1-1-2016

Doubly Robust Causal Inference With Complex Parameters

Edward Henry Kennedy
University of Pennsylvania, edwardh.kennedy@gmail.com

Doubly Robust Causal Inference With Complex Parameters

Abstract

Semiparametric doubly robust methods for causal inference help protect against bias due to model misspecification, while also reducing sensitivity to the curse of dimensionality (e.g., when high-dimensional covariate adjustment is necessary). However, doubly robust methods have not yet been developed in numerous important settings. In particular, standard semiparametric theory mostly only considers independent and identically distributed samples and smooth parameters that can be estimated at classical root-n rates. In this dissertation we extend this theory and develop novel methodology for three settings outside these bounds: (1) matched cohort studies, (2) nonparametric dose-response estimation, and (3) complex high-dimensional effects with continuous instrumental variables. After giving an introduction in Chapter 1, we show in Chapter 2 that, for matched cohort studies, efficient and doubly robust estimators of effects on the treated are computationally equivalent to standard estimators that ignore the non-standard sampling. We also show that matched cohort studies are often more efficient than random sampling for estimating effects on the treated, and derive the optimal number of matches for given matching variables. We apply our methods in a study of the effect of hysterectomy on the risk of cardiovascular disease. In Chapter 3 we develop a novel approach for causal dose-response curve estimation that is doubly robust without requiring any parametric assumptions, and which naturally incorporates general off-the-shelf machine learning. We derive asymptotic properties for a kernel-based version of our approach and propose a data-driven method for bandwidth selection. The methods are used to study the effect of hospital nurse staffing on excess readmissions penalties. In Chapter 4 we develop novel estimators of the local instrumental variable curve, which represents the treatment effect among compliers who would take treatment when the instrument passes some threshold. Our methods do not require parametric assumptions, allow for flexible data-adaptive estimation of effect modification, and are doubly robust. We derive asymptotic properties under weak conditions, and use the methods to study infant mortality effects of neonatal intensive care units with high versus low technical capacity, using travel time as an instrument.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Dylan S. Small

Keywords

causal inference, health policy, machine learning, nonparametric methods, semiparametric theory

DOUBLY ROBUST CAUSAL INFERENCE WITH COMPLEX PARAMETERS

Edward H. Kennedy

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Dylan S. Small, Professor of Statistics

Graduate Group Chairperson

John H. Holmes, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Harold I. Feldman, Professor of Epidemiology & Medicine

Zongming Ma, Assistant Professor of Statistics

R. Taki Shinohara, Assistant Professor of Biostatistics

DOUBLY ROBUST CAUSAL INFERENCE WITH COMPLEX PARAMETERS

Dedicated to my family.

ACKNOWLEDGEMENT

To my advisors, Dylan and Marshall: I'm so grateful to have had the opportunity to work with you over the last four years. You are two of the most intelligent, generous, and inspiring people I've ever met, and I will always look up to you both.

Unfortunately I do not have enough space to appropriately thank everyone who contributed to this dissertation work in some way. The long list includes:

- my dissertation committee (Eric, Harv, Taki, Zongming), who helped with everything from moral support to administrative tasks to discussing proofs;
- the causal group (Alisa, Ashkan, Dylan, Jason, Jesse, Marshall, Peter, etc.), who taught me a lot, generously listened to my ideas, and provided great discussions;
- many other faculty (Dick, Jason, Mary, Nandita, Peter, Taki, Zongming, etc.) who selflessly shared their time and were excellent mentors and teachers;
- kind and supportive administrative staff (Cathy, Marissa, Noelle);
- fellow students at Penn and elsewhere (Bret, Emin, Jared, Julie, Jun, Valerio, etc.), who kept my life balanced with good company, laughter, music, beer, and more.

I hope others who I have not listed by name still know how much I appreciate their help.

Lastly, I want to especially thank my family:

- To my brothers and sisters: You are all so important to me, and I would be lost without you guys. Thank you for everything.
- To my parents: Your unwavering guidance and support has meant so much to me over the years, and really made this work possible. I can't say enough how much you have inspired me, and how thankful I am for your faith in me.
- To Traci: Thank you for encouraging me, for being patient when I am in math mode, and for being my partner in everything. I am excited for our next steps together!

ABSTRACT

DOUBLY ROBUST CAUSAL INFERENCE WITH COMPLEX PARAMETERS

Edward H. Kennedy

Dylan S. Small

Semiparametric doubly robust methods for causal inference help protect against bias due to model misspecification, while also reducing sensitivity to the curse of dimensionality (e.g., when high-dimensional covariate adjustment is necessary). However, doubly robust methods have not yet been developed in numerous important settings. In particular, standard semiparametric theory mostly only considers independent and identically distributed samples and smooth parameters that can be estimated at classical root-n rates. In this dissertation we extend this theory and develop novel methodology for three settings outside these bounds: (1) matched cohort studies, (2) nonparametric dose-response estimation, and (3) complex high-dimensional effects with continuous instrumental variables. After giving an introduction in Chapter 1, we show in Chapter 2 that, for matched cohort studies, efficient and doubly robust estimators of effects on the treated are computationally equivalent to standard estimators that ignore the non-standard sampling. We also show that matched cohort studies are often more efficient than random sampling for estimating effects on the treated, and derive the optimal number of matches for given matching variables. We apply our methods in a study of the effect of hysterectomy on the risk of cardiovascular disease. In Chapter 3 we develop a novel approach for causal dose-response curve estimation that is doubly robust without requiring any parametric assumptions, and which naturally incorporates general off-the-shelf machine learning. We derive asymptotic properties for a kernel-based version of our approach and propose a data-driven method for bandwidth selection. The methods are used to study the effect of hospital nurse staffing on excess readmissions penalties. In Chapter 4 we develop novel estimators of the local instrumental variable curve, which represents the treatment effect among compliers who would take treatment when the instrument passes some threshold. Our methods do not require parametric assumptions, allow for flexible data-adaptive estimation of effect modification, and are doubly robust. We derive asymptotic properties under weak conditions, and use the methods to study infant mortality effects of neonatal intensive care units with high versus low technical capacity, using travel time as an instrument.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : INTRODUCTION

Many important problems in causal inference, missing data, and other settings lead to parameters that can be estimated doubly robustly. A full characterization of what it means to be doubly robust and when exactly double robustness is possible is an open problem. However, the following illustration covers many useful examples. Consider a target parameter ψ (e.g., an average treatment effect) and a corresponding estimator $\hat{\psi}$, which is constructed based on a sample of observed data $(Z_1, ..., Z_n)$. Suppose further that $\hat{\psi}$ is a regular asymptotically linear estimator with influence function φ, so that it has the representation $\hat{\psi} = \psi_0 + \mathbb{P}_n\{\varphi(Z)\} + o_p(1/\sqrt{n})$, where $\mathbb{P}_n$ denotes the empirical measure with $\mathbb{P}_n(f) = n^{-1}\sum_i f(Z_i)$, and $X_n = o_p(r_n)$ means X_n/r_n converges in probability to zero. Then the estimator $\hat{\psi}$ is doubly robust if the influence function $\varphi(\cdot) = \varphi(\cdot;\eta) = \varphi(\cdot;\pi,\mu)$ depends on two nuisance functions $\eta = (\pi, \mu)$ (e.g., a propensity score and outcome regression function) and satisfies $\mathbb{E}\{\varphi(Z;\pi_0,\overline{\mu})\} = \mathbb{E}\{\varphi(Z;\overline{\pi},\mu_0)\} = \mathbb{E}\{\varphi(Z;\pi_0,\mu_0)\} = 0$ for arbitrary $\overline{\eta} = (\overline{\pi}, \overline{\mu})$. Thus the influence function has mean zero (and thus is an unbiased estimating function) as long as one of the two nuisance functions is evaluated at the truth.

Doubly robust estimators have several crucial advantages. First, they give analysts two independent chances at arriving at the truth in large samples, since they are consistent as long as only one of two nuisance functions is consistently estimated (i.e., even if one is misspecified). This helps protect against bias from model misspecification, which is particularly important in complex high-dimensional data settings where simple parametric model assumptions are unrealistic. Second, doubly robust estimators are also less sensitive to the curse of dimensionality than more standard estimators. This follows from the fact that they can attain faster rates of convergence than the nuisance estimators they depend on (when both are consistently estimated); this is not the case for standard plug-in estimators, which rely on a single nuisance estimator. Thus, even after model selection and machine learning-based covariate adjustment, doubly robust estimators can yield fast rates of convergence and uniformly valid inference (e.g., confidence intervals).

However, despite their many advantages and increasing popularity, doubly robust methods have not yet been developed in numerous important settings. In particular, they have mostly only been established for independent and identically distributed samples, and for relatively straightforward target parameters that can be estimated at classical root-n rates of convergence. In this dissertation we extend double robustness theory and develop novel methodology for settings outside these bounds.

In Chapter 2 we consider semiparametric doubly robust estimation and inference in matched cohort studies, which are a popular but non-standard sampling design. We show that efficient doubly robust estimators of effects on the treated in such designs are computationally equivalent to standard estimators that ignore the sampling, and explore various issues related to efficiency and study design. We apply our methods in a matched cohort study of the effect of hysterectomy on the risk of cardiovascular disease. In Chapter 3 we develop a novel nonparametric doubly robust approach for causal dose-response curve estimation, which is an interesting but common example where double robustness is possible even though standard root-n rates are not achievable. Our approach naturally incorporates general off-the-shelf machine learning tools, and we explore its asymptotic properties under weak conditions. We use our estimator to study the effect of hospital nurse staffing on excess readmissions penalties. Finally in Chapter 4 we develop novel semiparametric doubly robust estimators of the local instrumental variable curve, which is a complex parameter representing the treatment effect among compliers who would take treatment when the instrument passes some threshold. We also develop an approach for doubly robust model selection, and apply our methods to study the effects on infant mortality of delivery at high-versus low-level neonatal intensive care units (using travel time as an instrument).

CHAPTER 2 : SEMIPARAMETRIC CAUSAL INFERENCE IN MATCHED COHORT STUDIES

2.1. Abstract

Odds ratios can be estimated in case-control studies using standard logistic regression, ignoring the outcome-dependent sampling. In this paper we discuss an analogous result for treatment effects on the treated in matched cohort studies. Specifically, in studies where a sample of treated subjects is observed along with a separate sample of possibly matched controls, we show that efficient and doubly robust estimators of effects on the treated are computationally equivalent to standard estimators, which ignore the matching and exposure-based sampling. This is not the case for general average effects. We also show that matched cohort studies are often more efficient than random sampling for estimating effects on the treated, and derive the optimal number of matches for a given set of matching variables. We illustrate our results via simulation and in a matched cohort study of the effect of hysterectomy on the risk of cardiovascular disease.

2.2. Introduction

In this paper we consider matched cohort studies in which a sample of treated subjects is observed along with a separate sample of possibly matched controls. Such studies are particularly useful in settings where the treatment is relatively uncommon and it is expensive to collect either the outcome data or the full set of covariates. These designs are also widely used; according to PubMed the number of articles including both terms "matched" and "cohort" has increased every year since 2000, and totals 19,581 as of 8 January 2015. For example, Ingelsson et al. (2011) used a matched cohort design to estimate the effect of hysterectomy on the risk of cardiovascular disease. They first identified all Swedish women who underwent hysterectomies between 1973 and 2003 using the Swedish Inpatient Register, and then for each of these women matched three additional women who did not have a hysterectomy but who were the same age and lived in the same county. In this study it was difficult to collect outcome data about cardiovascular events, as well as additional covariate information such as socioeconomic status, because linkage to numerous additional national health registers was required. More examples and general discussion of matched cohort studies can be found in Jewell (2003) and Rothman et al. (2008).

Matched cohort studies are most often used for estimating treatment effects on the treated. These effects can be of more interest than average effects, especially when treatment is relatively rare and some subjects are very unlikely to receive it. A primary contribution of this paper is to show that effects on the treated can be estimated in matched cohort studies using standard methods, ignoring the study design; this is a cohort study analog of the famous odds ratio result for case-control studies (Anderson, 1972; Prentice and Pyke, 1979). To the best of our knowledge, this fact has never before been mentioned in the literature. It means that, for example, even though propensity scores are not identified in matched cohort designs, usual semiparametric, e.g., propensity score-based, doubly robust, estimators of the effect on the treated can be applied without modification, and without requiring external information about treatment prevalence or matched covariate distributions. Thus much of the important literature on semiparametric estimation of effects on the treated (Heckman et al., 1997; Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003; Imbens, 2004; Abadie and Imbens, 2006; Kline, 2011) is also relevant for matched cohort studies, even though this work has mostly focused on simple random sampling.

A number of authors have considered causal inference in matched cohort study settings, but none seem to have mentioned the above result. Heckman and Todd (2009) gave some justification for using the propensity score in exposure-stratified studies without matching, but did not discuss semiparametric theory or double robustness. Tchetgen Tchetgen and Rotnitzky (2011) developed semiparametric theory and doubly robust estimators for the conditional odds ratio but did not consider general marginal effects or efficiency across study designs. Sjölander et al. (2012) and Sjölander and Greenland (2013) discussed using likelihood-based regression methods, but did not consider using propensity scores. van der

Laan et al. (2013) examined cohort studies for community-based interventions, but required external information beyond the sample.

