of the most well-known of these penalty functions are:

- LASSO (Tibshirani (1996)):

$$p_\lambda(|\beta_j|) = \lambda|\beta_j|$$

- SCAD (Fan & Li (2001)): for $a > 2$

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| < \lambda \\ -\dfrac{(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta_j| < a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda \end{cases}$$

so that the first derivative takes the form

$$p'_\lambda(|\beta_j|) = \lambda\left\{\mathbb{I}(|\beta_j| < \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda}\mathbb{I}(|\beta_j| > \lambda)\right\}$$

- EN (Zou & Hastie (2005)):

$$p_\lambda(|\beta_j|) = \lambda_1|\beta_j| + \lambda_2\beta_j^2$$

- HARD (Antoniadis (1997)):

$$p_\lambda(|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2\mathbb{I}(|\beta_j| < \lambda)$$

A penalty function $p_\lambda(.)$ serves its purpose if the corresponding MPLE possesses the sparsity property (sets true zero coefficients to zero) and behaves like the MLE for large samples. While LASSO exhibits sparsity, it does not behave like the MLE when $n \to \infty$. Fan & Li (2001) introduced the SCAD penalty function to avoid these deficiencies. The HARD penalty is important because it corresponds to the best subset selection in certain cases and it exhibits the oracle properties i.e., it satisfies the sparsity property and asymptotic normality simultaneously. Finally, Zou & Hastie (2005) presented the Elastic Net (EN) penalty which is a mixture of LASSO and RIDGE to effectively select grouped variables in large datasets.

The remainder of this paper is organized as follows. In Section 3.3, we show analytically how the variable selection can reduce the variance of the estimator. Section 3.4 introduces a penalized (pseudo-) causal likelihood and defines the penalized estimator corresponding to the derived likelihood. In Section 3.5, we study the large sample behaviour of the MPLE. The performance of the proposed method is studied via simulation in Section 3.5. We analyze the NSWD data in Section 3.6. The last section contains some concluding remarks.

### 3.3 An Analytic Example: Propensity Score Regression

As a simple example, assume the true structural conditional mean model is given by

$$\mathbb{E}[Y|D = d, \mathbf{X} = \mathbf{x}] = \beta d + \beta_1 x_1 + \beta_2 x_2,$$

with propensity score model $\text{logit}(\pi(x)) = \alpha_1 x_1$ where $X_1$ and $X_2$ are continuous covariates and $\beta$ is the causal parameter. Using the PS regression approach, the following model will be fitted

$$\mathbb{E}[Y|D = d, \mathbf{X} = \mathbf{x}] = \beta d + \theta\pi(\mathbf{x}), \tag{3.1}$$

where $\mathbf{X}$ is the vector of covariates.

In this regression setting, the usual sum of squares decomposition applies; the total sum of squares ($SST$) can be decomposed into the residual sum of squares ($SSE$)

and the regression sum of squares ($SSR$), $SST = SSE + SSR$. Adding $X_2$ into the fitted propensity score model in equation (3.1) increases the covariance between the modelled $\pi$ and $Y$, which decreases $SSE$, and hence increases $SSR$ which is indeed $\mathbb{V}\text{ar}[D|\mathbf{X}]$ in our setting. Therefore, it decreases the variance of the estimated causal effect, which explains the results obtained by Brookhart et al. (2006a). Clearly, adding $X_2$ to the propensity score model will not harm the balancing property and thus the estimator remains unbiased.

Conversely, consider the effect of including a spurious covariate in the propensity score model, that is, a covariate that is **not** related to $Y$ (even if it is related to $D$). Suppose for simplicity the data generating model has

$$\mathbb{E}[Y|D = d, \mathbf{X} = \mathbf{x}] = \beta d + \beta_1 x_1$$

and suppose that the fitted conditional model for $Y$ is

$$\mathbb{E}[Y|D = d, \mathbf{X} = \mathbf{x}] = \beta d + \theta\pi(\mathbf{x})$$

as before. For simplicity of exposition, suppose $\pi(\mathbf{X})$ is constructed using a linear (rather than logistic) model, with $\pi(\mathbf{x}) = \mathbb{E}[D|\mathbf{X} = \mathbf{x}] = \alpha_1 x_1 + \alpha_2 x_2$, with $D$ modelled using additive zero mean errors, $D = \pi(\mathbf{x}) + \epsilon_1$ say. In reduced form, substituting the model for $\pi$ into the conditional mean model for $Y$, the fitted model for $Y$ given $\mathbf{X} = \mathbf{x}$ is

$$Y = \beta d + \theta\pi(\mathbf{x}) + \epsilon = \beta d + \theta(\alpha_1 x_1 + \alpha_2 x_2) + \epsilon = \beta d + \beta_1^\star x_1 + \beta_2^\star x_2 + \epsilon,$$

say. Thus, comparing this with the data generating model, $X_2$ is a spurious covariate unless $\alpha_2 = 0$. Write the fitted model

$$Y = [Z_1 \; Z_2][\gamma_1 \; \gamma_2]' + \epsilon$$

where $Z_1 = \texttt{cbind}(D, X_1)$ and $Z_2 = X_2$, $\gamma_1(\beta, \beta_1^\star)$, $\gamma_2 = \beta_2^\star$. The OLS estimator of $[\gamma_1 \; \gamma_2]'$ is straightforward to derive and is an unbiased estimator of $\gamma_1 = (\beta, \beta_1^\star)$, and the estimator of $\gamma_2 = \beta_2^\star$ has expectation zero. The variance of the OLS estimator is

$$\mathbb{V}\mathrm{ar}[\widehat{\gamma}] = \mathbb{V}\mathrm{ar}_{Z_1,Z_2}\left[\mathbb{E}[\widehat{\gamma}|Z_1, Z_2]\right] + \mathbb{E}_{Z_1,Z_2}\left[\mathbb{V}\mathrm{ar}[\widehat{\gamma}|Z_1, Z_2]\right]$$

From above, the first term is zero. In the second term, $\mathbb{V}\mathrm{ar}[Y|Z_1, Z_2]\sigma^2\mathbf{1}_n$, so

$$\mathbb{V}\mathrm{ar}[\widehat{\gamma}|Z_1, Z_2] = \sigma^2 \begin{bmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}M\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}M \\ -M\Sigma_{21}\Sigma_{11}^{-1} & M \end{bmatrix}$$

where $M = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$, $\Sigma_{ij} = Z_i'Z_j$. Thus the variance of $\widehat{\gamma}_1$ is the $(1,1)$ element after taking expectations, if $\mathbb{E}_{Z_1}\left[\Sigma_{11}^{-1}\right]$ is the variance if $Z_2 \equiv X_2$ *is omitted from the model*,

$$\mathbb{E}_{Z_1,Z_2}\left[\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}M\Sigma_{21}\Sigma_{11}^{-1}\right] \geq \mathbb{E}_{Z_1}\left[\Sigma_{11}^{-1}\right]$$

in general. The inequality holds as $M$ is positive definite, as is $\Sigma_{12}M\Sigma_{21}$. Therefore, the inclusion of $X_2$ inflates the variance of the estimator of $\beta$. The exception to this case is where $\Sigma_{12} = Z_1'Z_2$ is near zero, or when $X_2$ is uncorrelated with $D$ and $X_1$. If $X_2$ is a predictor of $D$, it certainly is correlated with $D$.

These simple examples suggest that if our aim is to decrease the mean square error ($MSE$) of the estimator, we should include in the propensity score model those variables that are just related to the response, and not merely to the treatment mechanism. It also suggests that the common strategy of building rich (saturated) models for the propensity score will lead to inefficiency if covariates that are not confounders are included. Excluding confounders, of course, leads to bias and inconsistency; however, the impact on $MSE$ is difficult to quantify from these calculations.

### 3.4 Penalized Estimation for Causal Parameters

In the spirit of Robins et al. (1992), we proposed a slight modification of the conventional PSR. We define the conditional response mean model as follows:

$$\mathbb{E}[Y_i|S_i = s_i, \mathbf{X}_i = \mathbf{x}_i] = \beta_0 + \beta_1 S_i + g(\mathbf{x}; \alpha),$$

where $S_i = D_i - \mathbb{E}[D_i|\mathbf{x}_i] = D_i - \pi(\mathbf{x}_i)$, $g(\mathbf{x}; \alpha)$ is a function of covariates and $\pi$ is the propensity score. The quantity $S_i$ is used in the mean model in place of $D_i$; if $D_i$ is used the fitted model may result in a biased estimator for $\beta_1$ since $g(\mathbf{x})$ might not be the true function of covariates in the mean model. Fitting the incorrect $g(.)$ would not affect the consistency of the exposure effect if the treatment assignment were randomized, when $\text{cor}[D, \mathbf{X}_j] = 0$ for $j = 1, 2, .., p$. By defining $S_i = D_i - \pi(\mathbf{x}_i)$, we restore $\text{cor}[S, X_j] = 0$ for $j = 1, 2, .., p$ if $\pi(\mathbf{x}_i) = \mathbb{E}[D_i|\mathbf{x}_i]$ is correctly specified, as $\pi(\mathbf{x}_i)$ is the (fitted) expected value of $D_i$, and hence $\mathbf{x}_j'(D - \pi(\mathbf{x})) = 0$. Therefore misspecification of $g(.)$ will not result in an inconsistent estimator.

In general, this model is doubly robust since it results in a consistent exposure effect estimator $\beta_1$ if either the propensity score or the functional form of the association between the outcome and the covariates $g(.)$ is correctly specified. In our model, we use a linear function of covariates as $g(.)$. When the propensity score and the posited response mean models are correctly specified, it results in the most efficient estimator (Tsiatis (2006)). Some other forms of DR estimators can be found in Davidian et al. (2005), Schafer & Kang (2005) and Bang & Robins (2005). For additional details on the asymptotic and the small sample behaviour of DR estimators see Kang & Schafer (2007), Neugebauer & van der Laan (2005), van der Laan & Robins (2003) and Robins (1999).

### 3.4.1 Likelihood construction

In likelihood-based penalized estimation, a proper likelihood is required. Consider the following parametric likelihood (based on a specific parametric submodel)

$$\begin{aligned}\mathcal{L}(\eta; \mathbf{y}, \mathbf{d}, \mathbf{x}) &= \prod_{i=1}^{n} \{f_1(y_i|x_i, \beta)p(d=1|x_i, \alpha)f(x_i)\}^{d_i} \\ &\qquad \{f_0(y_i|x_i, \beta)(p(d=0|x_i, \alpha))f(x_i)\}^{1-d_i} \\ &= \prod_{i=1}^{n} \{f_1(y_i|\pi_i, g(x; \alpha), \beta)p(d=1|x_i, \alpha)\}^{d_i} \\ &\qquad \{f_0(y_i|\pi_i, g(x; \alpha), \beta)p(d=0|x_i, \alpha)\}^{1-d_i} f(x_i).\end{aligned}$$

In general, the counterfactual densities $f_j(y_i|\pi_i, g(x; \alpha), \beta)$, $j = 1, 2$, are unknown. Using the assumption of no unmeasured confounders, this density can be replaced by the observed conditional density, $f(y_i|D = j, \pi_i, g(x;\alpha), \beta)$,

$$\mathcal{L}(\eta; \mathbf{y}, \mathbf{d}, \mathbf{x}) \propto \prod_{i=1}^{n} f(y_i|D = d, \pi_i, g(x;\alpha), \beta) p(d = 1|x_i, \alpha)^{d_i} p(d = 0|x_i, \alpha)^{1-d_i}, \quad (3.2)$$

where $\beta$ is an $r_1$-dimensional vector and $\alpha$ is an $r_2$-dimensional vector containing parameters that appear in the model for $Y|X$ and $D|X$ respectively. Thus $\eta = (\beta, \alpha)$ is an $r$-dimensional vector, where $r = r_1 + r_2$. Since our goal is to select covariates for the propensity score model, we just impose a penalty on the propensity score parameters. Therefore, using (3.2) the penalized log likelihood can be written as

$$\tilde{l}_n(\eta) = l_{ny}(\eta) + l_{nd}(\eta) - np_{\lambda_n}(\alpha) \quad (3.3)$$

where

$$l_{ny}(\eta) = \sum_{i=1}^{n} \log f(y_i|d_i, \pi_i, g(x;\alpha), \beta)$$

$$l_{nd}(\eta) = \sum_{i=1}^{n} \left\{ d_i \log \frac{p(d = 1|x_i, \alpha)}{p(d = 0|x_i, \alpha)} + \log p(d = 0|x_i, \alpha) \right\}.$$

Thus, the penalized (pseudo-) density for each sample point is

$$f_p(z_i, \eta) = f(z_i; \eta) f(\alpha),$$

where $Z_i = (y_i, d_i, \mathbf{x}_i)$ and $f(\alpha) = \exp\{-p_{\lambda_n}(\alpha)\}$. The maximum penalized likelihood estimator, $\hat{\eta}_n$, is defined by

$$\hat{\eta}_n = \arg\sup_{\eta} \prod_{i=1}^{n} f_p(z_i; \eta) = \arg\sup_{\eta} \sum_{i=1}^{n} \log f_p(z_i; \eta).$$

Since $\alpha$ appears in both models $Y|X$ and $D|X$, not just the model $D|X$, its estimate is not the same as the parameters in the PS model. We do not require differentiability of the likelihood.

### 3.4.2 Omission of confounders

Missing confounders which are weakly related to the outcome (treatment) but strongly related to the treatment (outcome) may impose bias in the estimation of causal effect. Thus, any covariate selection technique must take into account both outcome and treatment assignment models simultaneously. By using the joint likelihood, we give the covariates a chance to represent their contribution twice, once in the conditional mean model and once in the treatment assignment model. This strategy then gives an equal chance to Type-I and Type-III covariates for selection as key covariates. To deal with this problem, we introduce the *boosting* parameter $\nu$ which boosts covariates Type-III relative to Type-I. The boosting parameter can be defined as $\nu = 1/|\tilde{\alpha}_Y|$, where $\tilde{\alpha}_Y$ is the least-squares estimate of the parameters in the response mean model. Our penalty function is proportional to the boosting parameter, $p_{\lambda_n}(.) = \nu p^*_{\lambda_n}(.)$, where $p^*_{\lambda_n}(.)$ is the conventional penalty function. Therefore, the magnitude of penalty on each parameter is proportional to its contribution to the response and treatment model. A similar argument can be found in the adaptive LASSO (Zou (2006)).

### 3.4.3 Main Theorems

The following conditions on penalty functions guarantee a consistent penalized estimating procedure which sets the small coefficients to zero for covariate selection.