2.3. Setup

We consider the following study setup. Covariates L and outcome Y are observed for n_1 treated subjects along with n_0 controls, where $n_0 = kn_1$ is fixed so that k controls are selected for each treated subject. In addition the controls can be matched to the treated on a subset of discrete covariates $W \subseteq L$. We use $\overline{W} = L \setminus W$ to denote the set of covariates not used in matching, so that $L = (W, \overline{W})$. The observed data are $(Z_1, ..., Z_n)$ with $Z = (L, A, Y)$ and A an indicator of treatment, where by design we have that $\sum_i A_i = n_1$, $\sum_i (1 - A_i) = n_0 = kn_1$, and $W_i = W_j$ if subjects i and j are matched. If there is no matching so that $W = \emptyset$ and $\overline{W} = L$ then we simply observe two separate random samples of treated and control subjects.

The main statistical issue in a matched cohort study is the fact that the observations are not independent and identically distributed from the population of interest. Specifically, the proportion treated in the sample is fixed due to the exposure-stratified sampling, and the distribution of the matched covariates is forced to be the same for the treated and control subjects due to the matching. Although the implications for causal inference are different, this setup is conceptually similiar to that of a case-control study, where sampling is stratified by outcome (Breslow et al., 2000). As in case-control studies, although the observations in a matched cohort study are not an independent and identically distributed sample from the population distribution of interest, they can be viewed as an independent and identically distributed sample from a particular modified distribution. This is called the biased sampling model framework (Jewell, 1985; Bickel et al., 1993). In a matched cohort study the observations $(Z_1, ..., Z_n)$ arise from a biased distribution Q with density

$$q(z) = p(y \mid l, a) p(\overline{w} \mid w, a) p(w \mid a = 1) q(a), \tag{2.1}$$

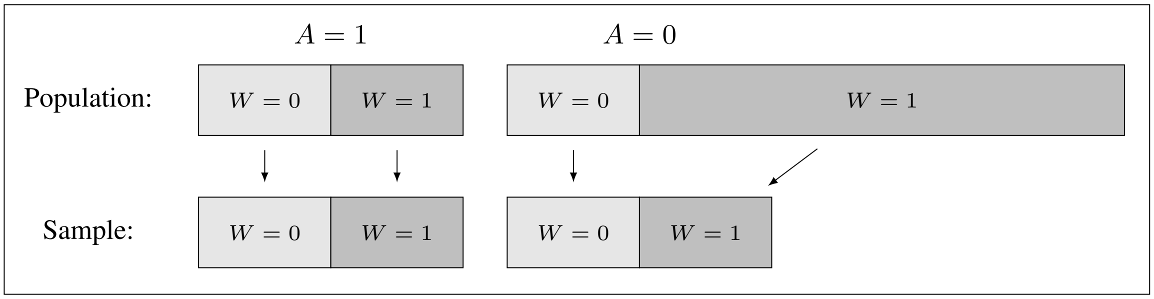

Figure 1: Schematic of matched cohort study design for 1:1 matching on a binary variable W. Arrows denote random samples of size $n_1/2$.

where P denotes the distribution of Z in a larger population of interest, with density given by $p(z) = p(y \mid l, a)p(l \mid a)p(a)$ with respect to some dominating measure, and $q(a)$ is the proportion of subjects in the sample receiving treatment level a. In general we write the density under distribution F of variable X evaluated at value c as $f(x = c)$, except when the density we are referring to is unambiguous, e.g., $f(x)$ denotes the density of X under F. The likelihood can be written as $\prod_i p(y_i \mid l_i, a_i)p(\overline{w}_i \mid w_i, a_i)q(a_i) \prod_j p(w_j \mid a = 1)$, where i references units and j references matched strata. For illustration Figure 1 gives a schematic of a matched cohort study in the simple case of 1:1 matching on a binary variable.

In subsequent sections we characterize causal treatment effects using potential outcome notation (Rubin, 1974), and so let Y^a denote the potential outcome that would have been observed had treatment level a been applied. We further make use of some simplifying notation. Specifically we use $\pi(l)$ to denote the propensity score under P given by $p(a = 1 \mid l)$, and we use $\xi(l)$ to denote the analog of the propensity score in the biased distribution Q given by $q(a = 1 \mid l)$. We also use $\mu(l, a)$ to denote the conditional mean of the outcome given covariates and treatment $\mathbb{E}(Y \mid L = l, A = a)$, which is the same under both P and Q whenever it exists. All expectations are taken under the distribution P of interest, unless otherwise noted with a subscript, as in $\mathbb{E}_Q$.

2.4. Identification and Estimation

Throughout we consider the following identifying assumptions, the third of which is commonly called no unmeasured confounding.

Assumption 2.1 (Consistency) *If $A = a$ then $Y = Y^a$ with probability one.*

Assumption 2.2 (Positivity) *For all l such that $p(l) > 0$, we have $0 < \pi(l) < 1$.*

Assumption 2.3 (Ignorability) *For $a \in \{0, 1\}$, $\mathbb{E}(Y^a \mid L, A = 1) = \mathbb{E}(Y^a \mid L, A = 0)$.*

These assumptions are all typically satisfied by design in randomized trials, but in observational studies they may be violated and are generally untestable. Consistency ensures that one potential outcome is observed for every subject, namely that potential outcome under the treatment that was actually received; it can fail to hold if different versions of treatment have different effects, or if there is interference, for example. Positivity says that treatment is not assigned deterministically, in the sense that every subject has some positive probability of receiving both treatment and control, regardless of covariates. Ignorability says that the mean potential outcomes are the same for both treatment groups once we condition on the covariates, and requires sufficiently many relevant covariates to be collected.

It is well-known and straightforward to show that $\mathbb{E}(Y^a) = \int \mu(l, a)p(l)d\nu(l)$ under Assumptions 2.1–2.3, where ν is a dominating measure for the distribution of L. Importantly, his expression is identified under P, but not under Q since we observe $q(l) \neq p(l)$ under Q. Note that

$$p(l) = q(\overline{w} \mid w, a = 0)p(w \mid a = 0)p(a = 0) + q(\overline{w} \mid w, a = 1)q(w \mid a = 1)p(a = 1)$$

since $q(\overline{w} \mid w, a) = p(\overline{w} \mid w, a)$ and $q(w \mid a = 1) = p(w \mid a = 1)$, but at least $p(a)$ is not identified under Q. Without matching, the covariate distributions given treatment $p(l \mid a)$ would be identified, but matching further removes identification of the covariate distribution among the controls since it forces $q(w \mid a = 0) = p(w \mid a = 1)$. Thus, identification of average effects $\mathbb{E}(Y^a)$ cannot be achieved under matched cohort sampling without external knowledge of the treatment proportions $p(a)$ and the matched covariate density $p(w \mid a = 0)$.

If $p(a)$ and $p(w \mid a = 0)$ are known from external data, however, one can construct estimators of $\mathbb{E}(Y^a)$, or any other parameter defined on P, based on appropriately weighted estimating functions, as in van der Laan et al. (2013). Weighting is necessary since estimating functions based on P will in general be biased, e.g., not have mean zero, under Q. For use in matched cohort studies, estimating functions under P should be weighted by $b(W, A) = \{p(A)/q(A)\}\{p(W \mid A)/p(W \mid a = 1)\}$ since $p(z) = q(z)b(w, a)$.

In many cases such external information is not available, especially when W is high-dimensional. But this is not problematic for estimation of the effect on the treated, which is given by $\psi = \mathbb{E}(Y^1 - Y^0 \mid A = 1)$. Under Assumptions 2.1–2.3 we have

$$\psi = \int y \ p(y \mid a = 1) \ d\eta(y) - \int \mu(l, 0) \ p(l \mid a = 1) \ d\nu(l),$$

where η is a dominating measure for the outcome distribution; this follows from the same logic as in Hahn (1998) and elsewhere. Thus ψ is identified under Assumptions 2.1–2.3 in any study design that identifies $p(y \mid l, a)$ and $p(l \mid a = 1)$. Since these densities are components of the density of distribution Q given in (2.1), it follows that ψ is identified under matched cohort sampling.

As discussed by Breslow et al. (2000) in the context of case-control studies, this fact alone also implies that influence functions for ψ under sampling from Q are equivalent to those under sampling from P, but with densities under distribution Q replacing those under P. For the sake of completeness, we follow Breslow et al. (2000) and prove this result explicitly in the Appendix. To do so we use the same approach as Hahn (1998), with theory developed by Robins and Rotnitzky (1995) and Robins et al. (1995) and discussed in more detail elsewhere (Bickel et al., 1993; van der Laan and Robins, 2003; Tsiatis, 2006). The result can also be derived by weighting the efficient influence function under P by the term $b(W, A)$ as discussed above.

Theorem 2.1 *The efficient influence function for the effect on the treated ψ under a nonparametric model with distribution Q is*

$$\varphi(\mu, \xi; \psi) = \frac{A}{q(a=1)}\Big\{Y - \mu(L,0) - \psi\Big\} - \frac{1-A}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\Big\{Y - \mu(L,0)\Big\}.$$

A simple estimator based on the efficient influence function can be formulated by using the efficient influence function φ as an estimating function, after plugging in estimates $\hat{\mu}$ and $\hat{\xi}$ of the nuisance functions, i.e., solving $\mathbb{Q}_n\{\varphi(\hat{\mu}, \hat{\xi}; \psi)\} = 0$ where $\mathbb{Q}_n$ is the empirical measure under Q. For example, $\hat{\mu}(l, 0)$ could be predicted values from a regression of the outcome on covariates using only control subjects, and $\hat{\xi}(l)$ could be predicted values from a logistic regression of treatment on covariates. We show that this estimator is doubly robust and derive its asymptotic properties in the Appendix.

Computationally, such estimators are exactly equivalent to those that would be used in a simple study with standard random sampling. Thus, just as in case-control studies where one can ignore the outcome-dependent sampling and regress outcome on exposure using logistic regression to obtain valid odds ratio estimates, the above result justifies using standard estimators of effects on the treated in cohort studies with exposure-dependent sampling and matching. In particular, one can use propensity score-based estimators as usual even though the propensity score $\pi(l)$ is not identified under matched cohort sampling. In the Appendix we discuss estimation of effect modification among the treated.

2.5. Efficiency and Design

The semiparametric efficiency bound under sampling from Q is the variance of the efficient influence function from Theorem 2.1. Letting $\sigma^2(l, a) = \text{var}(Y \mid L = l, A = a)$, it is shown in the Appendix that this efficiency bound can be expressed as

$$B_Q = \frac{\Omega + \Sigma_1}{q(a=1)} + \frac{p(a=0)}{p(a=1)} \frac{\Sigma_0^*}{q(a=0)}, \tag{2.2}$$

where $\Omega = \text{var}\{\mu(L,1) - \mu(L,0) \mid A = 1\}$, $\Sigma_1 = \mathbb{E}\{\sigma^2(L,1) \mid A = 1\}$, and $\Sigma_0^* = \mathbb{E}\{\varsigma(W) \mid A = 0\}$ with $\varsigma(w) = \mathbb{E}[\sigma^2(L,0)\pi(L)/\{1 - \pi(L)\} \mid W = w, A = 1]$. Letting $\Sigma_0 = \mathbb{E}[\sigma^2(L,0)\pi(L)/\{1-\pi(L)\} \mid A = 1] = \mathbb{E}\{\varsigma(W) \mid A = 1\}$, the efficiency bound under P can be similarly expressed as $B_P = (\Omega + \Sigma_1 + \Sigma_0)/p(a=1)$.

The expressions for the bounds B_Q and B_P can simplify in certain cases; we will consider three such settings here. The simplest is one in which there are no covariates, i.e., $L = \emptyset$. Then $\pi(l) = p(a=1)$ so that $\Sigma_0^* = \Sigma_0 = \text{var}(Y \mid A = 0)p(a=1)/p(a=0)$, and it also follows that $\Omega = 0$. Another setting of interest is when there are no matching variables, i.e., $W = \emptyset$. Then we again have $\Sigma_0^* = \Sigma_0$, but without further simplification. Lastly we also consider full matching, i.e., $W = L$. Then we have $\Sigma_0^* = \Sigma_0^r$, where $\Sigma_0^r = \mathbb{E}\{\sigma^2(L,0)p(a=1)/p(a=0) \mid A = 1\}$ is the value of Σ_0 we would see in a study had all subjects been randomized to treatment with probability $p(a=1)$ regardless of covariates.

Using the above expressions for B_Q and B_P, it follows that $B_Q < B_P$ if and only if

$$\Sigma_0^* < \frac{q(a=0)}{p(a=0)} \left\{ \Sigma_0 - \frac{p(a=1) - q(a=1)}{q(a=1)} \Big(\Omega + \Sigma_1 \Big) \right\}.$$

Clearly, there always exists a cohort study that can match the efficiency bound under random sampling, since random sampling is equivalent to a cohort study with no matching and with $q(a=1) = p(a=1)$. In the next theorem we show the more interesting result that there almost always exists a cohort study that is strictly more efficient than random sampling.