P1. For all $n$, $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\alpha)$ is symmetric and nonnegative. Moreover, it is twice differentiable with derivatives $p'_{\lambda_n}(.)$ and $p''_{\lambda_n}(.)$ and nondecreasing for $\alpha \neq 0$ with at most a few exceptions.

P2. As $n \to \infty$, $\max_{\alpha \neq 0}\{p''_{\lambda_n}(\alpha)\} \to 0$ and $\max_{\alpha \neq 0}\{\sqrt{n}p'_{\lambda_n}(\alpha)\} \to 0$.

P3. For $N_n \equiv (0, M_n)$, $\lim \inf_{\alpha \in N_n} \sqrt{n}p'_{\lambda_n}(\alpha) = \infty$, where $M_n \to 0$ as $n \to \infty$.

Assumption P1 is used to prove theorem 3.1, while P2 prevents the $j$th element of the penalized likelihood from being dominated by the penalty function since it vanishes when $n \to \infty$. If $\alpha_j = 0$, condition P3 allows the penalty function to dominate the penalized likelihood which leads to the sparsity property (Khalili & Chen (2007)).

Without loss of generality, we reorder the $r$-dimensional vector of parameters, $\eta$, to $\eta = (\eta_1, \eta_2)$ such that $\eta_2 = (\eta_j)$ for $j = s, ..., r$ corresponds to the zero coefficients. The true parameter values denoted by $\eta_0 = (\eta_{01}, 0)$. Note that since there is no penalty on the $\beta$s, $\eta_2$ consists of those $\alpha$ that should be shrunk to zero ($\alpha_j = 0$ for $j = s', ..., r_2$). The following Theorem states the existence of the consistent penalized maximum likelihood estimator under certain conditions. It assumes certain regularity conditions (C1-C4) that are common in the study of the asymptotic behaviour of likelihood-based estimates; see Appendix B, and Ibragimov & Has' Minskii (1981). Proofs are given in the Appendix.

**Theorem 3.1** *Suppose assumptions C1-C4 and P1-P2 are fulfilled. There exists a penalized maximum likelihood estimator $\hat{\eta}_n \to \eta_0$ as $n \to \infty$ almost surely.*

Lemma 5.2 in the Appendix shows that the proposed penalized estimation technique is able to detect zero coefficients and shrink them to zero through the penalized maximum likelihood estimation procedure. In the next theorem, we show that the proposed method asymptotically sets the detected zero coefficients to zero.

**Theorem 3.2** *Under conditions C1-C4 and P1-P3, $Pr(\hat{\eta}_2 = 0) \to 1$ as $n \to \infty$.*

The next result presents asymptotic normality of the score function. Let $I(\eta)$ be the Fisher information matrix, where

$$I(\eta) = \int_{z:f(z,\eta)\neq 0} \left[\frac{\partial \log f(z,\eta)}{\partial \eta}\right] \left[\frac{\partial \log f(z,\eta)}{\partial \eta}\right]' f(z,\eta)dz$$

**Theorem 3.3** *Suppose assumptions C1-C5 and P1-P3 are fulfilled and further* $\det[I(\eta)] \neq 0$ *for* $\eta \in \Xi$*. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\partial \log f(z,\eta_{01})}{\partial \eta_{01}}\right] - \sqrt{n} p'_{\lambda_n}(\alpha_{01}) \xrightarrow{d} N(0, I(\eta_{01}))$$

*where $\eta_{01} = (\beta, \alpha_{01})$ and $\alpha_{01}$ is the true vector of non-zero coefficients.*

The proof is omitted since it is similar to the proof of Theorem 2.1.1 in Ibragimov and Hes' Minskii (1981).

The asymptotic distribution of the MPL estimator is presented in the following Corollary which is the direct result of Theorem 3.3 and Theorem 2.5.2 in Bickel et al. (1993b).

**Corollary 3.4** *Under the assumptions of Theorem 3.3, we have*

$$\sqrt{n}(\widehat{\eta}_{01} - \eta_{01}) \xrightarrow{d} N(0, I^{-1}(\eta_{01})).$$

Corollary 3.4 states that asymptotically, under $P_2$ the penalized estimator is as efficient as the ML estimator when the important covariates are known a priori.

**Remark 1.** By Theorem 3.3, the estimator obtained by the LASSO penalty function will be a $\sqrt{n}$ consistent estimator for non-zero coefficients if $\sqrt{n}\lambda_n \to 0$, whereas it achieves the sparsity property if $\sqrt{n}\lambda_n \to \infty$. Therefore, the LASSO penalty function cannot satisfy both sparsity and consistency simultaneously. However, in adaptive LASSO, the sparsity property will be achieved if $n\lambda_n \to \infty$; therefore, it exhibits the oracle properties. For the HARD and SCAD penalties, P2 and P3 hold if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$.

### 3.4.4 Choosing the Tuning Parameter

Choosing the right tuning parameter $\lambda$ is essential in penalized likelihood estimation. Small values of $\lambda$ result in an overfitted model, while large values can set important coefficients to zero. We select the tuning parameter using the *Generalized Cross Validation* (GCV) method suggested by Tibshirani (1996) and Fan & Li (2001). The GCV measure is defined as

$$GCV(\lambda) = \frac{RSS(\lambda)/n}{\{1 - d(\lambda)/n\}^2},$$

where $RSS(\lambda) = ||y - \hat{\beta}_0 - \hat{\beta}_1 S - \mathbf{X}\hat{\alpha}||^2$, $d(\lambda) = \text{trace}\{\mathbf{X}(\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\hat{\eta}))^{-1})\mathbf{X}^T\}$ is the effective number of parameters and

$$\Sigma_\lambda(\eta) = \text{diag}\{p'_\lambda(|\eta_1|)/|\eta_1|, ..., p'_\lambda(|\eta_p|)/|\eta_p|\},$$

The selected tuning parameter $\hat{\lambda}$ is defined by $\hat{\lambda} = \arg\min_\lambda GCV(\lambda)$.

### 3.4.5 Estimation Procedure

The penalized exposure effect estimation process can be summarized as follows:

I. Fit a saturated (that is, rich) propensity score model by including all the variables in the model $\pi(\mathbf{X})$. This is to ensure no confounders are omitted, and to attempt to ensure that $\pi(\mathbf{X}) = \mathbb{E}[D|\mathbf{X}]$ is correctly specified.

II. Define a new random variable $S_i = D_i - \pi(\mathbf{X})$ and fit the response conditional mean model

$$\mathbb{E}[Y_i|d, \mathbf{x}] = \beta_0 + \beta_1 s_i + g(\mathbf{x}; \alpha),$$

where $g(\mathbf{x}; \alpha) = \alpha'\mathbf{x}$ and $\mathbf{x}$ is the whole vector of covariates. A richer conditional mean model can also be fitted by including the interaction and/or higher order terms of $g(.)$.

III. Estimate the vector of parameters $\eta = (\beta, \alpha)$ such that

$$\hat{\eta}_n = \arg\sup_{\eta} \prod_{i=1}^{n} f_p(y_i, s_i, \pi_i; \eta) = \arg\sup_{\eta} \sum_{i=1}^{n} \log f_p(y_i, s_i, \pi_i; \eta),$$

where $f_p(.)$ is the joint penalized density of $(Y, S, \pi)$. Note that in practice, the penalized counterfactual density $f_{jp}(y|\pi)$ for $j = 0, 1$ is unknown. Thus, it needs to be replaced by the observed penalized conditional density $f_p(y_i|s_i, \pi_i; \hat{\eta}_n)$.

We develop a one-step estimation procedure; parameters of the response and the function $g(.)$ are estimated simultaneously in one step (III).

## 3.5 Simulation Studies

### 3.5.1 Simulation Study: The joint penalized likelihood approach

In this section, we represent a simulation study based on 1000 data sets of sizes 500 and 5000 to check the performance of the proposed penalized covariate selection

Table 3–1: Penalized ATE estimators based on the SCAD and LASSO penalty functions.

| $\sigma = 2$ | | $n = 500$ | | | $n = 5000$ | | |
|---|---|---|---|---|---|---|---|
| Method | $\rho$ | Bias | S.D. | MSE | Bias | S.D. | MSE |
| SCAD | 0 | 0.012 | 0.230 | 0.053 | 0.000 | 0.065 | 0.004 |
| | 0.5 | 0.003 | 0.233 | 0.055 | 0.005 | 0.064 | 0.004 |
| LASSO | 0 | 0.003 | 0.254 | 0.065 | 0.002 | 0.065 | 0.004 |
| | 0.5 | 0.023 | 0.246 | 0.061 | 0.005 | 0.064 | 0.004 |
| ORACLE | 0 | 0.006 | 0.195 | 0.038 | 0.000 | 0.065 | 0.004 |
| | 0.5 | 0.010 | 0.187 | 0.035 | 0.001 | 0.064 | 0.004 |
| Saturated | 0 | 0.001 | 0.326 | 0.106 | 0.002 | 0.123 | 0.015 |
| | 0.5 | 0.031 | 0.329 | 0.109 | 0.003 | 0.125 | 0.016 |
| $\sigma = 4$ | | $n = 500$ | | | $n = 5000$ | | |
| SCAD | 0 | 0.025 | 0.541 | 0.29 | 0.006 | 0.127 | 0.016 |
| | 0.5 | 0.009 | 0.537 | 0.288 | 0.007 | 0.124 | 0.016 |
| LASSO | 0 | 0.055 | 0.618 | 0.385 | 0.003 | 0.127 | 0.016 |
| 0.5 | | 0.090 | 0.562 | 0.316 | 0.008 | 0.124 | 0.015 |
| ORACLE | 0 | 0.021 | 0.387 | 0.150 | 0.008 | 0.127 | 0.016 |
| | 0.5 | 0.012 | 0.372 | 0.139 | 0.002 | 0.121 | 0.015 |
| Saturated | 0 | 0.039 | 0.899 | 0.810 | 0.008 | 0.332 | 0.111 |
| | 0.5 | 0.026 | 0.902 | 0.814 | 0.014 | 0.343 | 0.118 |

method. We assume the following model:

$$D \sim \text{Bernoulli}\left(\frac{\exp\{0.2x_1 - x_2 + 3x_8 - x_9 + x_{10}\}}{1 + \exp\{0.2x_1 - x_2 + 3x_8 - x_9 + x_{10}\}}\right)$$

$$Y \sim \text{Normal}(d + 2x_1 + 0.5x_2 + 5x_3 + 5x_4, \sigma)$$

where $X_k$ for $k = 1, ..., 10$ are normal random variables with parameters $(\mu = 1, \sigma, \rho)$. Table 3–1 summarizes the results based on the LASSO and SCAD penalty functions for two sets of different values of standard deviation ($\sigma$) and correlation ($\rho$). ORACLE refers to the proposed model by Brookhart et al. (2006a) which includes just covariates related to the response in the propensity score model and "Saturated"

Table 3–2: Penalized ATE estimators based on the SCAD and LASSO penalty functions.

| Method | $\rho$ | MRME | Correct | Incorrect | MRME | Correct | Incorrect |
|---|---|---|---|---|---|---|---|
| $\sigma = 2$ | | $n = 500$ | | | $n = 5000$ | | |
| SCAD | 0 | 0.583 | 5.5 | 0.0 | 0.401 | 6 | 0 |
| | 0.5 | 0.596 | 5.4 | 0.0 | 0.413 | 6 | 0 |
| LASSO | 0 | 0.622 | 5.2 | 0.0 | 0.394 | 6 | 0 |
| | 0.5 | 0.538 | 5.3 | 0.0 | 0.316 | 6 | 0 |
| $\sigma = 4$ | | $n = 500$ | | | $n = 5000$ | | |
| SCAD | 0 | 0.355 | 5.1 | 0.03 | 0.240 | 6 | 0 |
| | 0.5 | 0.346 | 5.2 | 0.03 | 0.215 | 6 | 0 |
| LASSO | 0 | 0.377 | 5.1 | 0.02 | 0.220 | 6 | 0 |
| | 0.5 | 0.339 | 5.2 | 0.00 | 0.212 | 6 | 0 |

refers to the propensity score model including all the covariates $(X_1, ..., X_{10})$. The fitted propensity score model in ORACLE and Saturated are

$$\pi(\mathbf{x}) = p(D = 1|x_1, x_2, x_3, x_4),$$

$$\pi(\mathbf{x}) = p(D = 1|x_1, x_2, ..., x_{10}).$$

Table 3–1 confirms that our proposed technique converges to the ORACLE model as sample size increases. For moderate sample size, the covariate selection technique results in a lower MSE than the Saturated model, decreasing both bias and standard error. In fact, the standard error of the treatment effect obtained by the saturated model is roughly twice that obtained using the penalized technique. Table 3–2 summarizes the performance of LASSO and SCAD in terms of the median of relative model errors (MRME). The MRME is defined as the median of the following quantity

$$RME = \frac{(\hat{\beta} - \beta)^T \mathbb{E}[\mathbf{X}'\mathbf{X}](\hat{\beta} - \beta)}{(\hat{\beta}_{sat} - \beta)^T \mathbb{E}[\mathbf{X}'\mathbf{X}](\hat{\beta}_{sat} - \beta)},$$

where $\hat{\beta}$ is the estimator obtained from the penalized method and $\hat{\beta}_{sat}$ is the one obtained from the saturated model. The average number of coefficients set to zero correctly or incorrectly are also reported in Table 3–2. For moderate sample size, SCAD outperforms LASSO when covariates are independent, while LASSO performs better when there is correlation between covariates. However, for large sample size, LASSO outperforms SCAD in terms of the MRME for both correlated and uncorrelated covariates and both noise levels. In both cases, increasing the variance from 2 to 4 with moderate sample size reduces the MRME.

### 3.5.2 Simulation Study: The joint and conditional penalized likelihood approaches

Here, we want to compare the proposed joint penalized likelihood approach with the conditional response penalized likelihood approach. Our simulation study in this section reveals that missing a confounder which is weakly related to the response but strongly related to the treatment may lead to an inconsistent estimator. This type of confounder can be ignored by applying the penalized covariate selection methods. Assume the following data generating process

$$\begin{aligned} D &\sim \text{Bernoulli}\left(\frac{\exp\{0.2x_1 - 2x_2 + 3x_8 - x_9 + x_{10}\}}{1+\exp\{0.2x_1 - 2x_2 + 3x_8 - x_9 + x_{10}\}}\right) \\ Y &\sim \text{Normal}(d + 2x_1 + 0.05x_2 + 5x_3 + 5x_4, \sigma), \end{aligned}$$

where $X_k$ for $k = 1, ..., 10$ are independent normal random variables with parameters $(\mu = 1, \sigma = 4)$.