Theorem 2.2 *Suppose $p(a=1) \neq (\Omega + \Sigma_1)/(\Omega + \Sigma_1 + \Sigma_0)$. Then there exists a cohort study that is more efficient than random sampling for estimation of ψ. For example, an* *efficiency bound strictly smaller than B_P can be attained via an unmatched cohort study with*

$$\min\left\{p(a=1), \frac{\Omega+\Sigma_1}{\Omega+\Sigma_1+\Sigma_0}\right\} < q(a=1) < \max\left\{p(a=1), \frac{\Omega+\Sigma_1}{\Omega+\Sigma_1+\Sigma_0}\right\}.$$

A proof of the above result is given in the Appendix. To illustrate, consider a simple cohort study with no covariates and let $\sigma_a^2 = \text{var}(Y \mid A=a)$. Then any cohort study with

$$p(a=1) < q(a=1) < \frac{p(a=0)\sigma_1^2}{p(a=0)\sigma_1^2 + p(a=1)\sigma_0^2},$$

or the inequalities reversed, yields a smaller efficiency bound than random sampling. If treatment is very rare or very common then nearly any cohort study will be more efficient than random sampling, since then the condition approximates $0 < q(a=1) < 1$.

Matching can provide even more opportunities for efficiency gains. Consider two cohort studies, one without matching, i.e., $W=\emptyset$, yielding efficiency bound B_Q^u and the other fully matched, i.e., $W=L$, yielding efficiency bound B_Q^m. The difference between efficiency bounds then equals

$$B_Q^u - B_Q^m = \frac{1}{q(a=0)}\frac{p(a=0)}{p(a=1)}\mathbb{E}\left[\sigma^2(L,0)\left\{\frac{\pi(L)}{1-\pi(L)} - \frac{p(a=1)}{p(a=0)}\right\} \;\middle|\; A=1\right].$$

If there is no confounding so that $\pi(l) = p(a=1)$, then the bounds are clearly equal and matching does not provide any efficiency gains. However, when there is confounding the above will often be positive since $\pi(l)$ will generally be larger than $p(a=1)$ among the treated. For example, if $\sigma^2(l,0)$ is constant then $B_Q^u \geq B_q^m$ by Jensen's inequality. This suggests that matched cohort studies will in general provide better efficiency than unmatched cohort studies.

In principle one could design a fully efficient matched cohort study by minimizing the expression for B_Q given in (2.2) over choices of $q(a=1) = 1/(k+1)$ and different sets of matching variables W. Optimizing over different matching variables would be difficult in practice, but results for optimizing over $q(a = 1)$ are given in the following theorem.

Theorem 2.3 *Consider a cohort study with a fixed set of (possibly empty) matching variables and given sample size. The optimal number of matches that maximizes efficiency for estimation of* ψ *is* $k_{opt} = [\{p(a = 0)/p(a = 1)\}\{\Sigma_0^*/(\Omega + \Sigma_1)\}]^{1/2}$.

In the simplest matched cohort study with no covariates, this expression simplifies to $k_{opt} = \sigma_0/\sigma_1$. Thus for such studies the optimal matching ratio does not depend on the treatment prevalence, and in particular 1:1 matching is optimal if the variance of the outcome is constant across treatment groups. As intuition would suggest, if the variance of the outcome is greater among controls then more matched controls should be used, and if the variance is greater among the treated then fewer matched controls should be used.

2.6. Simulations and Illustration

2.6.1. Simulation Study

To explore finite-sample properties we adapt the simulation setup from Kang and Schafer (2007). Specifically we simulated $L_j \sim N(0, 1)$ for $j = 1, ..., 4$, $\pi(l) = \text{expit}(-1.7 - l_1 + 0.5l_2 - 0.25l_3 - 0.1l_4)$ so that $p(a = 1) = 0.20$, and $Y = \mu(L, A) + \epsilon$ for $\mu(l, a) = 200 + 13.7l_1 + 13.7\sum_j l_j + 10a$, and $\epsilon \sim N(0, 1)$ so that $\psi = 10$. We generated matched cohort studies with $q(a = 1) = 0.5$ and $W = I(L_1 > 0)$, which ensures that $q(a = 1 \mid l)$ follows a logistic model with covariates w and l. For each simulated dataset we applied inverse-probability-weighted, regression, and doubly robust estimators, with confidence intervals computed via sandwich standard errors. To misspecify models we transformed L as in Kang and Schafer (2007).

As shown in Table 1, the inverse-probability-weighted and regression estimators were biased when relying on misspecified models, while the doubly robust estimator performed well as long as at least one model was correct. The doubly robust and regression estimators had

Table 1: Bias, variance, and coverage based on 500 simulated 1:1 matched cohort studies

		Correct model							
		Neither		Treatment		Outcome		Both	
n	Est.	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov
100	IPW	−20 (129)	98	6 (172)	97	−20 (129)	98	6 (172)	97
	Reg	−54 (29)	68	−54 (29)	68	0 (2.3)	95	0 (2.3)	95
	DR	−41 (36)	76	−7 (35)	94	0 (2.5)	94	0 (2.7)	92
1000	IPW	−27 (83)	84	−1 (95)	96	−27 (83)	84	−1 (95)	96
	Reg	−55 (30)	0	−55 (30)	0	0 (2.2)	95	0 (2.2)	95
	DR	−41 (33)	4	−1 (30)	94	0 (2.4)	95	0 (2.5)	96

IPW, inverse probability weighted; Reg, regression; DR, doubly robust; n, sample size; SE, empirical standard error multiplied by $n^{1/2}$; Cov, coverage (%).

similar efficiency when the outcome model was correct; when only the treatment model was correct the doubly robust estimator was more efficient than the inverse-probability-weighted estimator. Coverage was near 95% except under misspecification. In the Appendix we give further results comparing with random sampling and different matching ratios.

2.6.2. Application

Here we analyze the 3:1 matched cohort study by Ingelsson et al. (2011) discussed in Section 2.2. We used the same three estimators as in the simulation study, with logistic regression models for the treatment, i.e., hysterectomy, and outcome, i.e., cardiovascular disease within 10 years after enrollment. The matching covariates were birthyear and county of residence, and the unmatched covariates were socioeconomic status and age at enrollment. For simplicity we assumed independent censoring. As shown in Table 2, assuming no unmeasured confounding we estimate that hysterectomy yielded a statistically significant 0.55% increased risk of cardiovascular disease within 10 years, among those who underwent hysterectomy.

We also used the formulas from Section 2.5 to analyze efficiency, by estimating the terms in the bound B_Q. For simplicity we assumed $p(w \mid a) = p(w)$ and focused on varying $p(a)$. We estimate that 3:1 matched cohort sampling yields a smaller efficiency bound than random sampling if $p(a = 1) < 23\%$, and is more than twice as efficient if $p(a = 1) < 7\%$.

Table 2: Hysterectomy and 10-year cardiovascular risk

Method	Estimate (%)	SE (%)	95% CI	p-value
IP-weighted	0.47	0.093	(0.29, 0.65)	< 0.001
Regression	0.55	0.092	(0.37, 0.73)	< 0.001
Doubly robust	0.55	0.092	(0.37, 0.73)	< 0.001

CI, confidence interval; SE, standard error; IP, inverse-probability.

We also estimate that 3:1 matching is optimal if $p(a = 1) = 3\%$, and that full matching using socioeconomic status and age is beneficial if $p(a = 1) < 23\%$. More details are in the Appendix.

CHAPTER 3 : NONPARAMETRIC METHODS FOR DOUBLY ROBUST ESTIMATION OF CONTINUOUS TREATMENT EFFECTS

3.1. Abstract

Continuous treatments (e.g., doses) arise often in practice, but many available causal effect estimators are limited by either requiring parametric models for the effect curve, or by not allowing doubly robust covariate adjustment. We develop a novel kernel smoothing approach that requires only mild smoothness assumptions on the effect curve, and still allows for misspecification of either the treatment density or outcome regression. We derive asymptotic properties and give a procedure for data-driven bandwidth selection. The methods are illustrated via simulation and in a study of the effect of nurse staffing on hospital readmissions penalties.

3.2. Introduction

Continuous treatments or exposures (such as dose, duration, and frequency) arise very often in practice, especially in observational studies. Importantly, such treatments lead to effects that are naturally described by curves (e.g., dose-response curves) rather than scalars, as might be the case for binary treatments. Two major methodological challenges in continuous treatment settings are (1) to allow for flexible estimation of the dose-response curve (for example to discover underlying structure without imposing a priori shape restrictions), and (2) to properly adjust for high-dimensional confounders (i.e., pre-treatment covariates related to treatment assignment and outcome).

Consider a recent example involving the Hospital Readmissions Reduction Program, instituted by the Centers for Medicare & Medicaid Services in 2012, which aimed to reduce preventable hospital readmissions by penalizing hospitals with excess readmissions. McHugh et al. (2013) were interested in whether nurse staffing (measured in nurse hours per patient day) affected hospitals' risk of excess readmissions penalty. The left panel of Figure 2 shows data for 2976 hospitals, with nurse staffing (the ‘treatment’) on the x-axis, whether each hospital was penalized (the outcome) on the y-axis, and a loess curve fit to the data (without any adjustment). One way to characterize effects is to imagine setting all hospitals’ nurse staffing to the same level, and seeing if changes in this level yield changes in excess readmissions risk. Such questions cannot be answered by simply comparing hospitals’ risk of penalty across levels of nurse staffing, since hospitals differ in many important ways that could be related to both nurse staffing and excess readmissions (e.g., size, location, teaching status, among many other factors). The right panel of Figure 2 displays the extent of these hospital differences, showing for example that hospitals with more nurse staffing are also more likely to be high-technology hospitals and see patients with higher socioeconomic status. To correctly estimate the effect curve, and fairly compare the risk of readmissions penalty at different nurse staffing levels, one must adjust for hospital characteristics appropriately.

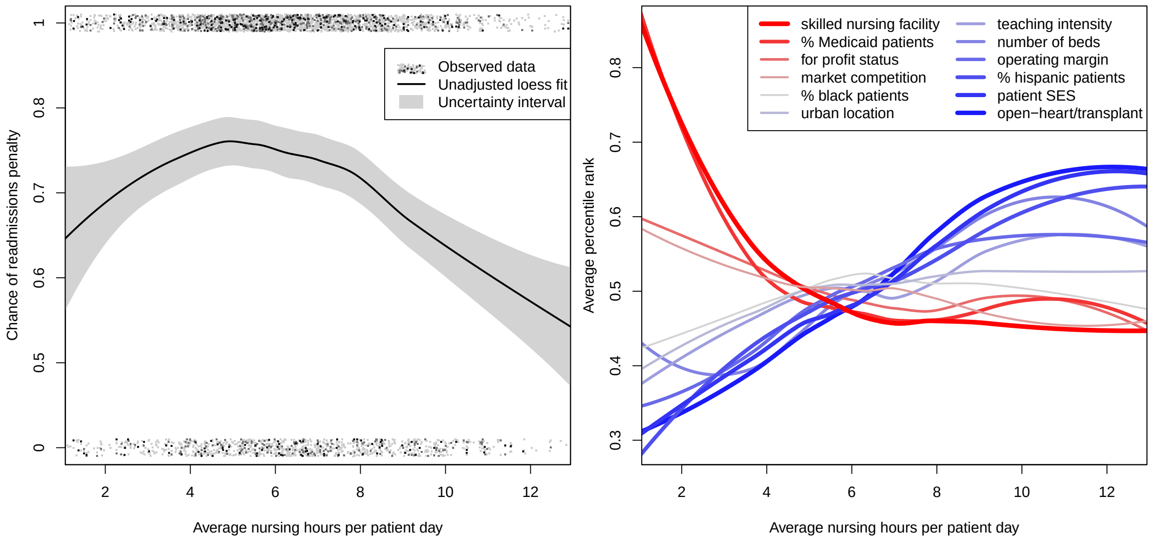

Figure 2: Left panel: Observed treatment and outcome data with unadjusted loess fit. Right panel: Average covariate value as a function of exposure, after transforming to percentiles to display on common scale.

In practice, the most common approach for estimating continuous treatment effects is based on regression modeling of how the outcome relates to covariates and treatment (e.g., Imbens (2004), Hill (2011)). However, this approach relies entirely on correct specification of the outcome model, does not incorporate available information about the treatment mechanism, and is sensitive to the curse of dimensionality by inheriting the rate of convergence of the outcome regression estimator. Hirano and Imbens (2004), Imai and van Dyk (2004), and Galvao and Wang (2015) adapted propensity score-based approaches to the continuous treatment setting, but these similarly rely on correct specification of at least a model for treatment (e.g., the conditional treatment density).

In contrast, semiparametric doubly robust estimators (Robins and Rotnitzky, 2001; van der Laan and Robins, 2003) are based on modeling both the treatment and outcome processes and, remarkably, give consistent estimates of effects as long as one of these two nuisance processes is modeled well enough (not necessarily both). Beyond giving two independent chances at consistent estimation, doubly robust methods can also attain faster rates of convergence than their nuisance (i.e., outcome and treatment process) estimators when both models are consistently estimated; this makes them less sensitive to the curse of dimensionality and can allow for inference even after using flexible machine learning-based adjustment. However, standard semiparametric doubly robust methods for dose-response estimation rely on parametric models for the effect curve, either by explicitly assuming a parametric dose-response curve (Robins, 2000; van der Laan and Robins, 2003), or else by projecting the true curve onto a parametric working model (Neugebauer and van der Laan, 2007). Unfortunately, the first approach can lead to substantial bias under model misspecification, and the second can be of limited practical use if the working model is far away from the truth.

Recent work has extended semiparametric doubly robust methods to more complicated nonparametric and high-dimensional settings. In a foundational paper, van der Laan and Dudoit (2003) proposed a powerful cross-validation framework for estimator selection in general censored data and causal inference problems. Their empirical risk minimization approach allows for global nonparametric modeling in general semiparametric settings involving complex nuisance parameters. For example, Díaz and van der Laan (2013) considered global modeling in the dose-response curve setting, and developed a doubly robust substitution estimator of risk. In nonparameric problems it is also important to consider non-global learning methods, e.g., via local and penalized modeling (Györfi et al., 2002). Rubin and van der Laan (2005, 2006a,b) proposed extensions to such paradigms in numerous important problems, but the former considered weighted averages of dose-response curves and the latter did not consider doubly robust estimation.

In this paper we present a new approach for causal dose-response estimation that is doubly robust without requiring parametric assumptions, and which can naturally incorporate general machine learning methods. The approach is motivated by semiparametric theory for a particular stochastic intervention effect and a corresponding doubly robust mapping. Our method has a simple two-stage implementation that is fast and easy to use with standard software: in the first stage a pseudo-outcome is constructed based on the doubly robust mapping, and in the second stage the pseudo-outcome is regressed on treatment via off-the-shelf nonparametric regression and machine learning tools. We provide asymptotic results for a kernel version of our approach under weak assumptions, which only require mild smoothness conditions on the effect curve and allow for flexible data-adaptive estimation of relevant nuisance functions. We also discuss a simple method for bandwidth selection based on cross-validation. The methods are illustrated via simulation, and in the study discussed earlier about the effect of hospital nurse staffing on excess readmission penalties.

3.3. Background

3.3.1. Data and notation

Suppose we observe an independent and identically distributed sample $(\mathbf{Z}_1, ..., \mathbf{Z}_n)$ where $\mathbf{Z} = (\mathbf{L}, A, Y)$ has support $\mathcal{Z} = (\mathcal{L} \times \mathcal{A} \times \mathcal{Y})$. Here $\mathbf{L}$ denotes a vector of covariates, A a continuous treatment or exposure, and Y some outcome of interest. We characterize causal effects using potential outcome notation (Rubin, 1974), and so let Y^a denote the potential outcome that would have been observed under treatment level a.

We denote the distribution of $\mathbf{Z}$ by P, with density $p(\mathbf{z}) = p(y \mid \mathbf{l}, a)p(a \mid \mathbf{l})p(\mathbf{l})$ with respect to some dominating measure. We let $\mathbb{P}_n$ denote the empirical measure so that empirical averages $n^{-1}\sum_i f(\mathbf{Z}_i)$ can be written as $\mathbb{P}_n\{f(\mathbf{Z})\} = \int f(\mathbf{z})d\mathbb{P}_n(\mathbf{z})$. To simplify the presentation we denote the mean outcome given covariates and treatment with $\mu(\mathbf{l}, a) = \mathbb{E}(Y \mid \mathbf{L} = \mathbf{l}, A = a)$, denote the conditional treatment density given covariates with $\pi(a \mid \mathbf{l}) = \frac{\partial}{\partial a}P(A \leq a \mid \mathbf{L} = \mathbf{l})$, and denote the marginal treatment density with $\varpi(a) = \frac{\partial}{\partial a}P(A \leq a)$. Finally, we use $||f|| = \{\int f(\mathbf{z})^2 dP(\mathbf{z})\}^{1/2}$ to denote the $L_2(P)$ norm, and we use $||f||_{\mathcal{X}} = \sup_{x \in \mathcal{X}} |f(x)|$ to denote the uniform norm of a generic function f over $x \in \mathcal{X}$.

3.3.2. Identification

In this paper our goal is to estimate the effect curve $\theta(a) = \mathbb{E}(Y^a)$. Since this quantity is defined in terms of potential outcomes that are not directly observed, we must consider assumptions under which it can be expressed in terms of observed data. A full treatment of identification in the presence of continuous random variables was given by Gill and Robins (2001); we refer the reader there for details. The assumptions most commonly employed for identification are as follows (the following must hold for any $a \in \mathcal{A}$ at which $\theta(a)$ is to be identified).

Assumption 3.1 *Consistency: $A = a$ implies $Y = Y^a$.*

Assumption 3.2 *Positivity: $\pi(a \mid \mathbf{l}) \geq \pi_{min} > 0$ for all $\mathbf{l} \in \mathcal{L}$.*

Assumption 3.3 *Ignorability: $\mathbb{E}(Y^a \mid \mathbf{L}, A) = \mathbb{E}(Y^a \mid \mathbf{L})$.*

Assumptions 3.1–3.3 can all be satisfied by design in randomized trials, but in observational studies they may be violated and are generally untestable. The consistency assumption ensures that potential outcomes are defined uniquely by a subject's own treatment level and not others' levels (i.e., no interference), and also not by the way treatment is administered (i.e., no different versions of treatment). Positivity says that treatment is not assigned deterministically, in the sense that every subject has some chance of receiving treatment level a, regardless of covariates; this can be a particularly strong assumption with continuous treatments. Ignorability says that the mean potential outcome under level a is the same across treatment levels once we condition on covariates (i.e., treatment assignment is unrelated to potential outcomes within strata of covariates), and requires sufficiently many relevant covariates to be collected. Using the same logic as with discrete treatments, it is straightforward to show that under Assumptions 3.1–3.3 the effect curve $\theta(a)$ can be identified with observed data as

$$\theta(a) = \mathbb{E}\{\mu(\mathbf{L}, a)\} = \int_{\mathcal{L}} \mu(\mathbf{l}, a) \ dP(\mathbf{l}). \tag{3.1}$$

Even if we are not willing to rely on Assumptions 3.1 and 3.3, it may often still be of interest to estimate $\theta(a)$ as an adjusted measure of association, defined purely in terms of observed data.

3.4. Main Results

In this section we develop doubly robust estimators of the effect curve $\theta(a)$ without relying on parametric models. First we describe the logic behind our proposed approach, which is based on finding a doubly robust mapping whose conditional expectation given treatment equals the effect curve of interest, as long as one of two nuisance parameters is correctly specified. To find this mapping, we derive a novel efficient influence function for a stochastic intervention parameter. Our proposed method is based on regressing this doubly robust mapping on treatment using off-the-shelf nonparametric regression and machine learning methods. We derive asymptotic properties for a particular version of this approach based on local-linear kernel smoothing. Specifically, we give conditions for consistency and asymptotic normality, and describe how to use cross-validation to select the bandwidth parameter in practice.

3.4.1. Setup and doubly robust mapping

If $\theta(a)$ is assumed known up to a finite-dimensional parameter, for example $\theta(a) = \psi_0 + \psi_1 a$ for $(\psi_0, \psi_1) \in \mathbb{R}^2$, then standard semiparametric theory can be used to derive the efficient influence function, from which one can obtain the efficiency bound and an efficient estimator (Bickel et al., 1993; van der Laan and Robins, 2003; Tsiatis, 2006). However, such theory is not directly available if we only assume, for example, mild smoothness conditions on $\theta(a)$ (e.g., differentiability). This is due to the fact that without parametric assumptions $\theta(a)$ is not pathwise differentiable, and root-n consistent estimators do not exist (Bickel et al., 1993; Díaz and van der Laan, 2013). In this case there is no developed efficiency theory.

To derive doubly robust estimators for $\theta(a)$ without relying on parametric models, we adapt semiparametric theory in a novel way similar to the approach of Rubin and van der Laan (2005, 2006a). Our goal is to find a function $\xi(\mathbf{Z}; \pi, \mu)$ of the observed data $\mathbf{Z}$ and nuisance functions (π, μ) such that

$$\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} = \theta(a)$$

if either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$ (not necessarily both). Given such a mapping, off-the-shelf nonparametric regression and machine learning methods could be used to estimate $\theta(a)$ by regressing $\xi(\mathbf{Z}; \hat{\pi}, \hat{\mu})$ on treatment A, based on estimates $\hat{\pi}$ and $\hat{\mu}$.

This doubly robust mapping is intimately related to semiparametric theory and especially the efficient influence function for a particular parameter. Specifically, if $\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} = \theta(a)$ then it follows that $\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu})\} = \psi$ for

$$\psi = \int_{\mathcal{A}} \int_{\mathcal{L}} \mu(\mathbf{l}, a) \varpi(a) \ dP(\mathbf{l}) \ da. \tag{3.2}$$

This indicates that a natural candidate for the unknown mapping $\xi(\mathbf{Z}; \pi, \mu)$ would be a component of the efficient influence function for the parameter ψ, since for regular parameters such as ψ in semi- or non-parametric models, the efficient influence function

$\phi(\mathbf{Z};\pi,\mu)$ will be doubly robust in the sense that $\mathbb{E}\{\phi(\mathbf{Z};\overline{\pi},\overline{\mu})\} = 0$, if either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$ (Robins and Rotnitzky, 2001; van der Laan and Robins, 2003). This implies $\mathbb{E}\{\phi(\mathbf{Z};\pi,\mu)\} = \mathbb{E}\{\xi(\mathbf{Z};\pi,\mu) - \psi\} = 0$ so that $\mathbb{E}\{\xi(\mathbf{Z};\overline{\pi},\overline{\mu})\} = \psi$ if either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$. This kind of logic was first used by Rubin and van der Laan (2005, 2006a) for full data parameters that are functions of covariates rather than treatment (i.e., censoring) variables.

The parameter ψ is also of interest in its own right. In particular, it represents the average outcome under an intervention that randomly assigns treatment based on the density ϖ (i.e., a randomized trial). Thus comparing the value of this parameter to the average observed outcome provides a test of treatment effect; if the values differ significantly, then there is evidence that the observational treatment mechanism impacts outcomes for at least some units. Stochastic interventions were discussed by Díaz and van der Laan (2012), for example, but the efficient influence function for ψ has not been given before under a nonparametric model. Thus in Theorem 3.1 below we give the efficient influence function for this parameter respecting the fact that the marginal density ϖ is unknown.

Theorem 3.1 *Under a nonparametric model, the efficient influence function for ψ defined in* (3.2) *is* $\xi(\mathbf{Z};\pi,\mu) - \psi + \int_{\mathcal{A}}\{\mu(\mathbf{L},a) - \int_{\mathcal{L}}\mu(\mathbf{l},a)dP(\mathbf{l})\}\varpi(a)da$, *where*

$$\xi(\mathbf{Z};\pi,\mu) = \frac{Y - \mu(\mathbf{L},A)}{\pi(A \mid \mathbf{L})}\int_{\mathcal{L}}\pi(A \mid \mathbf{l})\ dP(\mathbf{l}) + \int_{\mathcal{L}}\mu(\mathbf{l},A)\ dP(\mathbf{l}).$$

A proof of Theorem 3.1 is given in the Appendix. Importantly, we also prove that the function $\xi(\mathbf{Z};\pi,\mu)$ satisfies its desired double robustness property, i.e., that $\mathbb{E}\{\xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = a\} = \theta(a)$ if either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$. As mentioned earlier, this motivates estimating the effect curve $\theta(a)$ by estimating the nuisance functions (π,μ), and then regressing the estimated pseudo-outcome

$$\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) = \frac{Y - \hat{\mu}(\mathbf{L},A)}{\hat{\pi}(A \mid \mathbf{L})}\int_{\mathcal{L}}\hat{\pi}(A \mid \mathbf{l})\ d\mathbb{P}_n(\mathbf{l}) + \int_{\mathcal{L}}\hat{\mu}(\mathbf{l},A)\ d\mathbb{P}_n(\mathbf{l})$$

on treatment A using off-the-shelf nonparametric regression or machine learning methods. In the next subsection we describe our proposed approach in more detail, and analyze the properties of an estimator based on kernel estimation.

3.4.2. Proposed Approach

In the previous subsection we derived a doubly robust mapping $\xi(\mathbf{Z}; \pi, \mu)$ for which $\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} = \theta(a)$ as long as either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$. This indicates that doubly robust nonparametric estimation of $\theta(a)$ can proceed with a simple two-step procedure, where both steps can be accomplished with flexible machine learning. To summarize, our proposed method is:

1. Estimate nuisance functions (π, μ) and obtain predicted values.

2. Construct pseudo-outcome $\hat{\xi}(\mathbf{Z}; \hat{\pi}, \hat{\mu})$ and regress on treatment variable A.

We give sample code implementing the above in the Appendix.

In what follows we present results for an estimator that uses kernel smoothing in Step 2. Such an approach is related to kernel approximation of a full-data parameter in censored data settings. Robins and Rotnitzky (2001) gave general discussion and considered density estimation with missing data, while van der Laan and Robins (1998), van der Laan and Yu (2001), and van der Vaart and van der Laan (2006) used the approach for current status survival analysis; Wang et al. (2010) used it implicitly for nonparametric regression with missing outcomes.