Table 3–3: Comparing the joint and conditional penalized likelihood approaches. LASSO[1] represents the results based on the proposed penalized joint likelihood approach. LASSO[2] represents the results based on the conditional response penalized likelihood approach using the adaptive LASSO.

| Method | Bias | S.D | MSE | Bias | S.D | MSE |
|---|---|---|---|---|---|---|
| $\sigma = 4$ | | $n = 5000$ | | | $n = 30000$ | |
| LASSO[1] | 0.003 | 0.122 | 0.015 | 0.002 | 0.050 | 0.002 |
| LASSO[2] | 0.121 | 0.117 | 0.028 | 0.118 | 0.047 | 0.016 |
| Saturated | 0.014 | 0.288 | 0.083 | 0.002 | 0.119 | 0.014 |

Table 3–3 indicates that missing $x_2$ which is weakly related to the response produces a bias and as sample size increases, the bias dominates the variance, which results in an inconsistent estimator.

### 3.6 Application to Real Data

The National Supported Work Demonstration (NSWD) aimed to provide work experience to individuals who lacked basic skills. In 1970, a total of 6,616 workers were sampled at random from different cities across the United States by the Manpower Demonstration Research Corporation (MDRC). Qualified individuals were assigned to a treatment group to receive training provided by the NSW program. Some baseline variables such as age ($X_1$), education ($X_2$), African American ($X_3$), Hispanic ($X_4$), married status ($X_5$), no degree ($X_6$) and earnings in 1975 ($X_7$) were recorded for each participant. As in LaLonde (1986), we focus on the male workers among the participants and consider their earnings in 1978 as the response variable. Our treated observations group consists of a subset of the NSW with 297 workers

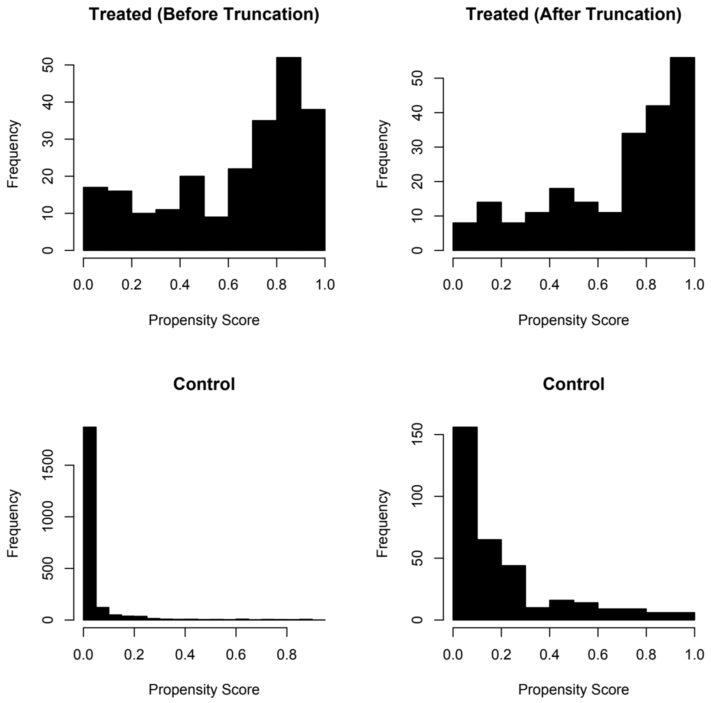


Figure 3–2: The NSWD data: Estimated propensity score overlap in treated and control groups before and after truncation.

and the control group includes individuals from the Panel Study of Income Dynamics (PSID) with 2490 workers. Additional details are provided in Dehejia & Wahba (1999).

The initial propensity score is estimated using logistic regression including all the covariates. The first step in the propensity score causal adjustment method is to check for sufficient overlap between treatment and control groups (Rubin (2008)). The two plots on the left side of Figure 3–2 reveal that there are some subclasses (bins) that include only treated or control individuals which may result in a poor

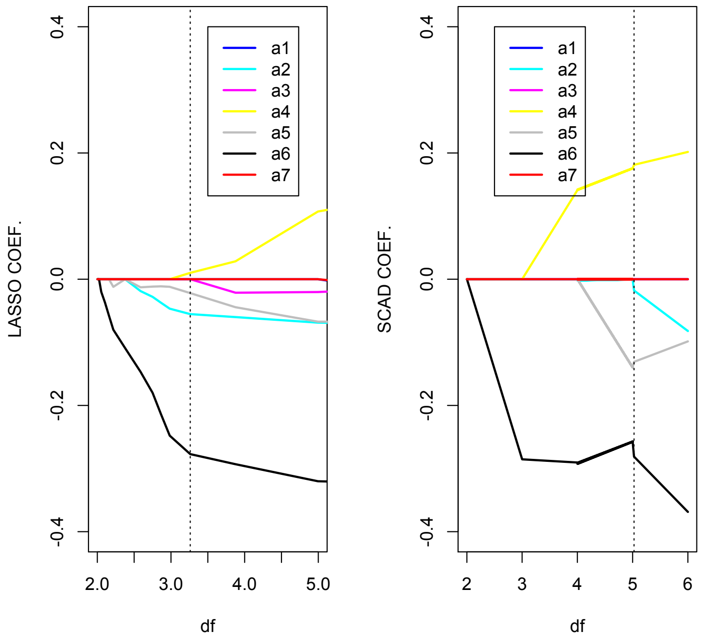


Figure 3–3: The NSW data: The left and right plots are the estimated coefficients using the LASSO and SCAD, respectively. The horizontal axis represents the effective number of coefficients (df).

causal effect estimate. More precisely, the treatment effect can not be estimated in those empty bins without some response model assumptions relating the counterfactuals to the covariates. To overcome this problem, we truncated those individuals who fall in empty bins. The right side of Figure 3–2 shows that sufficient overlap can be achieved by truncating the empty bins. Therefore, there are both treated and untreated workers in all the propensity score subclasses, which means that in each

Table 3–4: Penalized ATE estimators based on the SCAD and LASSO penalty functions.

| Method | ATE | S.D. | C.I. | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|---|---|---|
| SCAD | -0.552 | 0.091 | (-0.734,-0.370) | – | $\surd$ | – | $\surd$ | $\surd$ | $\surd$ | – |
| LASSO | -0.565 | 0.094 | (-0.753,-0.377) | – | $\surd$ | – | $\surd$ | $\surd$ | $\surd$ | – |
| Saturated | -0.502 | 0.107 | (-0.716,-0.288) | $\surd$ | $\surd$ | $\surd$ | $\surd$ | $\surd$ | $\surd$ | $\surd$ |
| Unadjusted | -1.376 | 0.072 | (-1.520,-1.232) | – | – | – | – | – | – | – |

bin the distributions of all the covariates are nearly the same for both treated and untreated workers.

After ensuring that the randomized experiments can be reconstructed inside each propensity subclass, we need to find the conditional expectation $\mathbb{E}[Y(j)|d, \pi(\mathbf{x})]$. Heckman et al. (1997) and Hardie & Linton (1994) suggested the use of nonparametric techniques to find the conditional expectation of the response given treatment and the propensity score. Here, we fit the following regression model for the observable response

$$Y = \beta_0 + \beta_1 s + g(\mathbf{x}; \alpha),$$

where $\beta_1$ is the treatment effect parameter. The interaction or the higher order of the propensity score can be added to the response model if needed.

Rubin (1997) stressed the importance of including all the confounders in the propensity score model. Accordingly, if there is any variable which is related to the response and not to the treatment, it will be excluded from the treatment mechanism model. In other words, Rubin believes that the propensity score should be used as a diagnostic tool for design rather than analysis and we should not look at the response variable while fitting a model for the propensity score. In their simulation studies

Brookhart & Van Der Laan (2006), however, suggested that by adding covariates that are related only to the response and not treatment, we can gain efficiency.

We applied our penalized technique to the NSWD data, with results as summarized in Table 3–4. *Saturated* and *unadjusted* rows in this Table refer to the propensity score model including all the variables and none of the variables, respectively. Figure 3–3 shows how estimated coefficients vary by changing the tuning parameter. Note that the maximum(minimum) effective number of coefficients (df) corresponds to $\lambda = 0(1)$. The constructed penalized covariate selection technique keeps "Education", "Hispanic", "Married Status" and "No degree" as the only significant confounders in the propensity score model. Although we do not have a large number of covariates in this example, our variable selection technique results in more than 27% variance reduction compared with the saturated model.

### 3.7 Future Directions

In this Chapter, we use a regression model to estimate the treatment effect by modelling the conditional expectation of the response given treatment and the propensity score. We establish a Maximum Penalized Likelihood Estimator (MPLE) which satisfies the oracle properties under certain assumptions. The proposed technique involves the treatment and the outcome models simultaneously such that it does not ignore covariates which are just related to the outcome and not to the treatment. Our covariate selection approach reduce the variance while attaining no residual confounding.

Although we have discussed the time point treatment case, the penalized technique can be used to estimate the direct treatment effect in longitudinal setting as well. However, in longitudinal setting, if the total effect is the parameter of interest other causal adjustment techniques such as IPTW need to be used. One of the drawbacks of propensity score regression based models is that any nonlinear term in the true response model has to be added in the propensity score model to ensure the consistency of the treatment effect estimator which is not easily detectable in some cases. Moreover, it is well-known that in the presence of time varying-confounders and treatments PSR adjustment may result in a biased estimator. In particular, when there is a unmeasured covariate which causes the outcome and a confounder, conditioning on the confounder results in a biased estimator of the treatment. In this case, the IPTW technique is a great alternative to solve the problem. So, adopting our proposed technique for longitudinal setting can be very helpful to reduce the bias of the estimator obtained by the IPTW. It seems reasonable to apply the cofounder selection method for the baseline covariates. It might, however, be risky to apply any covariate selection technique on the time varying confounders because of the possible causal path complications.

In general, confounding might not be the only source of bias in the estimation of the treatment effect. In some cases, we do not have a representative sample of the target population and it needs to be considered as another source of bias. In the next Chapter, we consider one of well-known biased sampling cases called length-biased sampling and present a weighted estimating equation which adjust for the length-biased sampling as well as the confounding effect for estimating the treatment effect on the survival time.

# CHAPTER 4
# Double Bias

## Chapter Summary

Biased sampling is often exercised when logistic or other constraints preclude the possibility of having a representative sample from the population of interest. It is, in particular, a common phenomenon in observational studies on disease duration when recruiting incident cases is infeasible. A second type of bias is a bias induced when we use classified sampling and the goal is to compare a response variable between different subpopulations. An example of such is encountered when comparing survival with dementia among institutionalized elderly citizens and those recruited from the community. From a slightly different perspective, this scenario may be viewed as estimating a causal effect of treatment when in addition to the lack of treatment randomization, the sample we is not a representative sample from the target population. While there is a vast literature addressing these two types of bias separately, there is no methodology to address them both simultaneously. We introduce two estimating equations, a weighted and a double robust for estimating grouping effect that can handle both types of bias. Large sample properties of the estimators are established and then small sample behaviour is studied using simulations. We apply the proposed method to a set of prevalent cohort survival data collected as a part of the Canadian Study of Health and Aging (CSHA) to compare survival with dementia among institutionalized patients versus those recruited from the community.

## 4.1 Introduction

Survival or failure time data typically comprise an initiating event, say onset of a disease, and a terminating event, say death. In an ideal situation, recruited subjects have not experienced the initiating event, the so called incident cases. These cases are then followed to a terminating event, say death, or censoring, say loss to follow up. It frequently happens, however, that recruiting the incident cases is infeasible, due to logistic or other constraints. As such, subjects who have experienced the initiating event prior to the start of the study, so called prevalent cases, are recruited. It is well-known that these cases tend to have a longer survival time, and hence they do not constitute a representative sample from the population of interest, the target population. When the initiating events are generated by a stationary Poisson process, the induced bias is called length-biased (Cox & Lewis (1966), Zelen & Feinlein (1969)). In other words, the probability of being observed for individuals in the population is proportional to their survival time.

Studies on length-biased sampling can be traced as far back as Wicksell (1925) and his corpuscles problem. Such phenomenon was later noticed by Fisher (1934) in his article on methods of ascertainment. Neyman (1955) discussed length-biased sampling further and coined the terminology incidence prevalence bias. Cox (1969) studied length-biased sampling in industrial applications and quality control, while Zelen & Feinlein (1969) observed the same bias in screening test for disease prevalence. Patil & Rao (1978) provide several other interesting examples of length-biased sampling. Vardi and his collaborators systematically studied nonparametric statistical inference when data are subject to length-biased sampling (Vardi (1982), Vardi

(1985), Gill et al. (1988), Vardi (1989), Vardi & Zhang (1992), and Gilbert et al. (1999) ). Asgharian et al. (2002) and Asgharian & Wolfson (2005) studied NPMLE from right censored prevalent cohort survival data. More recently, Shen et al. (2009) and Qin & Shen (2010) studied analysis of covariates under biased sampling.

Length-biased sampling can affect the sampling distribution of the covariates such that those covariates which are related to the longer survival have a higher chance of being selected. Bergeron et al. (2008) pointed out this issue and suggested the joint likelihood estimation based rather than the conventional conditional approach if either the the failure time distribution is parametrically specified or the covariate distribution in the target population is known. Note that when the sample is representative of the population, the marginal distribution of the covariates does not contain any information about the relationship between the survival and the covariates. As such, joint likelihood estimation does not add any extra information to the estimation procedure and it performs as well as conditional inference. Another source of bias often encountered in observational studies is the nonrandomized grouping assignment which is discussed in detail in Chapter 1.

To adjust for both types of bias simultaneously, we propose a modified version of the estimating equations introduced by Shen et al. (2009). This modified version in conjunction with IPTW or PSR can adjust the two types of bias discussed above. The IPTW based estimating equation does not need the response mean model to be correctly specified and it results in a consistent estimator if the grouping assignment model, the conditional probability of being in a group given the covariates, is correctly specified. The PSR estimating equation; however, needs both grouping and the response mean model to be correctly specified to estimate the grouping effect consistently. This stronger modelling assumption results in an estimator which has a lower variance than the IPTW method.

## 4.2 Preliminaries

### 4.2.1 Notations

Let $T^{pop}$ be the time measured from the onset to failure time in the population with distribution $F_U$, and $T^{obs}$ be the same measured time for observed subjects with distribution $F_{LB}$. It is well-known that

$$F_{LB}(t) = \frac{1}{\mu} \int_0^t s dF_U(s),$$

where $\mu = \int_0^\infty s dF_U(s)$ if the onset times are generated by a stationary Poisson process (the so-called stationarity assumption). The observed event time r. v., $T^{obs}$, can be decomposed into the time from the onset of the disease to the recruitment time, say $A$, and the residual life time which covers the time from recruitment to the event, $R$, also called backward and forward recurrence times, respectively. When the individuals are also subject to right-censoring, the observed survival time is $Y = A + min(R, C)$ where $C$ is the censoring time measured from the recruitment to loss to follow up.