As indicated above, however, a wide variety of flexible methods could be used in our Step 2, including local partitioning or nearest neighbor estimation, global series or spline methods with complexity penalties, or cross-validation-based combinations of methods, e.g., Super Learner (van der Laan et al., 2007). In general we expect the results we report in this paper to hold for many such methods. To see why, let $\hat{\theta}$ denote the proposed estimator described above (based on some initial nuisance estimators $(\hat{\pi}, \hat{\mu})$ and a particular regression method in Step 2), and let $\overline{\theta}$ denote an estimator based on an oracle version of the pseudo-outcome $\xi(\mathbf{Z};\overline{\pi},\overline{\mu})$ where $(\overline{\pi},\overline{\mu})$ are the unknown limits to which the estimators $(\hat{\pi},\hat{\mu})$ converge. Then $||\hat{\theta}-\theta|| \leq ||\hat{\theta}-\overline{\theta}|| + ||\overline{\theta}-\theta||$, where the second term on the right can be analyzed with standard theory since $\overline{\theta}$ is a regression of a simple fixed function $\xi(\mathbf{Z};\overline{\pi},\overline{\mu})$ on A, and the first term will be small depending on the convergence rates of $\hat{\pi}$ and $\hat{\mu}$. A similar point was discussed by Rubin and van der Laan (2005, 2006a).

The local linear kernel version of our estimator is $\hat{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_h(a)$, where $\mathbf{g}_{ha}(t) = (1, \frac{t-a}{h})^{\mathrm{T}}$ and

$$\hat{\boldsymbol{\beta}}_h(a) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\ \mathbb{P}_n\left[K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \mathbf{g}_{ha}(A)^{\mathrm{T}}\boldsymbol{\beta}\right\}^2\right] \tag{3.3}$$

for $K_{ha}(t) = h^{-1}K\{(t-a)/h\}$ with K a standard kernel function (e.g., a symmetric probability density) and h a scalar bandwidth parameter. This is a standard local linear kernel regression of $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ on A. For overviews of kernel smoothing see, e.g., Fan and Gijbels (1996), Wasserman (2006), and Li and Racine (2007). Under near-violations of positivity, the above estimator could potentially lie outside the range of possible values for $\theta(a)$ (e.g., if Y is binary); thus we present a targeted minimum loss-based estimator (TMLE) in the Appendix, which does not have this problem. Alternatively one could project onto a logistic model in (3.3).

3.4.3. Consistency of Kernel Estimator

In Theorem 3.2 below we give conditions under which the proposed kernel estimator $\hat{\theta}_h(a)$ is consistent for $\theta(a)$, and also give the corresponding rate of convergence. In general this result follows if the bandwidth decreases with sample size slowly enough, and if either of the nuisance functions π or μ is estimated well enough (not necessarily both). The rate of convergence is a sum of two rates: one from standard nonparametric regression problems (depending on the bandwidth h), and another coming from estimation of the nuisance functions π and μ.

Theorem 3.2 *Let $\overline{\pi}$ and $\overline{\mu}$ denote fixed functions to which $\hat{\pi}$ and $\hat{\mu}$ converge in the sense that $||\hat{\pi}-\overline{\pi}||_{\mathcal{Z}} = o_p(1)$ and $||\hat{\mu}-\overline{\mu}||_{\mathcal{Z}} = o_p(1)$, and let $a \in \mathcal{A}$ denote a point in the interior of the compact support $\mathcal{A}$ of A. Along with Assumption 3.2 (Positivity), assume the following:*

1. *Either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$.*
2. *The bandwidth $h = h_n$ satisfies $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$.*
3. *K is a continuous symmetric probability density with support $[-1, 1]$.*
4. *$\theta(a)$ is twice continuously differentiable, and both $\varpi(a)$ and the conditional density of $\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu})$ given $A = a$ are continuous as functions of a.*
5. *The estimators $(\hat{\pi}, \hat{\mu})$ and their limits $(\overline{\pi}, \overline{\mu})$ are contained in uniformly bounded function classes with finite uniform entropy integrals (as defined in the Appendix), with $1/\hat{\pi}$ and $1/\overline{\pi}$ also uniformly bounded.*

Then

$$|\hat{\theta}_h(a) - \theta(a)| = O_p\left(\frac{1}{\sqrt{nh}} + h^2 + r_n(a)s_n(a)\right)$$

where

$$\sup_{t:|t-a|\leq h} ||\hat{\pi}(t \mid \mathbf{L}) - \pi(t \mid \mathbf{L})|| = O_p\Big(r(n)\Big)$$

$$\sup_{t:|t-a|\leq h} ||\hat{\mu}(\mathbf{L}, t) - \mu(\mathbf{L}, t)|| = O_p\Big(s(n)\Big)$$

are the 'local' rates of convergence of $\hat{\pi}$ and $\hat{\mu}$ near $A = a$.

A proof of Theorem 3.2 is given in the Appendix. The required conditions are all quite weak. Condition (a) is arguably the most important of the conditions, and says that at least one of the estimators $\hat{\pi}$ or $\hat{\mu}$ must be consistent for the true π or μ in terms of the uniform norm. Since only one of the nuisance estimators is required to be consistent (not both), Theorem 3.2 shows the double robustness of the proposed estimator $\hat{\theta}_h(a)$. Conditions (b), (c), and (d) are all common in standard nonparametric regression problems, while condition (e) involves the complexity of the estimators $\hat{\pi}$ and $\hat{\mu}$ (and their limits), and is a usual minimal regularity condition for problems involving nuisance functions.

Condition (b) says that the bandwidth parameter h decreases with sample size but not too quickly (so that $nh^3 \to \infty$). This is a standard requirement in local linear kernel smoothing (Fan and Gijbels, 1996; Wasserman, 2006; Li and Racine, 2007). Note that since $nh = nh^3/h^2$, it is implied that $nh \to \infty$; thus one can view nh as a kind of effective or local sample size. Roughly speaking, the bandwidth h needs to go to zero in order to control bias, while the local sample size nh (and nh^3) needs to go to infinity in order to control variance. We postpone more detailed discussion of the bandwidth parameter until a later subsection, where we detail how it can be chosen in practice using cross-validation. Condition (c) puts some minimal restrictions on the kernel function. It is clearly satisfied for most common kernels, including the uniform kernel $K(u) = I(|u| \leq 1)/2$, the Epanechnikov kernel $K(u) = (3/4)(1 - u^2)I(|u| \leq 1)$, and a truncated version of the Gaussian kernel $K(u) = I(|u| \leq 1)\phi(u)/\{2\Phi(1)-1\}$ with ϕ and Φ the density and distribution functions for a standard normal random variable. Condition (d) restricts the smoothness of the effect curve $\theta(a)$, the density of $\varpi(a)$, and the conditional density given $A = a$ of the limiting pseudo-outcome $\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu})$. These are standard smoothness conditions imposed in nonparametric regression problems. By assuming more smoothness of $\theta(a)$, bias-reducing (higher-order) kernels could achieve faster rates of convergence and even approach the parametric root-n rate (see for example Fan and Gijbels (1996), Wasserman (2006), and others).

Condition (e) puts a mild restriction on how flexible the nuisance estimators (and their corresponding limits) can be, although such uniform entropy conditions still allow for a wide array of data-adaptive estimators and, importantly, do not require the use of parametric models. Andrews (1994) (Section 4), van der Vaart and Wellner (1996) (Sections 2.6–2.7), and van der Vaart (2000) (Examples 19.6–19.12) discuss a wide variety of function classes with finite uniform entropy integrals. Examples include standard parametric classes of functions indexed by Euclidean parameters (e.g., parametric functions satisfying a Lipschitz condition), smooth functions with uniformly bounded partial derivatives, Sobolev classes of functions, as well as convex combinations or Lipschitz transformations of any such sets of functions. The uniform entropy restriction in condition (e) is therefore not a very strong restriction in practice; however, it could be further weakened via sample splitting techniques (see Chapter 27 of van der Laan and Rose (2011)).

The convergence rate given in the result of Theorem 3.2 is a sum of two components. The first, $1/\sqrt{nh} + h^2$, is the rate achieved in standard nonparametric regression problems without nuisance functions. Note that if h tends to zero slowly, then $1/\sqrt{nh}$ will tend to zero quickly but h^2 will tend to zero more slowly; similarly if h tends to zero quickly, then h^2 will as well, but $1/\sqrt{nh}$ will tend to zero more slowly. Balancing these two terms requires $h \sim n^{-1/5}$ so that $1/\sqrt{nh} \sim h^2 \sim n^{-2/5}$. This is the optimal pointwise rate of convergence for standard nonparametric regression on a single covariate, for a twice continuously differentiable regression function.

The second component, $r_n(a)s_n(a)$, is the product of the local rates of convergence (around $A = a$) of the nuisance estimators $\hat{\pi}$ and $\hat{\mu}$ towards their targets π and μ. Thus if the nuisance function estimates converge slowly (due to the curse of dimensionality), then the convergence rate of $\hat{\theta}_h(a)$ will also be slow. However, since the term is a product, we have two chances at obtaining fast convergence rates, showing the bias-reducing benefit of doubly robust estimators. The usual explanation of double robustness is that, even if $\hat{\mu}$ is misspecified so that $s_n(a) = O(1)$, then as long as $\hat{\pi}$ is consistent, i.e., $r_n(a) = o(1)$, we will still have consistency since $r_n(a)s_n(a) = o(1)$. But this idea also extends to settings when both $\hat{\pi}$ and $\hat{\mu}$ are consistent. For example suppose $h \sim n^{-1/5}$ so that $1/\sqrt{nh} + h^2 \sim n^{-2/5}$, and suppose $\hat{\pi}$ and $\hat{\mu}$ are locally consistent with rates $r_n(a) = n^{-2/5}$ and $s_n(a) = n^{-1/10}$. Then the product is $r_n(a)s_n(a) = O(n^{-1/2}) = o(n^{-2/5})$ and the contribution from the nuisance functions is asymptotically negligible, in the sense that the proposed estimator has the same convergence rate as an infeasible estimator with known nuisance functions. Contrast this with non-doubly-robust plug-in estimators whose convergence rate generally matches that of the nuisance function estimator, rather than being faster (van der Vaart, 2014).

In the Appendix we give some discussion of uniform consistency, which, along with weak convergence, will be pursued in more detail in future work.

3.4.4. Asymptotic Normality of Kernel Estimator

In the next theorem we show that if one or both of the nuisance functions are estimated at fast enough rates, then the proposed estimator is asymptotically normal after appropriate scaling.

Theorem 3.3 *Consider the same setting as Theorem 3.2. Along with Assumption 3.2 (Positivity) and conditions (a)–(e) from Theorem 3.2, also assume that:*

(f) The local convergence rates satisfy $r_n(a)s_n(a) = o_p(1/\sqrt{nh})$.

Then

$$\sqrt{nh}\Big\{\hat{\theta}_h(a) - \theta(a) + b_h(a)\Big\} \rightsquigarrow N\left(0,\ \frac{\sigma^2(a)\int K(u)^2\ du}{\varpi(a)}\right)$$

where $b_h(a) = \theta''(a)(h^2/2)\int u^2K(u)\ du + o(h^2)$, *and*

$$\sigma^2(a) = \mathbb{E}\left[\frac{\tau^2(\mathbf{L},a) + \{\mu(\mathbf{L},a) - \overline{\mu}(\mathbf{L},a)\}^2}{\{\overline{\pi}(a \mid \mathbf{L})/\overline{\varpi}(a)\}^2/\{\pi(a \mid \mathbf{L})/\varpi(a)\}}\right] - \Big\{\theta(a) - \overline{m}(a)\Big\}^2$$

for $\tau^2(\mathbf{l},a) = var(Y \mid \mathbf{L}=\mathbf{l}, A=a)$, $\overline{\varpi}(a) = \mathbb{E}\{\overline{\pi}(a \mid \mathbf{L})\}$, $\overline{m}(a) = \mathbb{E}\{\overline{\mu}(\mathbf{L},a)\}$.

The proof of Theorem 3.3 is given in the Appendix. Conditions (a)–(e) are the same as in Theorem 3.2 and were discussed earlier. Condition (f) puts a restriction on the local convergence rates of the nuisance estimators. This will in general require at least some semiparametric modeling of the nuisance functions. Truly nonparametric estimators of π and μ will typically converge at slow rates due to the curse of dimensionality, and will generally not satisfy the rate requirement in the presence of multiple continuous covariates. Condition (f) basically ensures that estimation of the nuisance functions is irrelevant asymptotically; depending on the specific nuisance estimators used, it could be possible to give weaker but more complicated conditions that allow for a non-negligible asymptotic contribution while still yielding asymptotic normality.

Importantly, the rate of convergence required by condition (g) of Theorem 3.3 is slower than the root-n rate typically required in standard semiparametric settings where the parameter of interest is finite-dimensional and Euclidean. For example, in a standard setting where the support $\mathcal{A}$ is finite, a sufficient condition for yielding the requisite asymptotic negligibility for attaining efficiency is $r_n(a) = s_n(a) = o(n^{-1/4})$; however in our setting the weaker condition $r_n(a) = s_n(a) = o(n^{-1/5})$ would be sufficient if $h \sim n^{-1/5}$. Similarly, if one nuisance estimator $\hat{\pi}$ or $\hat{\mu}$ is computed with a correctly specified generalized additive model, then the other nuisance estimator would ony need to be consistent (without a rate condition). This is because, under regularity conditions and with optimal smoothing, a generalized additive model estimator converges at rate $O_p(n^{-2/5})$ (Horowitz, 2009), so that if the other nuisance estimator is merely consistent we have $r_n(a)s_n(a) = O(n^{-2/5})o(1) = o(n^{-2/5})$, which satisfies condition (f) as long as $h \sim n^{-1/5}$. In standard settings such flexible nuisance estimation would make a non-negligible contribution to the limiting behavior of the estimator, preventing asymptotic normality and root-n consistency.

Under the assumptions of Theorem 3.3, the proposed estimator is asymptotically normal after appropriate scaling and centering. However, the scaling is by the square root of the local sample size $\sqrt{nh}$ rather than the usual parametric rate $\sqrt{n}$. This slower convergence rate is a cost of making fewer assumptions (equivalently, the cost of better efficiency would be less robustness); thus we have a typical bias-variance trade-off. As in standard nonparametric regression, the estimator is consistent but not quite centered at $\theta(a)$; there is a bias term of order $O(h^2)$, denoted $b_h(a)$. In fact the estimator is centered at a smoothed version of the effect curve $\theta_h^*(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\boldsymbol{\beta}_h(a) = \theta(a) + b_h(a)$. This phenomenon is ubiquitous in nonparametric regression, and complicates the process of computing confidence bands. It is sometimes assumed that the bias term is $o(1/\sqrt{nh})$ and thus asymptotically negligible (e.g., by assuming $h = o(n^{-1/5})$ so that $nh^5 \to 0$); this is called undersmoothing and technically allows for the construction of valid confidence bands around $\theta(a)$. However, there is little guidance about how to actually undersmooth in practice, so it is mostly a technical device. We follow Wasserman (2006) and others by expressing uncertainty about the estimator $\hat{\theta}_h(a)$ using confidence intervals centered at the smoothed data-dependent parameter $\theta_h^*(a)$. For example, under the conditions of Theorem 3.3, pointwise Wald 95% confidence intervals can be constructed with $\hat{\theta}_h(a) \pm 1.96\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}^2$ is the $(1,1)$ element of the sandwich variance estimate $\mathbb{P}_n\{\hat{\boldsymbol{\varphi}}_{ha}(\mathbf{Z})^{\otimes 2}\}$ based on the estimated efficient influence function for $\boldsymbol{\beta}_h(a)$ given by

$$\begin{aligned}\hat{\boldsymbol{\varphi}}_{ha}(\mathbf{Z}) = \hat{\mathbf{D}}_{ha}^{-1}\bigg[&\mathbf{g}_{ha}(A)K_{ha}(A)\Big\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \mathbf{g}_{ha}(A)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_h(a)\Big\} \\ &+ \int_{\mathcal{A}} \mathbf{g}_{ha}(t)K_{ha}(t)\Big\{\hat{\mu}(\mathbf{L},t) - \hat{m}(t)\Big\}\hat{\varpi}(t)\ dt\bigg]\end{aligned}$$

for $\hat{\mathbf{D}}_{ha} = \mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\mathbf{g}_{ha}^{\mathrm{T}}\}$, $\hat{m}(t) = \mathbb{P}_n\{\hat{\mu}(\mathbf{L},t)\}$, $\hat{\varpi}(t) = \mathbb{P}_n\{\hat{\pi}(t \mid \mathbf{L})\}$.

3.4.5. Data-Driven Bandwidth Selection

The choice of smoothing parameter is critical for any nonparametric method; too much smoothing yields large biases and too little yields excessive variance. In this subsection we discuss how to use cross-validation to choose the relevant bandwidth parameter h. In general the method we propose parallels those used in standard nonparametric regression settings, and can give similar optimality properties.

We can exploit the fact that our method can be cast as a non-standard nonparametric regression problem, and borrow from the wealth of literature on bandwidth selection there. Specifically, the logic behind Theorem 3.3 (i.e., that nuisance function estimation can be asymptotically irrelevant) can be adapted to the bandwidth selection setting, by treating the pseudo-outcome $\xi(\mathbf{Z};\hat{\pi},\hat{\mu})$ as known and using for example the bandwidth selection framework from Härdle et al. (1988). These authors proposed a unified selection approach that includes generalized cross-validation, Akaike's information criterion, and leave-one-out cross-validation as special cases, and showed the asymptotic equivalence and optimality of such approaches. In our setting, leave-one-out cross-validation is attractive because of its computational ease. The simplest analog of leave-one-out cross-validation for our problem would be to select the optimal bandwidth from some set $\mathcal{H}$ with

$$\hat{h}_{opt} = \underset{h \in \mathcal{H}}{\arg\min} \sum_{i=1}^{n} \left\{ \frac{\hat{\xi}(\mathbf{Z}_i; \hat{\pi}, \hat{\mu}) - \hat{\theta}_h(A_i)}{1 - \hat{W}_h(A_i)} \right\}^2 ,$$

where $\hat{W}_h(a_i) = (1,0)\mathbb{P}_n\{\mathbf{g}_{ha_i}(A)K_{ha_i}(A)\mathbf{g}_{ha_i}(A)^{\mathrm{T}}\}^{-1}(1,0)^{\mathrm{T}}h^{-1}K(0)$ is the i^{th} diagonal of the so-called smoothing or hat matrix. The properties of this approach can be derived using logic similar to that in Theorem 3.3, e.g., by adapting results from Li and Racine (2004). Alternatively one could split the sample, estimate the nuisance functions in one half, and then do leave-one-out cross-validation in the other half, treating the pseudo-outcomes estimated in the other half as known. We expect these approaches to be asymptotically equivalent to an oracle selector.

An alternative option would be to use the k-fold cross-validation approach of van der Laan and Dudoit (2003) or Díaz and van der Laan (2013). This would entail randomly splitting the data into k parts, estimating the nuisance functions and the effect curve on $(k-1)$ training folds, using these estimates to compute measures of risk on the k^{th} test fold, and then repeating across all k folds and averaging the risk estimates. One would then repeat this process for each bandwidth value h in some set $\mathcal{H}$, and pick that which minimized the estimated cross-validated risk. van der Laan and Dudoit (2003) gave finite-sample and asymptotic results showing that the resulting estimator behaves similarly to an oracle estimator that minimizes the true unknown cross-validated risk. Unfortunately this cross-validation process can be more computationally intensive than the above leave-one-out method, especially if the nuisance functions are estimated with flexible computation-heavy methods. However this approach will be crucial when incorporating general machine learning and moving beyond linear kernel smoothers.

3.5. Simulation Study

We used simulation to examine the finite-sample properties of our proposed methods. Specifically we simulated from a model with normally distributed covariates

$$\mathbf{L} = (L_1, ..., L_4)^{\mathrm{T}} \sim N(0, \mathbf{I}_4),$$

Beta distributed exposure

$$(A/20) \mid \mathbf{L} \sim \text{Beta}\{\lambda(\mathbf{L}), 1 - \lambda(\mathbf{L})\},$$
$$\text{logit } \lambda(\mathbf{L}) = -0.8 + 0.1L_1 + 0.1L_2 - 0.1L_3 + 0.2L_4,$$

and a binary outcome

$$Y \mid \mathbf{L}, A \sim \text{Bernoulli}\{\mu(\mathbf{L}, A)\},$$
$$\text{logit } \mu(\mathbf{L}, A) = 1 + (0.2, 0.2, 0.3, -0.1)\mathbf{L} + A(0.1 - 0.1L_1 + 0.1L_3 - 0.13^2A^2).$$

The above setup roughly matches the data example from the next section. Figure 3 shows a plot of the effect curve $\theta(a) = \mathbb{E}\{\mu(\mathbf{L}, a)\}$ induced by the simulation setup, along with treatment versus outcome data for one simulated dataset (with $n = 1000$).

To analyze the simulated data we used three different estimators: a marginalized regression (plug-in) estimator $\hat{m}(a) = \mathbb{P}_n\{\hat{\mu}(\mathbf{L}, a)\}$, and two different versions of the proposed local linear kernel estimator. Specifically we used an inverse-probability-weighted approach first developed by Rubin and van der Laan (2006b), which relies solely on a treatment model estimator $\hat{\pi}$ (equivalent to setting $\hat{\mu} = 0$), and the standard doubly robust version that used both estimators $\hat{\pi}$ and $\hat{\mu}$. To model the conditional treatment density π we used logistic regression to estimate the parameters of the mean function $\lambda(\mathbf{l})$; we separately considered correctly specifying this mean function, and then also misspecifying the mean function by transforming the covariates with the same covariate transformations as in Kang and Schafer (2007). To estimate the outcome model μ we again used logistic regression, considering a

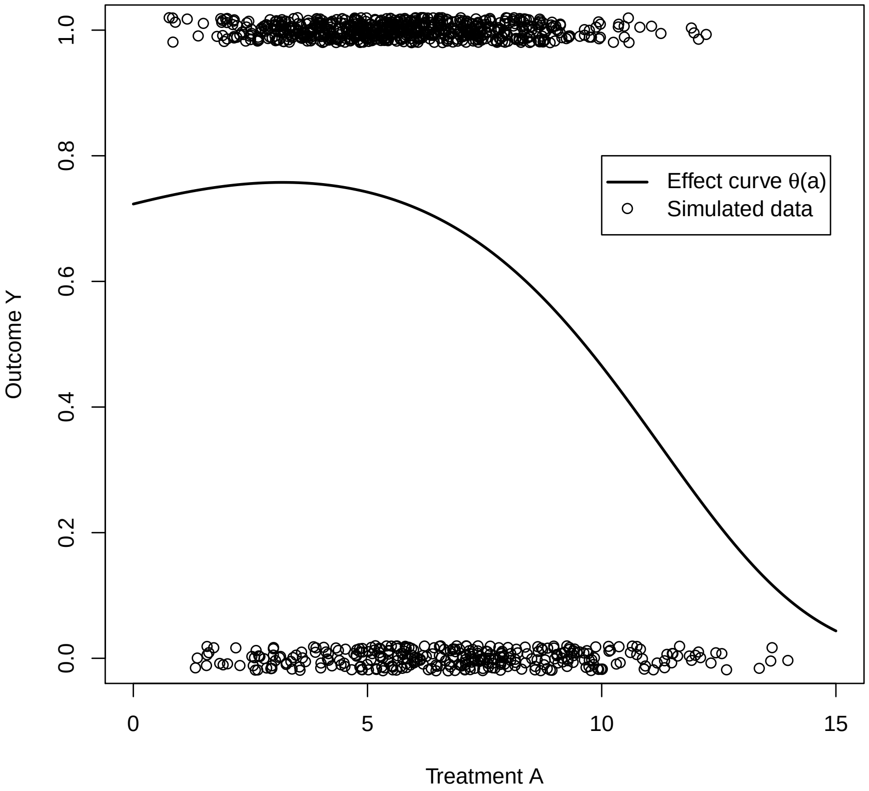

Figure 3: Plot of effect curve induced by simulation setup, with treatment and outcome data from one simulated dataset with $n = 1000$.

correctly specified model and then a misspecified model that used the same transformed covariates as with π and also left out the cubic term in a (but kept all other interactions). To select the bandwidth we used the leave-one-out approach proposed in Section 3.4.5, which treats the pseudo-outcomes as known. For comparison we also considered an oracle approach that picked the bandwidth by minimizing the oracle risk $\mathbb{P}_n[\{\theta(A) - \hat{\theta}_h(A)\}^2]$. In both cases we found the minimum bandwidth value in the range $\mathcal{H} = [0.01, 50]$ using numerical optimization.

We generated 500 simulated datasets for each of three sample sizes, $n = 100$, 1000, and 10000. To assess the quality of the estimates across simulations we calculated empirical

Table 3: Integrated bias and root mean squared error (500 simulations)

		Bias (RMSE) when correct model is:			
n	*Method*	*Neither*	*Treatment*	*Outcome*	*Both*
100	Reg	2.67 (5.54)	2.67 (5.54)	0.62 (5.25)	0.62 (5.25)
	IPW	2.26 (8.49)	1.64 (8.57)	2.26 (8.49)	1.64 (8.57)
	IPW*	2.26 (7.36)	1.58 (7.37)	2.26 (7.36)	1.58 (7.37)
	DR	2.23 (6.27)	1.01 (6.28)	1.12 (5.92)	1.10 (6.50)
	DR*	2.12 (5.48)	1.00 (5.36)	1.03 (5.08)	1.02 (5.65)
1000	Reg	2.62 (3.07)	2.62 (3.07)	0.06 (1.53)	0.06 (1.53)
	IPW	2.38 (3.97)	0.86 (2.94)	2.38 (3.97)	0.86 (2.94)
	IPW*	2.11 (3.44)	0.70 (2.34)	2.11 (3.44)	0.70 (2.34)
	DR	2.03 (3.11)	0.75 (2.39)	0.74 (2.53)	0.68 (2.25)
	DR*	1.84 (2.67)	0.64 (1.88)	0.61 (1.78)	0.58 (1.78)
10000	Reg	2.65 (2.70)	2.65 (2.70)	0.02 (0.47)	0.02 (0.47)
	IPW	2.36 (3.42)	0.33 (1.09)	2.36 (3.42)	0.33 (1.09)
	IPW*	2.24 (3.28)	0.35 (0.85)	2.24 (3.28)	0.35 (0.85)
	DR	1.81 (2.35)	0.26 (0.86)	0.20 (1.21)	0.25 (0.78)
	DR*	1.76 (2.27)	0.31 (0.68)	0.24 (1.10)	0.29 (0.64)
Notes:	Bias / RMSE = integrated mean bias / root mean squared error; IPW = inverse probability weighted; Reg = regression; DR = doubly robust; * = uses oracle bandwidth.				

bias and root mean squared error at each point, and integrated across the function with respect to the density of A. Specifically, letting $\hat{\theta}_s(a)$ denote the estimated curve at point a in simulation s ($s = 1, ..., S$ with $S = 500$), we estimated the integrated absolute mean bias and root mean squared error with

$$\widehat{\text{Bias}} = \int_{\mathcal{A}^*} \left| \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}_s(a) - \theta(a) \right| \varpi(a) \; da,$$

$$\widehat{\text{RMSE}} = \int_{\mathcal{A}^*} \left[\frac{1}{S} \sum_{s=1}^{S} \{\hat{\theta}_s(a) - \theta(a)\}^2 \right]^{1/2} \varpi(a) \; da.$$

In the above $\mathcal{A}^*$ denotes a trimmed version of the support of A, excluding 10% of mass at the boundaries. Note that the above integrands (except for the density) correspond to the usual definitions of absolute mean bias and root mean squared error for estimation of a single scalar parameter (e.g., the curve at a single point).