### 4.2.2 The Likelihood Based Approach

In the propensity score regression adjustment technique, the propensity score plays the role of the whole vector of covariates, and as noted by Bergeron et al. (2008), the left-truncation affects the distribution of the covariates/propensity score as well as the conditional distribution of the response. For example, if those who have a higher chance of being in the control group live longer then they will be over represented in the length-biased sample. Under the uniform truncation and the independency of $T^{pop}$ and $A$, it can be shown that

$$f_{\pi}^{LBj}(\pi) = \frac{\mu_j(\pi(\mathbf{x}), \theta) f_{\pi}(\pi)}{\mu_j(\theta)}$$

where $\mu_j(\pi(\mathbf{x}), \theta) = \mathbb{E}[T^{pop}(j)|\pi]$ and $\mu_j(\theta) = \mathbb{E}[\mu_j(\pi(\mathbf{x}), \theta)]$ for $j = 0, 1$. Hence, when the individuals are subject to left-truncation and right censoring, the first part of the (2.5) can be written as (Bergeron et al. (2008));

$$E^{-1}[T^{pop}(1)] = \mu_1(\theta)^{-1} = \int \frac{f_{1LB}(t|\pi, \theta)}{t} f_{\pi}^{LB1}(\pi)\, dt\, d\pi,$$

where $f_{1LB}(.)$ is the length-biased conditional density of the survival time given covariates if treated(Bergeron et al. (2008)).

The joint parametric likelihood of $(d_i, y_i, \pi_i)$ for $i = 1, ..., n$ is given by

$$\mathcal{L}(\theta; \mathbf{y}, \mathbf{d}, \pi) = \prod_{i=1}^{n} \{f_1(y_i|\pi_i, \beta) p(d_i|\pi, \alpha) f_{\pi}(\pi_i)\}^{d_i} \{f_0(y_i|\pi_i, \beta) p(d_i|\pi, \alpha) f_{\pi}(\pi_i)\}^{(1-d_i)}, \tag{4.1}$$

where $f_0$ and $f_1$ are the conditional densities of the variable response in the untreated and treated groups respectively, and under no unmeasured confounders assumption it can be written as

$$\mathcal{L}(\theta; \mathbf{y}, \mathbf{d}, \mathbf{x}) \propto \prod_{i=1}^{n} f(y_i|D = d, \pi_i, \beta) p(d = 1|x_i, \alpha)^{d_i} p(d = 0|x_i, \alpha)^{1-d_i}. \tag{4.2}$$

where $\theta = (\beta, \alpha)$.

Therefore, when observations are subject to length-biased sampling and right-censoring, we have

$$\mathcal{L}_{LB}(\theta; \mathbf{w}, \mathbf{d}, \pi) = \prod_{i=1}^{n} f(w_i, d, \pi_i; \theta), \tag{4.3}$$

where $w = (a, r, \delta)$ and

$$\begin{aligned} f(w, d, \pi; \theta) &= \left[\frac{f_U(a + r|d, \pi, \theta)}{\mu(\pi, \theta)}\right]^{\delta} \left[\frac{S_U(a + c|d, \pi, \theta)}{\mu(\pi, \theta)}\right]^{1-\delta} \\ &\qquad \times \left[\frac{\mu(\pi, \theta) f(d, \pi)}{p(T^{\text{pop}} \geq A, \pi)}\right] \left[\frac{p(T^{\text{pop}} \geq A, \pi)}{\mu(\theta)}\right] \\ &= \left[\frac{f_U(a + r|d, \pi, \theta)}{\mu(\theta)}\right]^{\delta} \left[\frac{S(a + c|d, \pi, \theta)}{\mu(\theta)}\right]^{1-\delta} f(d, \pi). \end{aligned}$$

Note that by uniformity of the left-truncation

$$p_B(d = 1|\pi, \theta) = p(d = 1|\pi, T^{pop} \geq A, \theta) = \frac{p(T^{pop} \geq A|D = 1, \pi, \theta) f(d, \pi)}{p(T^{pop} \geq A, \pi, \theta)},$$

$$f_B(\pi, \theta) = f(\pi|T^{pop} \geq A, \theta) = \frac{p(T^{pop} \geq A, \pi, \theta)}{\mu(\theta)}.$$

where $\mu(\theta) = \mathbb{E}[\mathbb{E}[T^{\text{pop}}|D, \pi]]$ and $\mu(D, \pi, \theta) = \mathbb{E}[T^{\text{pop}}|D, \pi]$. The vector of parameters of the derived joint likelihood can be estimated by MLE. Often, the joint density is unknown which makes the likelihood-based approach not possible. To cope with this problem, in the next section, we introduce estimating equation based approaches which do not depend on the information about the joint density function.

## 4.3 Accelerated Failure Time (AFT) Models

Inspired by AFT models introduced by Cox & Oakes (1984), we consider a general form of AFT models when we do not assume any functional form for the association of the confounders with the *log* scaled outcome nor do we assume a known error distribution. We, therefore, have the following model

$$\log(T^{pop}) = \beta D + g(\mathbf{x}) + \epsilon, \qquad \mathbb{E}[\epsilon | D, \mathbf{X}] = 0. \tag{4.4}$$

where $\epsilon$ has a unknown distribution. One may replace the function $g(.)$ by the conditional expectation of the exposure given the confounders, $\pi(\mathbf{X})$, and fit the following model

$$\log(T^{pop}) = \beta D + \theta \pi(\mathbf{X}) + \epsilon.$$

We refer to this model as the AFT Propensity Score Regression Model (AFTPSR). Robins et al. (1992) discuss about the possible efficiency loss in the estimation of treatment effect, $\beta$, when the $g(.)$ is replaced by the propensity score. The higher order and interaction terms can also be included in the model if needed. While AFT models may suffer from lack of robustness w.r.t. the transformation, being the logarithmic function, they are more advantageous than other models such as transformation models when parameter interpretability is the concern (Kalbfleisch & Prentice (1980)).

In the next section, we introduce a weighted estimating equation which estimate the treatment effect without specifying any functional form for the association of the confounders and the log response.

### 4.3.1 Weighted Estimating Equation

Recall that inverse probability weighting is efficient when no restriction is imposed on the joint distribution of $(Y, D, X)$ (see, Rotnitzky et al. (2010)). This estimator adjusts for the bias induced by confounders using appropriate weights, it associates to each observed outcome.

We generalize the IPTW estimator to account for length-biased sampling as well as non-randomization. In our setting, those weights reflect the inverse of the chance of being in the group that the individuals actually belong to. Unlike the conventional estimating equations in AFT models, our Weighted Estimating Equation (WEE) does not rely on the correct model specification of the response mean model to estimate the treatment effect consistently.

Let

$$w(y) = \int_0^y S_C(s)\,ds,$$

where $S_C(s)$ is the survivor function of the censoring variable. Assuming that censoring time is independent of the covariates and under the conditional independence assumption of the exposure and the counterfactual response given measured covariates among uncensored subjects, the class of influence functions corresponding to the equation for estimating the grouping effect when $\pi(\mathbf{X})$ and $w(Y)$ are known is given by

$$\mathcal{G}_1 = \left\{ \frac{\delta}{w(Y)} \left[ \frac{D\log(Y)}{\pi} - \frac{(1-D)\log(Y)}{(1-\pi)} - \beta \right] \right\}, \tag{4.5}$$

This is a modified version of the influence function introduced by Tsiatis (2006). We define

$$M(s) = \mathbb{1}(Y - A < s, \delta = 0) - \int_0^s \mathbb{1}(Y - A > u)\,d\Lambda_C(u),$$

where $\Lambda_C(.)$ is the cumulative hazard function of the censoring variable. $M(s)$ can be estimated by replacing the $\Lambda_C(.)$ by its estimate, $\widehat{\Lambda}_C(.)$.

It is well known (Hirano et al. (2003)) that estimating $\pi(\mathbf{x})$ using a parametric model results in an efficiency gain in estimating the treatment effect. The class of influence functions when $w(Y)$ is known and $\pi(\mathbf{x})$ is replaced by $\widehat{\pi}(\mathbf{x})$ is

$$\mathcal{G}_2^{(\mathrm{AFT})} = \left\{ \frac{\delta}{w(Y)} \left[ \frac{D[\log(Y) - \mu_1(\mathbf{X},\theta)]}{\widehat{\pi}(\mathbf{X})} - \frac{(1-D)[\log(Y) - \mu_0(\mathbf{X},\theta)]}{(1-\widehat{\pi}(\mathbf{X}))} \right.\right.$$
$$\left.\left. + \mu_1(\mathbf{X},\theta) - \mu_0(\mathbf{X},\theta) - \beta \right] \right\}, \qquad (4.6)$$

where $\mu_j(\mathbf{x},\theta)$ for $j = 0, 1$ is the posited response mean model for untreated and treated, respectively. The causal effect estimator corresponding to $\mathcal{G}_2$ is called a *Double Robust* (DR) estimator in the sense that $\mathcal{G}_2$ results in a consistent estimator if either the propensity score or the conditional response mean models are correctly specified (see Tsiatis (2006), Neugebauer & van der Laan (2005), Robins (1999) and van der Laan & Robins (2003)). When the propensity score and the posited response mean models are correctly specified, the $\mathcal{G}_2$ results in the most efficient estimator (Tsiatis (2006)).

### 4.3.2 Asymptotic Properties of the WEE estimator

The following theorem presents the asymptotic properties of the double robust treatment effect estimator obtained by (4.6) in the presence of length-biased sampling using the AFT models when both the treatment assignment and $w(.)$ are replaced by their estimated values.

**Theorem 4.1** *Let $\widehat{\beta}_{DR}^{AFT}$ be a double robust estimator corresponding to the class of influence functions $\mathcal{G}_2^{(AFT)}$. Then under regularity conditions $C.1 - C.5$,*

$$n^{1/2}(\widehat{\beta}_{DR}^{AFT} - \beta) \xrightarrow{d} \mathcal{N}(0, \eta(\theta)),$$

*where*

$$\eta(\theta) = \mathbb{E}\{\widehat{V}_0^2(\theta) + \widehat{V}_1^2(\theta) + \widehat{V}_2^2(\theta) + \widehat{V}_0(\theta)\widehat{V}_2(\theta) + \widehat{V}_1(\theta)\widehat{V}_2(\theta)\}$$

*and*

$$\begin{aligned}
\widehat{V}_j(\theta) &= \mathbb{1}(d=j)\delta\left[\frac{\log(Y)-\mu_j(\mathbf{X},\theta)}{p(D=j|\widehat{\pi}(\mathbf{X}))w(Y)}\right] + \int_0^Y \frac{\kappa_j(t)}{S_C(t)S_R(t)}dM(t), \quad j=0,1 \\
\kappa_j(t) &= \mathbb{E}\left\{\frac{\mathbb{1}(D=j)\delta\mathbb{1}(Y>t)[\log(Y)-\mu_j(\mathbf{X},\theta)]\int_t^Y S_C(v)dv}{p(D=j|\widehat{\pi}(\mathbf{X}))w^2(Y)}\right\} \\
\widehat{V}_2(\theta) &= \frac{\delta}{w(Y)}[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta] + \int_0^Y \frac{\kappa(t)}{S_C(t)S_R(t)}dM_i(t), \\
\kappa(t) &= \mathbb{E}\left\{\frac{\delta\mathbb{1}(Y>t)[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta]\int_t^Y S_C(v)dv}{w^2(Y)}\right\}.
\end{aligned}$$

### 4.3.3 Propensity Score Estimation

In most of cases, the propensity score is unknown and it needs to be estimated. It has been also shown that even if the propensity score is known, one may gain efficiency by estimating it using the data available (Hirano et al. (2003)). In this section, we construct an unbiased estimating equation for estimating the parameters of the propensity score, $\alpha$.

We have that if $f_U(y|\mathbf{X}, D)$ and $F_U(d|\mathbf{X})$ are the unbiased conditional density of the survival time given the covariates and the unbiased distribution of the exposure given the covariates, and

$$\mu(\theta, \mathbf{X}) = \int p(T^{\text{pop}} \geq a|\mathbf{X}, \theta)\, da,$$

then

$$\begin{aligned}
\mathbb{E}\left[\delta\frac{(D-\pi(\mathbf{X},\alpha))}{w(Y)}\;\middle|\;\mathbf{X}\right] &= \mathbb{E}\left[\mathbb{E}\left\{\delta\frac{(D-\pi(\mathbf{X},\alpha))}{w(Y)}|\mathbf{D},\mathbf{X}\right\}\right] \\
&= \int[D-\pi(\mathbf{X},\alpha)]\int\frac{f_U(y|\mathbf{X},D)w(y)}{w(y)\mu(\theta,D,\mathbf{X})}\,dy \\
&\qquad\qquad\times\frac{\mu(\theta,D,\mathbf{X})dF_U(D|\mathbf{X})}{\mu(\theta,\mathbf{X})} \\
&= \frac{1}{\mu(\theta,\mathbf{X})}\int[D-\pi(\mathbf{X},\alpha)]F_U(D|\mathbf{X}) \\
&= 0.
\end{aligned}$$

The last equality holds, since $f_U(y|\mathbf{X},D)$ is a proper density and

$$\pi(\mathbf{X},\alpha)=\int d\,dF_U(d|\mathbf{X}).$$

An unbiased estimating equation for $\alpha$ is therefore

$$\sum_{i=1}^{n}\delta_i\mathbf{x}_i\frac{(d_i-\pi(\mathbf{x}_i,\alpha))}{w(y_i)}=0,$$

which is equivalent to the weighted logistic regression among the uncensored subjects. In the simulation studies and the real data analysis, we use the above estimating equation to estimate the parameters of the treatment assignment model.

## 4.4 Simulation Studies

In the this section, we conduct a simulation study to examine the performance of the proposed estimating equations under the accelerated failure time. We simulate 1000 datasets consisting of 200, 400 and 800 observations to study the performance of the proposed estimating equations for estimating the unmediated causal effect.

We generated the failure times using the following model,

$$\log(T^{pop}) = 2.5d + \frac{x_2}{1 + 2x_2 + x_1} + \exp\{-x_1/2\} - 3dx_2 + \epsilon,$$

where $\epsilon$ is uniformly distributed on (-1,1), $X_1$ is uniformly distributed on (0,1), $X_2$ is a Bernoulli random variable with 0.5, and

$$D \sim \text{Bernouli}\left(\frac{\exp\{2 - x_1 - 3x_2\}}{1 + \exp\{2 - x_1 - 3x_2\}}\right).$$

The estimated treatment effects and their standard errors are listed in Table 4–1. We consider three different unadjusted scenarios: Unadjusted$^{lc}$ is an estimator for which neither the length-biased nor the non-randomization is adjusted, Unadjusted$^{c}$ is obtained by adjusting for the length-biased sampling but the non-randomization left unadjusted, and Unadjusted$^{l}$ is carried out by adjusting for the non-randomization while the length-biased sampling left unadjusted. We have used a correct conditional mean model in the DR estimating equation. The DR estimator dominates the two other estimators in terms of the standard deviation and the MSE. Increasing the censoring proportion, increases the bias in the PSR, IPTW and DR estimators while maintaining the unbiasedness. As we expected all the *unadjusted* estimators are biased and in our parameter setting it seems that the failure to account for the length-biased sampling leads to a more biased estimator comparing to the Unadjusted$^{c}$.

## 4.5 Real Data Analysis

The Canadian Study of Health and Aging (CSHA), initiated in 1989, is a nationwide study on aging. One of the objects of CSHA is to study dementia in Canada.

The CSHA included phases in 1991, 1996 and 2001. In the first phase, 10,263 individuals aged 65 or over were sampled at random across Canada, from both rural and urban areas, from communities and institutions for the elderly. Among the participants, 1,132 people were diagnosed with dementia. The ages of dementia onset were assessed from each individual's medical history. We analyze the data collected during the first phase of the study which began in 1991 by sampling prevalent cases and examining their types of dementia mainly, probable Alzheimer's disease, possible Alzheimer's disease and vascular dementia. The age of death or censoring were recorded for each subject from the time of screening, while the of onset was ascertained retrospectively using CAMDEX from care givers (Wolfson et al. (2001)).

One of the collected covariates is the dichotomous institutionalization (exposure) indicator which is one if being institutionalized at the time of sampling and zero otherwise. Since there are some covariates which confound the effect of the exposure on the survival time, the crude difference estimator will be biased. The challenge is to estimate the institutionalization effect on the survival time while having confounding and length-biased sampling as two sources of estimator bias. Our data includes 818 subjects after excluding patients with missing information; of which 180 subjects were right censored (Wolfson et al. (2001)). The validity of stationarity assumption has been verified by Addona & Wolfson (2006) and Asgharian et al. (2006).

Table 4–2 presents the estimated institution effect on the survival time using different estimating equations proposed in this Chapter under the AFT modelling assumptions. Similar to our simulation study, we also consider three different unadjusted scenarios: Unadjusted$^{lc}$, Unadjusted$^{c}$ and Unadjusted$^{l}$. The results reveal that the institutionalization has a positive effect on the survival time at the 10% level while the *unadjusted* estimators show a small negative effect. In other words, without adjusting for either the length-biased sampling or the treatment adjustment, we might conclude institutionalized subjects may tend to have shorter survival time.

### 4.6 Future Directions

We introduce a set of weighted estimating equations based on the AFT models which obtain an unbiased estimator of the exposure effect in the presence of length-biased sampling without assuming any functional form for the association of the confounders and the outcome. This method can be generalized to the longitudinal setting in the presence of time-varying confounders. Hernan et al. (2005) introduce a method called Structural Accelerated Failure Time Models (SAFTM) which accounts for time-dependent confounders and treatments (Robins (1992)). They, however, assum that the samples are representative sample of the target population. The SAFTM can be generalized to adjust for the length-biased sampling as well as the time-varying confounders.

In AFT models, we assume that the survival time is linearly related to the covariates under log transformation. Cheng et al. (1995) weaken this assumption by introducing the transformation models which leave the transformation function completely unknown. In particular, they assume that there exists an unknown function increasing function, $g(.)$, of the population failure times which is linearly related to the vector of covariates. Specifically,

$$g(T^{pop}) = \theta_1 D + \theta_2 \mathbf{X} + \epsilon$$

where the distribution of $\epsilon$ is known. Shen et al. (2009) introduce an estimating equation for estimating based on the transformation models assumption which account for the length-biased sampling. Adopting our proposed weighted estimating equation for the transformation models seems very helpful. The collapsibility issues, however, may arise under some transformation functions; that is, the marginal association of the exposure effect with the survival time may not be the same as conditional association given the covariates (Greenland et al. (1999), Gail et al. (1984) and Gail (1986)).

Table 4–1: Accelerated failure time simulation study. Unadjusted$^{lc}$: neither the length-biased nor the non-randomization are adjusted for. Unadjusted$^{c}$: The length-biased is adjusted whereas non-randomization left unadjusted. Unadjusted$^{l}$: The non-randomization is adjusted whereas the length-biased left unadjusted.

| Censored(%)=0 | $n = 200$ | | | $n = 400$ | | | $n = 800$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Bias | S.D. | MSE | Bias | S.D. | MSE | Bias | S.D. | MSE |
| PSR | 0.023 | 0.246 | 0.061 | 0.004 | 0.182 | 0.033 | 0.001 | 0.134 | 0.018 |
| DR | 0.014 | 0.203 | 0.042 | 0.003 | 0.150 | 0.022 | 0.001 | 0.109 | 0.012 |
| IPW | 0.007 | 0.220 | 0.048 | 0.009 | 0.170 | 0.029 | 0.008 | 0.123 | 0.015 |
| Unadjusted$^{lc}$ | 0.730 | 0.107 | 0.544 | 0.738 | 0.078 | 0.551 | 0.740 | 0.057 | 0.550 |
| Unadjusted$^{c}$ | 0.354 | 0.194 | 0.163 | 0.366 | 0.142 | 0.154 | 0.368 | 0.105 | 0.147 |
| Unadjusted$^{l}$ | 0.511 | 0.122 | 0.277 | 0.525 | 0.089 | 0.284 | 0.529 | 0.062 | 0.284 |
| Censored(%)=20 | $n = 200$ | | | $n = 400$ | | | $n = 800$ | | |
| PSR | 0.018 | 0.226 | 0.051 | 0.011 | 0.164 | 0.027 | 0.011 | 0.113 | 0.013 |
| DR | 0.022 | 0.219 | 0.049 | 0.012 | 0.158 | 0.025 | 0.013 | 0.110 | 0.012 |
| IPW | 0.002 | 0.245 | 0.060 | 0.003 | 0.176 | 0.031 | 0.005 | 0.125 | 0.016 |
| Unadjusted$^{lc}$ | 0.713 | 0.127 | 0.525 | 0.722 | 0.089 | 0.529 | 0.724 | 0.062 | 0.529 |
| Unadjusted$^{c}$ | 0.344 | 0.214 | 0.164 | 0.355 | 0.150 | 0.148 | 0.348 | 0.107 | 0.132 |
| Unadjusted$^{l}$ | 0.490 | 0.141 | 0.260 | 0.503 | 0.098 | 0.262 | 0.512 | 0.067 | 0.269 |
| Censored(%)=30 | $n = 200$ | | | $n = 400$ | | | $n = 800$ | | |
| PSR | 0.060 | 0.223 | 0.053 | 0.057 | 0.166 | 0.031 | 0.052 | 0.116 | 0.016 |
| DR | 0.069 | 0.216 | 0.051 | 0.063 | 0.161 | 0.030 | 0.058 | 0.113 | 0.016 |
| IPW | 0.046 | 0.230 | 0.054 | 0.044 | 0.180 | 0.034 | 0.043 | 0.132 | 0.019 |
| Unadjusted$^{lc}$ | 0.556 | 0.128 | 0.326 | 0.556 | 0.095 | 0.318 | 0.565 | 0.064 | 0.323 |
| Unadjusted$^{c}$ | 0.287 | 0.207 | 0.126 | 0.279 | 0.161 | 0.104 | 0.288 | 0.110 | 0.095 |
| Unadjusted$^{l}$ | 0.333 | 0.145 | 0.132 | 0.349 | 0.103 | 0.132 | 0.360 | 0.069 | 0.135 |

Table 4–2: CSHA data analysis: Estimation of the institutionalization effect on the survival time. Unadjusted$^{lc}$: neither the length-biased nor the non-randomization are adjusted for. Unadjusted$^{c}$: The length-biased is adjusted whereas non-randomization left unadjusted. Unadjusted$^{l}$: The non-randomization is adjusted whereas the length-biased left unadjusted.

| | | AFT | | |
|---|---|---|---|---|
| Method | Inst. Eff. | S.D. | CI 95% | CI 90% |
| PSR | 0.185 | 0.106 | (-0.027 , 0.397) | (0.011 , 0.359) |
| DR | 0.192 | 0.121 | (-0.050 , 0.434) | (-0.006 , 0.390) |
| IPW | 0.208 | 0.125 | (-0.042 , 0.458) | (0.003 , 0.413) |
| Unadjusted$^{lc}$ | -0.212 | 0.074 | (-0.360, -0.064) | (-0.333 , -0.090) |
| Unadjusted$^{c}$ | -0.079 | 0.085 | (-0.249 , 0.091) | (-0.218 , 0.060) |
| Unadjusted$^{l}$ | 0.084 | 0.103 | (-0.122 , 0.290) | (-0.085 , 0.252) |

# CHAPTER 5
# Concluding Remarks

Drawing inference about the possible effect of a treatment must often be done without the benefits of randomization. As a result, one needs to take into account the potential for bias in the estimation of the treatment effect. In this thesis, we focus on *Propensity Score Regression* (PSR) as a tool to adjust for the confounding bias where propensity score replaces the whole vector of covariates. In the PSR method, the exposure effect can be estimated by regressing the response variable on the exposure and the fitted propensity score. It is not hard to show that under the assumption of no unmeasured confounders the PSR method results in a consistent estimator.

Two other well-known causal adjustment methods are *Inverse Probability Treatment Weighting* (IPTW) and *Augmented Inverse Probability Weighed Complete Case* (AIPWCC). In the first Chapter, we derived the semiparametric variance bound for the estimated causal effect using propensity score regression adjustment, and showed that the obtained bound is equal to the efficiency bound introduced in the literature on semiparametric regression. A parametric model is assumed for the propensity score and the parameters of this model are treated as nuisance parameters. The nuisance tangent space is built based on the parameters in the treatment mechanism model. Using the theory of semiparametric inference, the efficient influence function corresponding to the estimator of interest has been constructed as a residual from projecting any influence function onto the nuisance tangent space. We compared the obtained efficiency bound with those already exist in the literature on IPTW and AIPWCC, and showed that the variance of the PSR estimator is lower than the one obtained by the IPTW or AIPWCC. We described an alternative approach to the classical binary treatment propensity score termed the Generalized Propensity Score (GPS) and extended the semiparametric result to continuous-valued treatments. We also discussed the treatment effect estimation in a multi-interval setting.

In longitudinal studies the covariates can vary during time since they depend on the previous assigned treatment. The treatment effect is decomposed to direct and indirect effect in such studies. Roughly speaking, the indirect effect is the effect of the treatment on the outcome through other time varying covariates and the direct effect is the effect of treatment which goes directly to the response. We conducted extensive simulation studies to assess the performance of the three causal effect estimators in longitudinal setting for estimating the magnitude of the direct effect of treatment. Our simulation studies reveal that the direct treatment effect estimator using propensity score regression adjustment has a smaller variance and is more successful in removing bias than corresponding methods that use weighting, under correct model specification.

In causal inference framework, choosing the right covariates to adjust for is a challenging problem which is discussed in Chapter 2. In general, ignoring important covariates leads to residual confounding and results in an inconsistent estimator. It, therefore, seems beneficial to adjust for all the covariates by fitting a rich propensity score model. Including both related and unrelated covariates in the treatment assignment model may, however, affect the efficiency of important covariates and inflate the variance of the estimators. Confounder selection methods based on either the propensity score or the response model may result in failure to account for important confounders which barely predict the treatment or the response, respectively. Moreover, it has been shown that the confounder selection procedures based on $AIC$ and $BIC$ can be sub-optimal. To overcome these issues, we presented a penalization technique based on the joint likelihood of the treatment and response variables to select the key covariates that need to be included in the treatment assignment/PS model for estimation of the causal effect. We further developed the appropriate methodology and studied the theoretical properties of the proposed method and showed that it satisfies the oracle properties under certain assumptions. Our simulation studies reveals that taking the confounder selection approach can in practice lead to significant efficiency gain. Although we discussed the time point treatment case, the penalized technique can be used to select the key baseline covariates for estimating the direct treatment effect in a longitudinal setting as well. However, in a longitudinal setting if the total effect is the parameter of interest, other causal adjustment techniques such as IPTW in a marginal structural models need to be used. In general, in multi-interval settings, selecting time-varying confounders can be quite involved and needs to be carried out cautiously because of the complicated confounding patterns.

In Chapters 1-3, we assumed that samples are representing the target population. In many cases, however, our sample is not a representative sample of the population of interest. Biased sampling is, in particular, a common phenomenon in observational studies on disease duration where the recruited subjects are prevalence cases. In particular, when the initiating events are generated by a stationary Poisson process, the induced bias is called length-bias. In length-biased sampling the observed individuals are more likely to be sampled from the right tail of the true density function. As such, leaving it unadjusted may result in overestimating the survival time. Chapter 4 is motivated by the Canadian Study of Health and Aging (CSHA) data to study dementia across Canada. One of the collected covariates is the dichotomous institutionalization (exposure) indicator which is one if being institutionalized at the time of sampling and zero otherwise. Since there are some covariates which confound the effect of the exposure on the survival time, the crude difference estimator will be biased. The challenge is to estimate the institutionalization effect on the survival time while having confounding and length-biased sampling as two sources of bias. We have focused on the stationary case, i.e., length-biased sampling, the methodology presented in this manuscript can be extended to the general left-truncation using Wang (1991) and references cited therein. This latter approach is, of course, robust against departure from stationarity, though it is less efficient when the stationarity assumption holds (Wang (1991), Asgharian et al. (2002)). We presented two estimating equations called weighting and double robust estimating equations. Unlike the regression-based version, the weighted estimating equation does not require the correct specification of the failure time model to estimate the exposure effect consistently; it just requires that the PS model be correctly specified. The double robust estimating equation obtains the consistent estimator if either the failure time or the PS models are correctly specified. We also studied the asymptotic properties of the estimators obtained by the estimating equations. The proposed estimating equations are applied to the CSHA data to estimate the institutionalization effect on the survival time of patients with dementia. Our real data analysis results highlights the importance of adjusting for the two sources of bias. Omitting either the length-biased sampling or the non-randomization may lead to a misleading results.