The simulation results are given in Table 3 (both the integrated bias and root mean squared error are multiplied by 100 for easier interpretation). Estimators with stars (e.g., IPW*) denote those with bandwidths selected using the oracle risk. When both $\hat{\pi}$ and $\hat{\mu}$ were misspecified, all estimators gave substantial integrated bias and mean squared error (although the doubly robust estimator consistently performed better than the other estimators in this setting). Similarly, all estimators had relatively large mean squared error in the small sample size setting ($n = 100$) due to lack of precision, although differences in bias were similar to those at moderate and small sample sizes ($n = 1000, 10000$). Specifically the regression estimator gave small bias when $\hat{\mu}$ was correct and large bias when $\hat{\mu}$ was misspecified, while the inverse-probability-weighted estimator gave small bias when $\hat{\pi}$ was correct and large bias when $\hat{\pi}$ was misspecified. However, the doubly robust estimator gave small bias as long as either $\hat{\pi}$ or $\hat{\mu}$ was correctly specified, even if one was misspecified.

The inverse-probability-weighted estimator was least precise, although it had smaller mean squared error than the misspecified regression estimator for moderate and large sample sizes. The doubly robust estimator was roughly similar to the inverse-probability-weighted estimator when the treatment model was correct, but gave less bias and was more precise, and was similar to the regression estimator when the outcome model was correct (but slightly more biased and less precise). In general the estimators based on the oracle-selected bandwidth were similar to those using the simple leave-one-out approach, but gave marginally less bias and mean squared error for small and moderate sample sizes. The benefits of the oracle bandwidth were relatively diminished with larger sample sizes.

3.6. Application

In this section we apply the proposed methodology to estimate the effect of nurse staffing on hospital readmissions penalties, as discussed in the Introduction. In the original paper, McHugh et al. (2013) used a matching approach to control for hospital differences, and found that hospitals with more nurse staffing were less likely to be penalized; this suggests increasing nurse staffing to help curb excess readmissions. However, their analysis only considered the effect of higher nurse staffing versus lower nurse staffing, and did not explore the full effect curve; it also relied solely on matching for covariate adjustment, i.e., was not doubly robust.

In this analysis we use the proposed kernel smoothing approach to estimate the full effect curve flexibly, while also allowing for doubly robust covariate adjustment. We use the same data on 2976 acute care hospitals as in McHugh et al. (2013); full details are given in the original paper. The covariates $\mathbf{L}$ include hospital size, teaching intensity, not-for-profit status, urban versus rural location, patient race proportions, proportion of patients on Medicaid, average socioeconomic status, operating margins, a measure of market competition, and whether open heart or organ transplant surgery is performed. The treatment A is nurse staffing hours, measured as the ratio of registered nurse hours to adjusted patient days, and the outcome Y indicates whether the hospital was penalized due to excess readmissions. Excess readmissions are calculated by the Centers for Medicare & Medicaid Services and aim to adjust for the fact that different hospitals see different patient populations. Without unmeasured confounding, the quantity $\theta(a)$ represents the proportion of hospitals that would have been penalized had all hospitals changed their nurse staffing hours to level a. Otherwise $\theta(a)$ can be viewed as an adjusted measure of the relationship between nurse staffing and readmissions penalties.

For the conditional density $\pi(a \mid \mathbf{l})$ we assumed a model $A = \lambda(\mathbf{L}) + \gamma(\mathbf{L})\varepsilon$, where ε has mean zero and unit variance given the covariates, but otherwise has an unspecified density. We flexibly estimated the conditional mean function $\lambda(\mathbf{l}) = \mathbb{E}(A \mid \mathbf{L} = \mathbf{l})$ and variance function $\gamma(\mathbf{l}) = \text{var}(A \mid \mathbf{L} = \mathbf{l})$ by combining linear regression, multivariate adaptive regression splines, generalized additive models, Lasso, and boosting, using the cross-validation-based Super Learner algorithm (van der Laan et al., 2007), in order to reduce chances of model misspecification. A standard kernel approach was used to estimate the density of ε.

For the outcome regression $\mu(\mathbf{l}, a)$ we again used the Super Learner approach, combining logistic regression, multivariate adaptive regression splines, generalized additive models, Lasso, and boosting. To select the bandwidth parameter h we used the leave-one-out approach discussed in Section 3.4.5. We considered regression, inverse-probability-weighted, and doubly robust estimators, as in the simulation study in Section 3.5. The two hospitals (<0.1%) with smallest inverse-probability weights were removed as outliers. For the doubly robust estimator we also computed pointwise uncertainty intervals (i.e., confidence intervals around the smoothed parameter $\theta_h^*(a)$; see Section 3.4.4) using a Wald approach based on the empirical variance of the estimating function values.

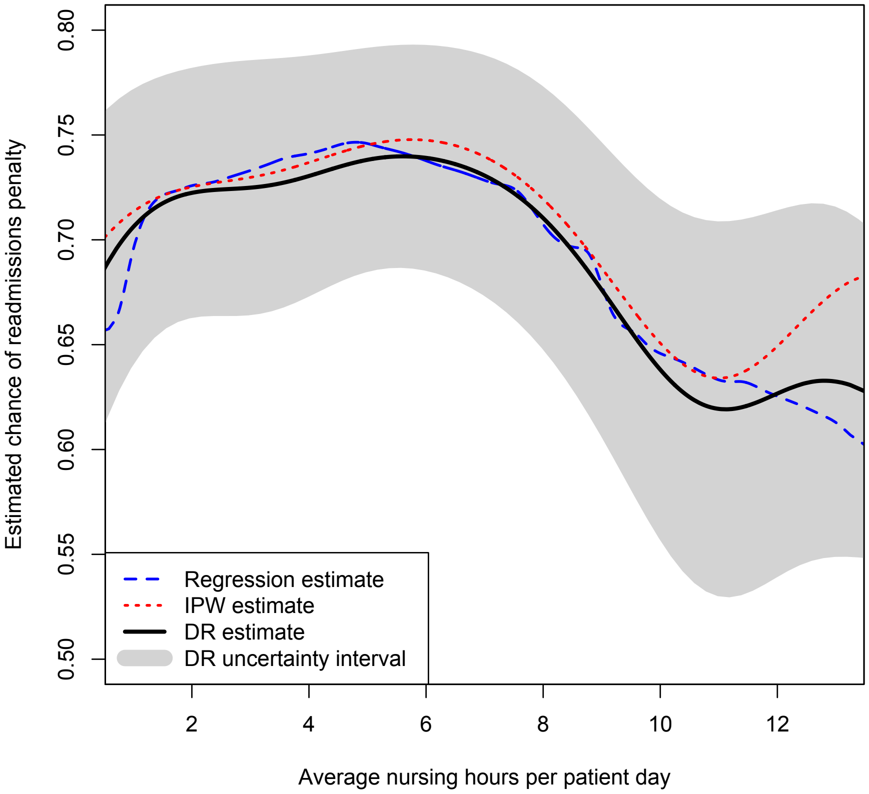

Figure 4: Estimated effects of nurse staffing on readmissions penalties.

A plot showing the results from the three estimators (with uncertainty intervals for the proposed doubly robust estimator) is given in Figure 4. In general the three estimators were very similar. For less than 5 average nurse staffing hours the adjusted chance of penalization was estimated to be roughly constant, around 73%, but at 5 hours chances of penalization began decreasing, reaching approximately 60% when nurse staffing reached 11 hours. This suggests that adding nurse staffing hours may be particularly beneficial in the 5-10 hour range, in terms of reducing risk of readmissions penalization; most hospitals (65%) lie in this range in our data.

3.7. Discussion

In this paper we developed a novel approach for estimating the average effect of a continuous treatment; importantly the approach allows for flexible doubly robust covariate adjustment without requiring any parametric assumptions about the form of the effect curve, and can incorporate general machine learning and nonparametric regression methods. We presented a novel efficient influence function for a stochastic intervention parameter defined within a nonparametric model; this influence function motivated the proposed approach, but may also be useful to estimate on its own. In addition we provided asymptotic results (including rates of convergence and asymptotic normality) for a particular kernel estimation version of our method, which only require the effect curve to be twice continuously differentiable, and allows for flexible data-adaptive estimation of nuisance functions. These results show the double robustness of the proposed approach, since either a conditional treatment density or outcome regression model can be misspecified and the proposed estimator will still be consistent as long as one such nuisance function is correctly specified. We also showed how double robustness can result in smaller second-order bias even when both nuisance functions are consistently estimated. Finally, we proposed a simple and fast data-driven cross-validation approach for bandwidth selection, found favorable finite sample properties of our proposed approach in a simulation study, and applied the kernel estimator to examine the effects of hospital nurse staffing on excess readmissions penalty.

This paper integrates semiparametric (doubly robust) causal inference with nonparametric function estimation and machine learning, helping to bridge the “huge gap between classical semiparametric models and the model in which nothing is assumed” (van der Vaart, 2014). In particular our work extends standard nonparametric regression by allowing for complex covariate adjustment and doubly robust estimation, and extends standard doubly robust causal inference methods by allowing for nonparametric smoothing. Many interesting problems arise in this gap between standard nonparametric and semiparametric inference, leading to many opportunities for important future work, especially for complex non-regular target parameters that are not pathwise differentiable. In the context of this paper, in future work it will be useful to study uniform distributional properties of our proposed estimator (e.g., weak convergence), as well as its role in testing and inference (e.g., for constructing tests that have power to detect a wide array of deviations from the null hypothesis of no effect of a continuous treatment).

CHAPTER 4 : SEMIPARAMETRIC CAUSAL INFERENCE WITH THE LOCAL INSTRUMENTAL VARIABLE CURVE

4.1. Abstract

Instrumental variables are commonly used to estimate effects of a treatment afflicted by unmeasured confounding, and in practice instrumental variables are often continuous (e.g., measures of distance, or treatment preference). However, available methods for continuous instrumental variables have important limitations: they either require restrictive parametric assumptions for identification, or else rely on modeling both the outcome and treatment process well. In this work we develop robust semiparametric estimators of a "local" effect curve among compliers, i.e., the effect among those who would take treatment for instrument values above some threshold and not below. The proposed methods do not require parametric assumptions, incorporate information about the instrument mechanism, allow for flexible data-adaptive estimation of effect modification, and are robust to misspecification of either the instrument or treatment/outcome processes (i.e., are doubly robust). We discuss asymptotic properties under weak conditions, and use the methods to study infant mortality effects of neonatal intensive care units with high versus low technical capacity, using travel time as an instrument.

4.2. Introduction

Instrumental variables are commonly used to estimate effects of treatments that are afflicted by unmeasured confounding. Instruments are special variables that influence treatment, but are themselves unconfounded and do not directly affect outcomes, allowing the recovery of some causal information from data that might otherwise be unusable. In practice, instruments are often continuous (e.g., measures of distance, or treatment preference), but most available methods only consider instruments that are discrete (and typically binary). Further, methods that do allow for continuous instruments have important limitations.

Classical instrumental variable methods (e.g., standard two-stage least squares), which were developed in a structural equation model framework, allow for continuous instruments but require strong parametric assumptions for identification, assume that treatment effects do not vary across units, and also require correct parametric models for how both the treatment and outcome processes depend on covariates and instruments (Wooldridge, 2010). Alternatively, Robins and others (Robins, 1989, 1994; Hernán and Robins, 2006; Tan, 2010) developed approaches in the potential outcomes framework that can also handle continuous instruments, but which allow heterogeneous treatment effects, and also permit doubly robust covariate adjustment. Doubly robust instrumental variable methods are consistent as long as either the instrument mechanism or the treatment/outcome mechanisms are correctly modeled (not necessarily both), and they can also yield fast root-n convergence rates and inference even when using flexible nonparametric methods for covariate adjustment. However, the methods of Robins et al. still require parametric assumptions for identification; they target treatment effects on the treated, and achieve identification with dimension-reducing parametric assumptions that restrict how heterogeneous treatment effects can be. As noted for example by Tchetgen Tchetgen and Vansteelandt (2013), this kind of approach is problematic because a priori information about the parametric form of underlying causal structure is rarely available, and misspecification could lead to large biases that cannot be detected with data.