# Appendix A

## 5.1 Differentiability of the Causal Effect

Following Newey (1990), we term the parameter $\mu(\beta)$ a *differentiable parameter* if there exists a random quantity $\xi$ such that $\mathbb{E}[\xi^2]$ is finite and for all parametric submodels

$$\frac{\partial\mu(\beta)}{\partial\beta} = \mathbb{E}[\xi(D, Y, \pi)S_\beta(D, Y, \pi|\beta)].$$

where $S_\beta$ is the $(k \times 1)$ score function in the parametric submodel based on $\theta$

$$S_\beta(D, Y, \pi|\beta) = \frac{\partial}{\partial\beta}\{\log f(Y|D, \pi, \beta)\}$$

These definitions extend to the case of a $p$-dimensional differentiable parameter, but the scalar case suffices here, as $\mu(\beta)$ represents the causal parameter.

Consider the function $\xi(D_i, Y_i, t)$ defined by

$$\xi(d, y, t) = \frac{d}{t}(y - \mu_1(t)) - \frac{1-d}{1-t}(y - \mu_0(t))$$

and the function

$$S_\beta(d, y, t|\beta) = dS_{1\beta}(y|t, \beta) + (1-d)S_{0\beta}(y|t, \beta) \tag{5.1}$$

where, for $j = 0, 1$,

$$S_{j\beta}(y|t, \beta) = \frac{\partial \log f_j(y|\beta)}{\partial\beta}$$

are the score functions derived from the density $f_j$ of subjects that received ($j = 1$) and did not receive ($j = 0$) treatment. Now

$$\begin{aligned}\xi(d, y, t)S_\beta(d, y, t|\beta) &= \frac{d^2}{t}(y - \mu_1(t))S_{1\beta}(y|t, \beta) - \frac{(1-d)^2}{(1-t)}(y - \mu_0(t))S_{0\beta}(y|t, \beta) \\ &\quad + \frac{d(1-d)}{1-t}(y - \mu_1(t))S_{0\beta}(y|t, \beta) - \frac{d(1-d)}{t}(y - \mu_0(t))S_{1\beta}(y|t, \beta)\end{aligned}$$

Given $\pi$, $D$ and $Y$ are independent by weak unconfoundedness. By elementary calculation,

$$\mathbb{E}[D^2|\pi] = \pi \qquad \mathbb{E}[(1-D)^2|\pi] = (1-\pi) \qquad \mathbb{E}[D(1-D)] = 0.$$

Thus, conditional on $\pi$,

$$\begin{aligned}\mathbb{E}\left[\frac{D(1-D)}{1-\pi}(Y - \mu_1(\pi))S_{0\beta}(Y|\pi, \beta)\middle|\pi\right] &= \mathbb{E}\left[\frac{D(1-D)}{\pi}(Y - \mu_0(\pi))S_{1\beta}(Y|\pi, \beta)\middle|\pi\right] \\ &= 0\end{aligned}$$

so therefore

$$\begin{aligned}\mathbb{E}[\xi(D, Y, \pi)S_\beta(D, Y, \pi|\beta)] = \mathbb{E}\Bigg[&\frac{D^2}{\pi}(Y - \mu_1(\pi))S_{1\beta}(Y|\pi, \beta) \\ &- \frac{(1-D)^2}{(1-\pi)}(Y - \mu_0(\pi))S_{0\beta}(Y|\pi, \beta)\Bigg]\end{aligned}$$

In addition,

$$\begin{aligned}\mathbb{E}[(Y - \mu_1(\pi))S_{1\beta}(Y|\pi, \beta)|\pi] &= \int (y - \mu_1(\pi))S_{1\beta}(y|\pi, \beta)f_1(y|\pi, \beta)\, dy \\ \mathbb{E}[(Y - \mu_0(\pi))S_{0\beta}(Y|\pi, \beta)|\pi] &= \int (y - \mu_0(\pi))S_{0\beta}(y|\pi, \beta)f_0(y|\pi, \beta)\, dy\end{aligned}$$

Now, note from (2.5) that

$$\frac{\partial\mu(\beta)}{\partial\beta} = \int\int yS_{1\beta}(y|\pi,\beta)f_1(y|\pi,\beta)f_\pi(\pi)\,dy\,d\pi \\ - \int\int yS_{0\beta}(y|\pi,\beta)f_0(y|\pi,\beta)f_\pi(\pi)\,dy\,d\pi$$

and hence

$$\mathbb{E}[\xi(D,Y,\pi)S_\beta(D,Y,\pi|\beta)] = \frac{\partial\mu(\beta)}{\partial\beta}$$

as, by the usual manipulation, for $j = 0, 1$,

$$\int \mu_j(\pi)S_j(y|\pi,\beta)f_j(y|\pi,\beta)\,dy = \mu_j(\pi)\frac{\partial}{\partial\beta}\left\{\int f_j(y|\pi,\beta)\,dy\right\} = 0.$$

Thus $\mu(\beta)$ is a differentiable parameter; this ensures that the efficient score can be constructed for $\mu(\beta)$, using the techniques outlined in Newey (1990), and hence that the semiparametric efficiency bound is well-defined and can be achieved.

The likelihood that corresponds to (5.1) can be obtained by undoing the construction of the score function. It is proportional to

$$\{f_1(y|\pi,\beta)p(d|\pi,\alpha)\}^d\{f_0(y|\pi,\beta)p(d|\pi,\alpha)\}^{(1-d)}$$

This likelihood forms the basis for our subsequent efficient influence function. In fact, to cover the most general setting where $f_\pi$ is not known, we will use the likelihood

$$\{f_1(y|\pi,\beta)p(d|\pi,\alpha)f_\pi(\pi)\}^d\{f_0(y|\pi,\beta)p(d|\pi,\alpha)f_\pi(\pi)\}^{(1-d)} \qquad (5.2)$$

and allow for the possibility that $f_\pi(\pi)$ contains unknown parameters.

## 5.2 Deriving the Efficient Score

Here we show that

$$\mathbb{E}[S_\beta \epsilon|\pi] = \frac{\partial \mu^*(\pi,\beta)}{\partial \beta} = \mu^*_\beta(\pi,\beta) = d\mu_{1\beta}(\pi,\beta) + (1-d)\mu_{0\beta}(\pi,\beta),$$

say. Let $S_\beta = dS_{1\beta} + (1-d)S_{0\beta}$ and $\epsilon = d\epsilon_1 + (1-d)\epsilon_0$. As $\mathbb{E}[\epsilon] = \mathbb{E}[\epsilon|\pi] = \int \epsilon \, dF(y|d,\pi) = 0$, we have

$$\frac{\partial}{\partial \beta}\int \epsilon \, f(y|d,\pi)dy = 0$$

where $f(y|d,\pi) = \{f_1(y|\pi,\beta)\}^d\{f_0(y|\pi,\beta)\}^{(1-d)}$ is the conditional density of $Y$ given $d$ and $\pi$. Thus

$$\int \epsilon f(y|d,\pi)\, dy = \int d\epsilon_1 f_1(y|\pi,\beta)\, dy + \int (1-d)\epsilon_0 f_0(y|\pi,\beta)\, dy = 0$$

hence

$$\begin{aligned}\frac{\partial}{\partial \beta}\int \epsilon \, dy &= \frac{\partial}{\partial \beta}\int d\epsilon_1 f_1(y|\pi,\beta)\, dy + \frac{\partial}{\partial \beta}\int (1-d)\epsilon_0 f_0(y|\pi,\beta)\, dy \\ &= d\int -\mu_{1\theta}(\pi,\beta) f_1(y|\pi,\beta)\, dy + (1-d)\int -\mu_{0\beta}(\pi,\beta) f_0(y|\pi,\beta)\, dy \\ &+ d\int \epsilon_1 S_{1\beta} f_1(y|\pi,\beta)\, dy + (1-d)\int \epsilon_0 S_{0\beta} f_0(y|\pi,\beta)\, dy = 0.\end{aligned}$$

Therefore,

$$\begin{aligned}d\mu_{1\beta}(\pi,\beta) + (1-d)\mu_{0\beta}(\pi,\beta) &= d\int \epsilon_1 S_{1\beta} f_1(y|\pi,\beta)\, dy + (1-d)\int \epsilon_0 S_{0\beta} f_0(y|\pi,\beta)\, dy \\ &= \mathbb{E}[S_\beta \epsilon|d,\pi].\end{aligned}$$

## 5.3 Propensity Score Stratification

In propensity score stratification, we have

$$\mathbb{E}[Y|D,\pi] = \beta D + \sum_{k=1}^{K} \theta_k \mathbb{I}_{B_k}(\pi)$$

where $B_1, \ldots, B_K$ form a (presumed fixed) partition of $(0,1)$. In the above notation, we have

$$\mu_1(\pi,\beta,\theta) = \beta + \sum_{k=1}^{K} \theta_k \mathbb{I}_{B_k}(\pi) \qquad \mu_0(\pi,\beta,\theta) = \sum_{k=1}^{K} \theta_k \mathbb{I}_{B_k}(\pi)$$

and $\mu(\beta,\theta) = \beta$. Further, we have

$$\mu^\star(\pi,\beta,\theta) = d\mu_1(\pi,\beta,\theta) + (1-d)\mu_0(\pi,\beta,\theta) = d\beta + \sum_{k=1}^{K} \theta_k \mathbb{I}_{B_k}(\pi)$$

so that

$$\mu^\star_\beta(\pi,\beta,\theta) = d \qquad \mu^\star_{\theta_k}(\pi,\beta,\theta) = \pi \quad W(\pi) = \begin{bmatrix} d \\ \mathbb{I}_{B_1}(\pi) \\ \mathbb{I}_{B_2}(\pi) \\ \vdots \\ \mathbb{I}_{B_K}(\pi) \end{bmatrix}$$

and $\mathbb{E}[WV^{-1}W']^{-1} = \sigma^2 \mathbb{E}[WW']^{-1}$. We have

$$[WW'] = \begin{bmatrix} D^2 & D\mathbb{I}_{B_1}(\pi) & D\mathbb{I}_{B_2}(\pi) & \cdots & D\mathbb{I}_{B_K}(\pi) \\ D\mathbb{I}_{B_1}(\pi) & \mathbb{I}^2_{B_1}(\pi) & \mathbb{I}_{B_1}(\pi)\mathbb{I}_{B_2}(\pi) & \cdots & \mathbb{I}_{B_1}(\pi)\mathbb{I}_{B_K}(\pi) \\ D\mathbb{I}_{B_2}(\pi) & \mathbb{I}_{B_1}(\pi)\mathbb{I}_{B_2}(\pi) & \mathbb{I}^2_{B_2}(\pi) & \cdots & \mathbb{I}_{B_2}(\pi)\mathbb{I}_{B_K}(\pi) \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ D\mathbb{I}_{B_K}(\pi) & \mathbb{I}_{B_1}(\pi)\mathbb{I}_{B_K}(\pi) & \mathbb{I}_{B_2}(\pi)\mathbb{I}_{B_K}(\pi) & \cdots & \mathbb{I}^2_{B_K}(\pi) \end{bmatrix}$$

which can be rewritten

$$[WW'] = \begin{bmatrix} D & D\mathbb{I}_{B_1}(\pi) & D\mathbb{I}_{B_2}(\pi) & \cdots & D\mathbb{I}_{B_K}(\pi) \\ D\mathbb{I}_{B_1}(\pi) & \mathbb{I}_{B_2}(\pi) & 0 & \cdots & 0 \\ D\mathbb{I}_{B_2}(\pi) & 0 & \mathbb{I}_{B_2}(\pi) & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ D\mathbb{I}_{B_K}(\pi) & 0 & 0 & \cdots & \mathbb{I}_{B_K}(\pi) \end{bmatrix}$$

Taking expectations with respect to the joint pdf of $D$ and $\pi$, we have

$$\begin{bmatrix} E & E_1 & E_2 & \cdots & E_K \\ E_1 & P_1 & 0 & \cdots & 0 \\ E_2 & 0 & P_2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ E_K & 0 & 0 & \cdots & P_K \end{bmatrix}$$

where $E = \mathbb{E}[\pi]$, $E_k = \mathbb{E}[\pi \mathbb{I}_{B_k}(\pi)]$ and $P_k = \mathbb{P}(\pi \in B_k)$. We need to access the $(1, 1)$ element of the inverse of this matrix. Note that $E_k = \mathbb{E}[\pi \mathbb{I}_{B_k}(\pi)] = \mathbb{E}[\pi|\pi \in B_k]\mathbb{P}(\pi \in B_k) = \widetilde{E}_k P_k$ say. Using the inversion formula, we have that the required expression is

$$\left( E - \sum_{k=1}^{K} E_k^2 P_k^{-1} \right)^{-1} \tag{5.3}$$

Note that by the theorem of total probability

$$E = \mathbb{E}[\pi] = \sum_{i=1}^{K} \mathbb{E}[\pi|\pi \in B_k]\mathbb{P}(\pi \in B_k) = \sum_{i=1}^{K} \widetilde{E}_k P_k$$

and thus the variance bound for $\theta$ is from equation (5.3)

$$\frac{\sigma^2}{\sum_{k=1}^{K} P_k \widetilde{E}_k (1 - \widetilde{E}_k)}$$

Note that

$$\begin{aligned}
\mathbb{E}[\pi(1-\pi)] &= \sum_{k=1}^{K} \int_{B_k} \pi(1-\pi) f_\pi(\pi)\, d\pi \\
&= \sum_{k=1}^{K} \left\{ \frac{\displaystyle\int_{B_k} \pi(1-\pi) f_\pi(\pi)\, d\pi}{\displaystyle\int_{B_k} f_\pi(\pi)\, d\pi} \right\} \int_{B_k} f_\pi(\pi)\, d\pi \\
&= \sum_{k=1}^{K} \mathbb{E}[\pi(1-\pi) | \pi \in B_k] \mathbb{P}(\pi \in B_k) = \sum_{k=1}^{K} \mathbb{E}[\pi(1-\pi) | \pi \in B_k] P_k \\
&= \sum_{k=1}^{K} \widetilde{E}_k (1 - \widetilde{E}_k) P_k
\end{aligned}$$

so the bound for propensity score stratification is identical to that for propensity score regression.

# Appendix B

In this appendix, we prove the results stated in the text. Lemma 5.1 is an adaptation of the results given by Ibragimov & Has' Minskii (1981). It holds under the following assumptions:

- C1. The parameter space $\Xi$ is a bounded open set in $\mathcal{R}^p$.
- C2. The joint penalized density $f_p(z;\eta)$, where $z_i = (y_i, d_i, x_i)$ is a continuous function of $\eta$ on $\Xi^c$ for almost all $z \in \mathcal{Z}$, where $\mathcal{Z}$ and $\Xi^c$ represent the sample space $(y_i, d_i, \mathbf{x}_i)$ and the closure of the parameter space $\Xi$, respectively.
- C3. For all $\eta \in \Xi$ and all $\gamma > 0$,

  $$\kappa_{\eta,n}(\gamma) = \inf_{||\eta-\eta^*||>\gamma} r^2(\eta, \eta^*),$$

  where

  $$r^2(\eta, \eta^*) = \int_{\mathcal{Z}} \{f^{1/2}(z;\eta) - f^{1/2}(z;\eta^*)\}^2 d\tau$$

- C4. For $\eta \in \Xi^c$

  $$w_\eta(\delta) = \left[ \int_{\mathcal{Z}} \sup_{||h||\leq\delta} \{f^{1/2}(z;\eta) - f^{1/2}(z;\eta+h)\}^2 d\tau \right] \to 0 \text{ as } \delta \to 0.$$

- C5. The joint density $f(z;\eta)$ has a finite Fisher's information at each point $\eta \in \Xi$.

Assumption $C3$ is the identifiability condition, essentially requiring the distance between the averaged densities over the response and the covariates for two different values of the parameters $\eta$ and $\eta^*$ be positive. Assumption $C4$ is referred to as the smoothness condition; it states that the distance of the joint densities over $\eta$ and $\eta^*$ when $\eta \to \eta^*$ should approach zero as the sample size goes to infinity.

**Lemma 5.1** *Suppose assumption C1-C4 are satisfied. Then for any fixed $\eta \in \Xi$*

$$\mathbb{E}_\eta \left[ \sup_\Gamma \prod_{i=1}^n \frac{f_p^{1/2}(z_i;\eta+b)}{f_p^{1/2}(z_i;\eta)} \right] \leq \exp\left\{ -\frac{n}{2} \left[ \kappa_{\eta,n}(\frac{\gamma}{2}) - 2w_{\eta+b_0,n}(\delta) \right.\right.$$
$$\left.\left. +p_{\lambda_n}(|\eta+b^m|) - p_{\lambda_n}(|\eta|) \right] \right\}, \qquad (5.4)$$

*where $\Gamma$ is the sphere of radius $\delta$, situated in its entirely in the region $||b|| > \gamma/2$, $b_0$ is the center of $\Gamma$ and $\inf_\Gamma p_{\lambda_n}(|\eta+b|) = p_{\lambda_n}(|\eta+b^m|)$.*

**Proof** The proof follows from the proof of Theorem 4.3 in Ibragimov & Has' Minskii (1981). Let

$$R_n(b) = \prod_{i=1}^n \frac{f_p(z_i;\eta+b)}{f_p(z_i;\eta)} = \prod_{i=1}^n \frac{f(z_i;\eta+b)e^{-p_{\lambda_n}(|\eta+b|)}}{f(z_i;\eta)e^{-p_{\lambda_n}(|\eta|)}}.$$

We want to find an upper bound for the expectation $\mathbb{E}_\eta \left[ \sup_\Gamma R_n^{1/2}(b) \right]$ , where $\Gamma$ is the sphere of a radius $\delta$ situated in its entirety in the region $||b|| > \frac{1}{2}\gamma$. If $b_0$ is the centre of $\Gamma$, then

$$\sup_\Gamma R_n^{1/2}(b) = \sup_\Gamma \prod_{i=1}^n \left( \frac{f(z_i;\eta+b)e^{-p_{\lambda_n}(|\eta+b|)}}{f(z_i;\eta)e^{-p_{\lambda_n}(|\eta|)}} \right)^{1/2} \leq \prod_{i=1}^n \sup_\Gamma e^{\frac{1}{2}p_{\lambda_n}(|\eta|)-\frac{1}{2}p_{\lambda_n}(|\eta+b|)}$$
$$\prod_{i=1}^n f^{-1/2}(z_i;\eta) \left( f^{1/2}(z_i;\eta+b_0) + \sup_{h\leq\delta} |f^{1/2}(z_i;\eta+b_0+h) - f^{1/2}(z_i;\eta+b_0)| \right).$$

Thus,

$$\mathbb{E}_\beta\left[\sup_{\Gamma} R_n^{1/2}(b)\right] \le \prod_{i=1}^{n} \sup_{\Gamma} e^{\frac{1}{2}p_{\lambda_n}(|\eta|)-\frac{1}{2}p_{\lambda_n}(|\eta+b|)}\left(\int_{\mathcal{Z}} f^{1/2}(z_i;\eta)f^{1/2}(z_i;\eta+b_0)d\tau\right.$$
$$\left. + \int_{\mathcal{Z}} \sup_{|h|\le\delta} f^{1/2}(z_i;\eta)|f^{1/2}(z_i;\eta+b_0+h) - f^{1/2}(z_i;\eta+b_0)|)^n d\tau\right).$$

We further note that

$$\begin{aligned}\int_{\mathcal{Z}} f^{1/2}(z;\eta)f^{1/2}(z;\eta+b_0)d\tau =& \frac{1}{2}\left(\int_{\mathcal{Z}} f(z;\eta)d\tau + \int_{\mathcal{Z}} f(z;\eta+b_0)d\tau\right. \\ &\left. - \int_{\mathcal{Z}}[f^{1/2}(z;\eta) - f^{1/2}(z;\eta+b_0)]^2 d\tau\right) \\ &\le 1 - \frac{1}{2}r^2(\eta+b_0) \le 1 - \frac{\kappa_\eta(\frac{\gamma}{2})}{2}\end{aligned} \tag{5.5}$$

and

$$\int \sup_{|h|\le\delta} f^{1/2}(z_i;\eta)|f^{1/2}(z_i;\eta+b_0+h) - f^{1/2}(z_i;\eta+b_0)|d\tau \le w_{b_0}(\delta). \tag{5.6}$$

The last inequality follows from the Cauchy-Schwarz inequality. Finally, using the inequality $1+a \le e^a$,

$$\mathbb{E}_\beta\left[\sup_{\Gamma} R_n^{1/2}(b)\right] \le \exp\left\{-\frac{n}{2}\left[\kappa_\eta(\frac{\gamma}{2}) - 2w_{b_0}(\delta) + p_{\lambda_n}(|\eta+b^m|) - p_{\lambda_n}(|\eta|)\right]\right\}$$

where $\sup_\Gamma e^{-p_{\lambda_n}(|\eta+b|)} = e^{-p_{\lambda_n}(|\eta+b^m|)}$. ■

**Lemma 5.2** *Let $Z_1, ..., Z_n$ be independent and identically distributed with a density $f(Z,\eta)$ that satisfies the conditions of lemma 5.1. If the penalty function satisfies P3, then*

$$R_n(\eta_2) = \prod_{i=1}^{n}\left(\frac{f_p(z_i;\eta_1,\eta_2)}{f_p(z_i;\eta_1,0)}\right) < 1. \tag{5.7}$$

**Proof** $R_n(\eta_2)$ can be written as

$$\prod_{i=1}^{n}\left(\frac{f(z_i;\eta_1,\eta_2)e^{-\sum_{j=s}^{p}p_{\lambda_n}(|\eta_j|)}}{f(z_i;\eta_1,0)}\right)$$

By theorem 1.1 in Chapter II of Ibragimov & Has' Minskii (1981), it can be written as

$$R_n(\eta_2)=\exp\left\{\left[\sum_{i=1}^{n}\frac{\partial\ln f_p(z_i;\eta_1,0)}{\partial\eta_2}\right]||\eta_2||-n\sum_{j=s}^{p}p'_{\lambda_n}(|\eta_j|)-\frac{1}{2}\eta_2 I(\eta_1,0)\eta_2+\psi_n(\eta_2)\right\},$$

where $p(|\psi_n(\eta_2)|>\epsilon)\to 0$. Since $\sum_{i=1}^{n}\partial\ln f(z_i;\eta_1,0)/\partial\eta_2=O_p(\sqrt{n})$, the desired inequality holds if

$$\sqrt{n}\sum_{j=s}^{p}p'_{\lambda_n}(|\eta_j|)>||\eta_2||O_p(1),$$

which is equivalent to the condition $P_3$.

*Proof of Theorem 3.1*: For fixed $\gamma>0$, the exterior of the sphere $||b||\leq\gamma$ can be covered by $N$ spheres $\Gamma_k$, $k=1,...,N$ of radius $\delta$ with centers $b_k$. The small value $\delta$ is chosen such that all the $N$ spheres are located in the $|b|>\gamma/2$ and also

$2w_{b_k} \leq \kappa_\eta(\gamma/2)/2$. In view of the inequality (5.4), we have

$$\begin{aligned}
P(|\hat{\eta}_n - \eta_0| > \gamma) &\leq \sum_{k=1}^{N} P(|\hat{\eta}_n - \eta_0| \in \Gamma_k) \\
&\leq \sum_{k=1}^{N} P(\sup_{\Gamma} R_n(b_k) \geq R(0)) \\
&\leq \sum_{k=1}^{N} \exp\left\{-\frac{n}{2}\left[\kappa_\eta(\frac{\gamma}{2}) - 2w_{b_k}(\delta) + p_{\lambda_n}(|\eta + b^m|) - p_{\lambda_n}(|\eta|)\right]\right\} \\
\text{(Since } p_{\lambda_n}(0) = 0\text{)} \quad &\leq N\exp\left\{-\frac{n}{2}\left[\frac{1}{2}\kappa_\eta(\frac{\gamma}{2}) + \sum_{j=1}^{s} p_{\lambda_n}(|\eta_j + b_j^m|) - p_{\lambda_n}(|\eta_j|)\right]\right\},
\end{aligned}$$

where the second inequality follows from lemma 5.1 and Markov's inequality. We need to show that $\kappa_\eta(\frac{\gamma}{2})/4$ dominates $\sum_{j=s}^{p} p_{\lambda_n}(|\eta_j + b_j^m|) - p_{\lambda_n}(|\eta_j|)$. Using Taylor's expansion, we have

$$\begin{aligned}
\sum_{j=1}^{s} p_{\lambda_n}(|\eta_j + b_j^m|) - p_{\lambda_n}(|\eta_j|) &= \sum_{j=1}^{s} p'_{\lambda_n}(\eta_j)\text{sign}(\eta_j)b_j^m + \frac{1}{2}p_{\lambda_n}(\eta_j)(b_j^m)^2 \\
&\leq s\max_{\eta_j \neq 0}\{p'_{\lambda_n}(|\eta_j|)\}||b|| + \frac{s}{2}\max_{\eta_j \neq 0}\{p''_{\lambda_n}(|\eta_j|)\}||b||^2 \\
&\leq s\gamma\max_{\eta_j \neq 0}\{p'_{\lambda_n}(|\eta_j|)\} + \frac{s\gamma^2}{2}\max_{\eta_j \neq 0}\{p''_{\lambda_n}(|\eta_j|)\} \qquad (5.8)
\end{aligned}$$

By choosing $\lambda_n$ such that condition P2 holds, $\kappa_\eta(\frac{\gamma}{2})/4$ dominates the RHS of (5.8). Thus,

$$P(|\hat{\eta}_n - \eta_0| > \gamma) \leq N\exp\left\{-\frac{n}{4}\kappa_\eta(\frac{\gamma}{2})\right\},$$

and hence

$$P\left(\bigcup_{m=n}^{\infty} |\hat{\eta}_{2m}|\right) \leq \frac{N\exp\left\{-\frac{n}{4}\kappa_\eta(\frac{\gamma}{2})\right\}}{1 - \exp\left\{-\frac{1}{4}\kappa_\eta(\frac{\gamma}{2})\right\}} \to 0 \text{ as } n \to \infty.$$

This completes our proof of strong consistency of the penalized maximum likelihood estimator. $\blacksquare$

*Proof of Theorem 3.2*: Consider $\eta_t = (\eta_{t1}, 0)$ and partition $\eta = (\eta_1, \eta_2)$. We need to show that in the neighbourhood $||\eta - \eta_t|| < O(h_n)$ where $h_n \to 0$ as $n \to \infty$,

$$\prod_{i=1}^{n} \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \hat{\eta}_1, 0)} < 1.$$

It can be written as

$$\prod_{i=1}^{n} \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \hat{\eta}_1, 0)} = \prod_{i=1}^{n} \left( \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} \right) \left( \frac{f_p(z_i; \eta_1, 0)}{f_p(z_i; \hat{\eta}_1, 0)} \right) < \prod_{i=1}^{n} \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} < 1.$$

By the result of Lemma 5.2, the last inequality holds with probability one as $n \to \infty$.

$\blacksquare$

# Appendix C

In this section, we present the assumptions, proofs of the main and other auxiliary results. The regularity conditions required in this paper are as follows:

C.1 $\mu_j(.)$ for $j = 0, 1$ is a twice continuously differentiable function.

C.2 $\pi(.)$ is not on the boundaries ($\gamma < \pi(.) < 1 - \gamma$ where $\gamma > 0$).

C.3 $\sup[t : p(R > t) > 0] \geq \sup[t : p(C > t) > 0]$ and $p(\delta = 1) > 0$.

C.4 $\int_0^s [(\int_t^s S_C(v)dv)^2/(S_C^2(t)S_V(t))]dS_C(t) < \infty$.

C.5 $\int_0^s \kappa_j^2(t)/(S_C^2(t)S_R(t))dS_C(v) < \infty$ and $\int_0^s \kappa^2(t)/(S_C^2(t)S_R(t))dS_C(v) < \infty$ where

$$\kappa_j(t) = \mathbb{E}\left[\frac{I(D = j)\delta I(Y > t)[\log(Y) - \mu_j(\mathbf{X}, \theta)]\int_t^Y S_C(v)dv}{\pi(\mathbf{X})w^2(Y)}\right],$$

$$\kappa(t) = \mathbb{E}\left[\frac{\delta I(Y > t)[\mu_1(\mathbf{X}, \theta) - \mu_0(\mathbf{X}, \theta) - \beta]\int_t^Y S_C(v)dv}{\pi(\mathbf{X})w^2(Y)}\right].$$

Condition C.1 is a smoothness assumption of the mean function. C.1-C.2 are termed *positivity* assumptions, meaning that there is a positive chance that a subject falls in either the treatment or the control groups and being not censored, respectively. C.3 is an identifiability condition (Wang (1991)) and C.4-C.5 are required to obtain an estimator with a finite variance.

*Proof of Theorem 4.1.* The class of efficient influence functions is

$$\mathcal{G}_2^{(\mathrm{AFT})} = \left\{\frac{\delta}{\widehat{w}(Y)}\left[\frac{D[\log(Y) - \mu_1(\mathbf{X}, \theta)]}{\widehat{\pi}(\mathbf{X})} - \frac{(1 - D)[\log(Y) - \mu_0(\mathbf{X}, \theta)]}{(1 - \widehat{\pi}(\mathbf{X}))}\right.\right.$$
$$\left.\left. + \mu_1(\mathbf{X}, \theta) - \mu_0(\mathbf{X}, \theta) - \beta\right]\right\}.$$

Let $\mathcal{G}_2^{(\mathrm{AFT})} = \{\widehat{V}_0(\theta) - \widehat{V}_1(\theta) + \widehat{V}_2(\theta)\}$ where

$$\begin{aligned}
\widehat{V}_0(\theta) &= \frac{D\delta[\log(Y) - \mu_1(\mathbf{X},\theta)]}{\widehat{\pi}(\mathbf{X})\widehat{w}(Y)} \\
\widehat{V}_1(\theta) &= \frac{(1-D)\delta[\log(Y) - \mu_0(\mathbf{X},\theta)]}{(1-\widehat{\pi}(\mathbf{X}))\widehat{w}(Y)} \\
\widehat{V}_2(\theta) &= \frac{\delta}{\widehat{w}(Y)}[\mu_1(\mathbf{X},\theta) - \mu_0(\mathbf{X},\theta) - \beta].
\end{aligned}$$

In order to show that $\mathcal{G}_2^{(\mathrm{AFT})}$ results in an unbiased estimator, we need to show that $E[\widehat{V}_0(\theta)] = \mathbb{E}[\widehat{V}_1(\theta)] = \mathbb{E}[\widehat{V}_2(\theta)] = 0$. For the first expectation, we have

$$\begin{aligned}
\mathbb{E}\left[V_0(\theta)|\mathbf{X}\right] &\propto \mathbb{E}\Bigg[\int_0^\infty \int_0^\infty f(Y=y, A=a, \delta=1, D=1|\mathbf{X}) \\
&\qquad \times \frac{D[\log(Y) - \mu_1(\mathbf{X},\theta)]}{\widehat{\pi}(\mathbf{X})w(Y)}\, da\, dy\Bigg] \\
&\propto \mathbb{E}\left[\frac{D}{\widehat{\pi}}|\mathbf{X}\right] \mathbb{E}\Bigg[\int_0^\infty \int_0^y f_U(Y=y|D=1,\mathbf{X})S_c(y-a) \\
&\qquad \times \frac{[\log(Y) - \mu_1(\mathbf{X},\theta)]}{w(Y)}\, da\, dy\Bigg] \\
&\propto \mathbb{E}\left[\frac{1}{\mu(1,\mathbf{X})}\int_0^\infty f_U(y|\mathbf{X}, D=1)[\log(Y) - \mu_1(\mathbf{X},\theta)]dy\right] \\
&= 0,
\end{aligned}$$

where $\mu(1,\mathbf{X}) = \int y f_U(y|D=1,\mathbf{X})\, dy$. If we replace the $S_c(.)$ by its estimate, we can show that $E[\widehat{V}_0(\theta)] = 0$. Similarly, we can show that $E[\widehat{V}_1(\theta)] = E[\widehat{V}_2(\theta)] = 0$. It can be shown that $\widehat{V}_0(\theta)$ and $\widehat{V}_1(\theta)$ are uncorrelated. Then using the strong consistency of $\widehat{w}(y)$ to $w(y)$ (Pepe & Fleming (1991)) and following the martingale integral representation $\sqrt{n}(\widehat{w}(y) - w(y))$ introduced by Shen et al. (2009), we can write the asymptotic variance of the corresponding estimator as

$$\begin{aligned}\mathbb{E}\left[\widehat{V}_0^2(\theta)\right] &= \mathbb{E}\left[\left\{\frac{D\delta[\log(Y)-\mu_1(\mathbf{X},\theta)]}{\pi(\mathbf{X})w(Y)}\right\}^2\left\{1+\frac{w(Y)-\widehat{w}(Y)}{w(Y)}\right\}^2\right]\\ &= \mathbb{E}\left[\left\{\frac{D\delta[\log(Y)-\mu_1(\mathbf{X},\theta)]}{\pi(\mathbf{X})w(Y)}+\int_0^Y\frac{\kappa_1(t)dM_i(t)}{S_C(t)S_R(t)}\right\}^2\right],\end{aligned}$$

and similarly the second and third parts are

$$\begin{aligned}\mathbb{E}\left[\widehat{V}_1^2(\theta)\right] &= \mathbb{E}\left[\left\{\frac{(1-D)\delta[\log(Y)-\mu_0(\mathbf{X},\theta)]}{(1-\pi(\mathbf{X}))w(Y)}+\int_0^Y\frac{\kappa_0(t)dM_i(t)}{S_C(t)S_R(t)}\right\}^2\right],\\ \mathbb{E}\left[\widehat{V}_2^2(\theta)\right] &= \mathbb{E}\left[\left\{\frac{\delta}{w(Y)}[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta]+\int_0^Y\frac{\kappa(t)dM_i(t)}{S_C(t)S_R(t)}\right\}^2\right],\end{aligned}$$

where

$$\begin{aligned}\kappa_j(t) &= \mathbb{E}\left[\frac{I(D=j)\delta I(Y>t)[\log(Y)-\mu_j(\mathbf{X},\theta)]\int_t^Y S_C(v)dv}{\pi(\mathbf{X})w^2(Y)}\right]\\ \kappa(t) &= \mathbb{E}\left[\frac{\delta I(Y>t)[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta]\int_t^Y S_C(v)dv}{w^2(Y)}\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\widehat{V}_1(\theta)\widehat{V}_2(\theta)\right] &= \mathbb{E}\left[\left\{\frac{(1-D)\delta[\log(Y)-\mu_0(\mathbf{X},\theta)]}{(1-\pi(\mathbf{X}))w(Y)}+\int_0^Y\frac{\kappa_0(t)dM_i(t)}{S_C(t)S_R(t)}\right\}\right.\\ &\qquad\left.\times\left\{\frac{\delta}{w(Y)}[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta]+\int_0^Y\frac{\kappa(t)dM_i(t)}{S_C(t)S_R(t)}\right\}\right],\\ \mathbb{E}\left[\widehat{V}_0(\theta)\widehat{V}_2(\theta)\right] &= \mathbb{E}\left[\left\{\frac{\delta}{w(Y)}[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta]+\int_0^Y\frac{\kappa(t)dM_i(t)}{S_C(t)S_R(t)}\right\}\right.\\ &\qquad\left.\times\left\{\frac{\delta}{w(Y)}[\mu_1(\mathbf{X},\theta)-\mu_0(\mathbf{X},\theta)-\beta]+\int_0^Y\frac{\kappa(t)dM_i(t)}{S_C(t)S_R(t)}\right\}\right].\end{aligned}$$

# References

Addona, V. & Wolfson, D. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Analysis* **12**, 267–284.

Alexander, C. & Markowitz, R. (1986). Maternal employment and use of pediatric clinic services. *Medical Care* **24**, 134–147.

Antoniadis, A. (1997). Wavelets in statistics: a review. *Statistical Methods and Applications* **6**, 97–130.

Asgharian, M., M'Lan, C. & Wolfson, D. (2002). Length-biased sampling with right censoring. *Journal of the American Statistical Association* **97**, 201–209.

Asgharian, M. & Wolfson, D. (2005). Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics* **33**, 2109–2131.

Asgharian, M., Wolfson, D. & Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in medicine* **25**, 1751–1767.

Bang, H. & Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–972.

BEGUN, J., HALL, W., HUANG, W. & WELLNER, J. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics* **11**, 432–452.

BERGERON, P., ASGHARIAN, M. & WOLFSON, D. (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association* **103**, 737–742.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. & WELLNER, J. A. (1993a). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Springer.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. & WELLNER, J. A. (1993b). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Johns Hopkins University Press.

BROOKHART, M. & VAN DER LAAN, M. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational statistics & data analysis* **50**, 475–498.

BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. & STURMER, T. (2006a). Variable selection for propensity score models. *American journal of epidemiology* **163**, 1149–1156.

CHAMBERLAIN, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305–334.

CHAMBERLAIN, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society* **60**, 567–596.

Cheng, S., Wei, L. & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

Cox, D. (1969). *Some sampling problems in technology.* in New Developments in Survey Sampling: New York: Wiley Interscience, pp. 506–527.

Cox, D. & Lewis, P. (1966). *The statistical analysis of series of events.* John Wiley and Sons.

Cox, D. & Oakes, D. (1984). *Analysis of survival data.* Chapman & Hall/CRC.

Crainiceanu, C., Dominici, F. & Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* **95**, 635–651.

Davidian, M., Tsiatis, A. & Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science: a review journal of the Institute of Mathematical Statistics* **20**, 261–301.

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

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

Ertefaie, A. & Stephens, D. (2010). Comparing approaches to causal inference for longitudinal data: inverse probability weighting versus propensity scores. *The International Journal of Biostatistics* **6**, 14.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fisher, R. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Human Genetics* **6**, 13–25.

Gail, M. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. *Modern statistical methods in chronic disease epidemiology* , 3–18.

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

Gilbert, P., Lele, S. & Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. *Biometrika* **86**, 27–43.

Gill, R., Vardi, Y. & Wellner, J. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* , 1069–1112.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology* **167**, 523–529.

Greenland, S., Robins, J. & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46.

Hafeman, D. & VanderWeele, T. (2010). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* **21**.

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

Hajek, J. (1970). A characterization of limiting distributions of regular estimates. *robability Theory and Related Fields* **14**, 323–330.

Hardie, W. & Linton, O. (1994). Applied nonparametric methods. *Handbook of econometrics* **4**, 2295–2339.

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

Heckman, J., Ichimura, H. & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic studies* **65**, 261–294.

Hernan, M., Cole, S., Margolick, J., Cohen, M. & Robins, J. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and drug safety* **14**, 477–491.

Hernán, M., Hernández-Díaz, S. & Robins, J. (2004). A structural approach to selection bias. *Epidemiology* **15**, 615–625.

Hirano, K. & Imbens, G. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives* , 73–84.

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

Huber, P., Ronchetti, E. & Corporation, E. (1981). *Robust statistics*, vol. 1. Wiley Online Library.

Ibragimov, I. & Has' Minskii, R. (1981). *Statistical estimation–asymptotic theory*. Springer.

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

Joffe, M. & Colditz, G. (1998). Restriction as a method for reducing bias in the estimation of direct effects. *Statistics in Medicine* **17**, 2233–2249.

Kalbfleisch, J. & Prentice, R. (1980). *The statistical analysis of failure time data*, vol. 5. Wiley New York.

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

Kaufman, J., MacLehose, R. & Kaufman, S. (2004). A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives & Innovations* **1**, 4.

Kaufman, S., Kaufman, J., MacLehose, R., Greenland, S. & Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine* **24**, 1683–1702.

Khalili, A. & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**, 1025–1038.

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

Little, R. J. A. & An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica* **14**, 949—968.

Moodie, E. & Stephens, D. (2010). Estimation of dose-response functions for longitudinal data. *Statistical Methods in Medical Research* .

Neugebauer, R. & van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* **129**, 405–426.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

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

Neyman, J. (1955). Statistics–servant of all science. *Science* **122**, 401–406.

Patil, G. & Rao, C. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34**, 179–189.

Pepe, M. & Fleming, T. (1991). Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 341–352.

Petersen, M., Sinisi, S. & van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology-Baltimore* **17**, 276–284.

Qin, J. & Shen, Y. (2010). Statistical methods for analyzing right-censored length-biased data under Cox model. *Biometrics* **66**, 382–392.

Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79**, 321–334.

Robins, J. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, M. Berkane, ed., Lecture Notes in Statistics (120). Springer, Berlin, pp. 69–117.

Robins, J. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, vol. 6.

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

Robins, J. & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155.

Robins, J. & Richardson, T. (2010). Alternative graphical causal models and the identification of direct effects.

Robins, J., Richardson, T. & Spirtes, P. (2010). On identification and inference for direct effects .

Robins, J. M., Mark, S. D. & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–495.

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

Rosenbaum, P. (2010). Causal inference in randomized experiments. *Design of Observational Studies* , 21–63.

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

Rotnitzky, A., Li, L. & Li, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika* , 997–1001.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127**, 757–763.

Rubin, D. (2004a). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* , 161–170.

Rubin, D. (2004b). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety* **13**, 855–857.

Rubin, D. (2005). Bayesian inference for causal effects. In *Bayesian thinking: modeling and computation*, D. Dey & C. Rao, eds. North-Holland, pp. 1–14.

Rubin, D. (2008). For objective causal inference, design trumps analysis. *Annals* **2**, 808–840.

Schafer, J. L. & Kang, J. D. Y. (2005). Discussion of "semiparametric estimation of treatment effect in a pretest–postest study with missing data" by m. davidian et al. *Statist. Sci.* **20**, 292—295.

Schisterman, E., Cole, S. & Platt, R. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* **20**, 488–495.

Shen, Y., Ning, J. & Qin, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association* **104**, 1192–1202.

Stewart, C., Fielder, A., Stephens, D. & Moseley, M. (2002). Design of the monitored occlusion treatment of amblyopia study (MOTAS). *British Journal of Ophthalmology* **86**, 915–919.

Stewart, C., Moseley, M., Stephens, D. & Fielder, A. (2004). Treatment dose-response in amblyopia therapy: the monitored occlusion treatment of amblyopia study (MOTAS). *Investigative Ophthalmology & Visual Science* **45**, 3048–3054.

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

Tsiatis, A. (2006). *Semiparametric theory and missing data.* Springer Verlag.

Van Der Laan, M. & Petersen, M. (2004). Estimation of direct and indirect causal effects in longitudinal studies. *UC Berkeley Division of Biostatistics Working Paper Series* , 155.

van der Laan, M. & Robins, J. (2003). *Unified methods for censored longitudinal data and causality.* Springer Verlag.

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20**, 18–26.

VanderWeele, T. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21**, 540–551.

Vansteelandt, S., Bekaert, M. & Claeskens, G. (2010). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* , 1477–0334.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* **10**, 616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13**, 178–203.

VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76**, 751–761.

VARDI, Y. & ZHANG, C. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *The Annals of Statistics* **20**, 1022–1039.

WANG, M. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86**, 130–143.

WICKSELL, S. (1925). The corpuscle problem: a mathematical study of a biometric problem. *Biometrika* **17**, 84–99.

WOLFSON, C., WOLFSON, D., ASGHARIAN, M., M'LAN, C., ØSTBYE, T., ROCKWOOD, K. & HOGAN, D. (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine* **344**, 1111–1116.

ZELEN, M. & FEINLEIN, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56**, 601–614.

ZOU, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **67**, 301–320.