An alternative approach is to replace dimension-reducing homogeneity assumptions with a monotonicity assumption (Robins, 1989; Imbens and Angrist, 1994), which rules out the possibility that any units would respond oppositely to encouragement from the instrument. In other words, there can be units who are encouraged by the instrument (e.g., by taking treatment if the instrument is received and taking control if not, in the binary instrument case), as well as units who do not respond at all to the instrument (e.g., by always taking treatment or always taking control, regardless of received instrument value), but there cannot be units who defy encouragement from the instrument (e.g., by taking control if the instrument is received and taking treatment if not). This assumption is often plausible in practice and, importantly, permits nonparametric identification of causal effects among compliers (i.e., those who do respond to encouragement from the instrument). However, most work relying on monotonicity assumes binary or discrete instruments (Imbens and Angrist, 1994; Abadie, 2003; Tan, 2006; Ogburn et al., 2015). An important exception is a strand of work that has focused on estimating local instrumental variable curves, i.e., effects among units who would comply right at a given threshold value of the instrument (Heckman, 1997; Heckman and Vytlacil, 1999; Glickman and Normand, 2000; Heckman and Vytlacil, 2005).

This literature on local instrumental variable approaches arose out of a latent index or selection model framework (Vytlacil, 2002), and is unique in allowing for continuous instruments while still permitting nonparametric identification. However there are important limitations. First, current local instrumental variable estimands are fully conditional on all measured covariates, even though in many cases effect modification is not of particular scientific interest, or else it is only of interest for a small subset of covariates. In addition to having closer ties to the scientific question, marginal effects are also less ambitious parameters that can typically be estimated more robustly than fully conditional parameters. Thus it is advantageous to specifically target such marginal effects when they are of interest. Also, as noted by van der Laan and Robins (2003), when we target fully conditional effects, we are at the whim of whatever confounders happen to arise in our given dataset; in contrast, targeting marginal parameters allows us to frame our scientific questions a priori. However, adapting current methods to settings where full effect modification is not of interest (by marginalizing available conditional estimators) is very difficult if effect modification is of interest for some covariate subset. Further, even when effect modification is not of interest, marginalization leads to awkward and uninterpretable models, as discussed in a related setting by Ogburn et al. (2015).

Second, available approaches for estimating the local instrumental variable curve rely on modeling how both the treatment and outcome depend on covariates and instrument (Basu et al., 2007; Carneiro and Lee, 2009; Carneiro et al., 2010), and typically use restrictive parametric models. This is problematic since parametric models are often relied upon based on convenience, rather than real substantive knowledge, and can yield severe bias if misspecified. Conversely, fully nonparametric approaches are sensitive to the curse of dimensionality and typically do not yield estimators that attain root-n convergence rates or valid inference (without impractical undersmoothing). Further, in the instrumental variable setting there may be some information available about how the instrument depends on covariates, but this is not incorporated at all in approaches that rely solely on treatment and outcome models (whether parametric or nonparametric).

In this work we make several new advances. First, we formulate marginal versions of the local instrumental variable curve within a nonparametric potential outcomes framework, and consider working models for this curve. This eases interpretability by allowing analysts to incorporate background knowledge on the actual parameter of interest, without having to specify models for quantities that are not of direct scientific interest. Second, we develop semiparametric theory for such settings. Third, we propose semiparametric efficient and doubly robust estimators, which incorporate information about the instrument mechanism, give two chances at consistency, and also can converge at fast root-n rates even after machine learning-based covariate adjustment. We also derive asymptotic properties under weak empirical process conditions. Fourth, we develop a doubly robust cross-validation approach for model selection in high-dimensional settings, which is crucial for learning the instrumental variable curve from data. To the best of our knowledge, these results are all novel, and we believe our paper is the first to use cross-validation for model selection in an instrumental variable setting. We explore finite-sample properties via simulation, and use our methods in a study of effects of high-level neonatal intensive care units on infant mortality, using travel time as an instrument.

4.3. Preliminaries

4.3.1. Data & Notation

Suppose we observe an independent and identically distributed sample $(\mathbf{O}_1, ..., \mathbf{O}_n)$ with $\mathbf{O} = (\mathbf{X}, Z, A, Y)$, where $\mathbf{X}$ is a vector of covariates, Z is a continuous instrument for a binary treatment A, and Y is some real-valued outcome of interest. The covariates $\mathbf{X} = (\mathbf{V}, \mathbf{W})$ are partitioned into potential effect modifiers of interest $\mathbf{V}$ and other covariates $\mathbf{W} = \mathbf{X} \setminus \mathbf{V}$ not of interest but for which adjustment is still necessary. The choice of $\mathbf{V}$ is based purely on the scientific question, so that if effect modification is not of interest one can simply select $\mathbf{V} = \emptyset$. We characterize causal effects using potential outcome notation (Rubin, 1974), and so let Y^a (and Y^{za}) denote the potential outcomes that would have been observed had treatment level $A = a$ (and instrument level $Z = z$) been received. Similarly we let A^z denote the potential treatment that would have been observed under instrument level $Z = z$. A directed acyclic graph showing the data structure is given in Figure 5.

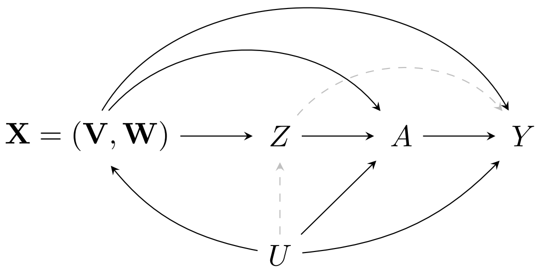

Figure 5: Directed acyclic graph showing covariates $\mathbf{X}$ (partitioned into potential effect modifiers of interest $\mathbf{V}$ and other variables $\mathbf{W}$), instrument Z, treatment A, outcome Y, and unmeasured variables U. Gray dotted arrows indicate relationships that are assumed absent by identifying assumptions.

We let P denote the distribution of $\mathbf{O}$, and for simplicity suppose it has a density with respect to some dominating measure given by $p(\mathbf{o}) = p(y \mid \mathbf{x}, z, a)p(a \mid \mathbf{x}, z)p(z \mid \mathbf{x})p(\mathbf{x})$. In general we write the density of a variable T under P evaluated at value z as $p(T = z)$, except when the density we are referring to is unambiguous (e.g., $p(t)$ denotes the density of T at t), and we denote the support of a variable T as $\text{supp}(T)$. Finally we use some additional notation to simplify the presentation. Specifically we use $\pi(z \mid \mathbf{x}) = p(Z = z \mid \mathbf{X} = \mathbf{x})$ to denote the density of the instrument given covariates (i.e., the instrument propensity score), $\mu(\mathbf{x}, z) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}, Z = z)$ and $\lambda(\mathbf{x}, z) = \mathbb{E}(A \mid \mathbf{X} = \mathbf{x}, Z = z)$ to denote the outcome and treatment regression functions, respectively, and $m(\mathbf{v}, z) = \mathbb{E}\{\mu(\mathbf{X}, z) \mid \mathbf{V} = \mathbf{v}\}$ and $\ell(\mathbf{v}, z) = \mathbb{E}\{\lambda(\mathbf{X}, z) \mid \mathbf{V} = \mathbf{v}\}$ to denote the marginalized versions of the regression functions. We let $\mathbb{P}_n$ denote the empirical measure so that empirical averages can be written as $n^{-1} \sum_i f(\mathbf{O}_i) = \int f(\mathbf{o})\, d\mathbb{P}_n(\mathbf{o}) = \mathbb{P}_n\{f(\mathbf{O})\}$. The notation $|| \cdot ||$ denotes the Euclidean norm $||\boldsymbol{\beta}|| = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{1/2}$ when applied to a vector, but denotes the $L_2(P)$ norm $||f|| = \{\int f(\mathbf{o})^2 \; dP(\mathbf{o})\}^{1/2}$ when applied to a (scalar) function.

4.3.2. Monotonicity

Before defining the causal estimand of interest and considering identifying assumptions, it will be helpful to discuss the concept of monotonicity, which was introduced in various forms by Robins (1989) and Imbens and Angrist (1994), among others. In a classical binary instrument setting (where $\text{supp}(Z) = \{0, 1\}$), monotonicity can be stated succinctly as

$$A^1 \geq A^0 \text{ with probability one.}$$

Monotonicity rules out the possibility that there are troublesome units in the population with $A^0 = 1$ but $A^1 = 0$. Such units are called 'defiers' since they take treatment $A = 1$ when not encouraged by the instrument (i.e., when $Z = 0$), but take control $A = 0$ when they are in fact encouraged (i.e., when $Z = 1$). Thus monotonicity ensures that the population only comprises never-takers ($A^0 = A^1 = 0$), always-takers ($A^0 = A^1 = 1$), and compliers ($A^0 = 0, A^1 = 1$). Monotonicity can often be a reasonable assumption in practice, but not always; it has been discussed extensively in previous work, particularly for binary instruments (see Imbens (2014) and discussion for a nice overview).

A natural way to extend monotonicity to the continuous instrument setting is as follows.

Assumption 4.1 (Monotonicity) *If $z' > z$ then $A^{z'} \geq A^z$ with probability one.*

Under Assumption 4.1, no unit would ever change from treatment to control with an increase in the instrument value; increasing the instrument can either encourage treatment over control or have no effect at all, but it cannot discourage treatment relative to lesser instrument values. Thus the population still comprises never-takers, always-takers, and compliers, but with continuous instruments the compliers can be further partitioned into compliers at given instrument values. In particular, a complier at $Z = z$ would be a unit for which $A^z = 1$ but $A^{z-\delta} = 0$ for any $\delta > 0$. The above continuous version of monotonicity has been employed and discussed by Glickman and Normand (2000) and Vytlacil (2002), for example. Importantly, these authors showed that (when coupled with standard identifying assumptions to be discussed shortly) the above monotonicity assumption is equivalent to the following latent threshold model.

Assumption 1′ (Latent Threshold) $A^z = \mathbb{1}(z \geq T)$ *for an unobserved random threshold T.*

Under the latent threshold model, each complier has some instrument value at which they are encouraged to take treatment, while for any lesser value they would take control. Larger values of the threshold T indicate units that are less willing to take treatment, i.e., less susceptible to encouragement by the instrument. We can thus define the latent threshold T as

$$T = \begin{cases} -\infty & \text{if } A^z = 1 \text{ for all } z \text{ (always-takers)} \\ \inf\{z : A^z = 1\} & \text{if } A^{z'} > A^z \text{ for some } z' > z \text{ (compliers)} \\ \infty & \text{if } A^z = 0 \text{ for all } z \text{ (never-takers).} \end{cases}$$

See Vytlacil (2002) for further discussion and detail.

4.3.3. Estimand & Identification

In this paper our main goal is estimation and inference for the local instrumental variable curve, which we define as

$$\gamma(t, \mathbf{v}) = \mathbb{E}(Y^1 - Y^0 \mid T = t, \mathbf{V} = \mathbf{v}). \tag{4.1}$$

This is the average treatment effect among those with latent threshold $T = t$ (and baseline covariates $\mathbf{V} = \mathbf{v}$), i.e., the effect among those units with $\mathbf{V} = \mathbf{v}$ who would be encouraged to take treatment right when the instrument passes $Z = t$ but not for lesser values. A fully conditional version of the local instrumental variable curve with $\mathbf{V} = \mathbf{X}$ was proposed in the latent index or selection model framework by Heckman (1997), and discussed in detail by Heckman and Vytlacil (1999), Heckman and Vytlacil (2005), and Heckman and Vytlacil (2007), among others. In this framework, the parameter in (4.1) with $\mathbf{V} = \mathbf{X}$ was termed the "marginal treatment effect", and its observed data counterpart the "local instrumental variable" estimand (after employing identifying assumptions). We follow the latter usage in both cases to avoid confusion, since "marginal" often means "averaged".

Throughout we consider standard instrumental variable identifying assumptions, which have been employed for example by Angrist et al. (1996), Tan (2006), Ogburn et al. (2015), and many others; useful overviews and discussions are given for example by Hernán and Robins (2006), Imbens (2014) (with discussion), and Baiocchi et al. (2014).

Assumption 4.2 (Consistency) *$A = A^Z$ and $Y = Y^A$ with probability one.*

Assumption 4.3 (Positivity) *$(z, \mathbf{x}) \in supp(Z, \mathbf{X})$ if $\mathbf{x} \in supp(\mathbf{X})$.*

Assumption 4.4 (Unconfoundedness of Z) *$(Y^z, A^z) \perp\!\!\!\perp Z \mid \mathbf{X}$.*

Assumption 4.5 (Exclusion Restriction) *$Y^{za} = Y^a$ with probability one.*

Consistency means potential treatments A^z and outcomes Y^a are uniquely defined by a unit's own instrument and treatment levels, respectively, and not by others' levels (i.e., no interference), and also not by the way the instrument or treatment are administered (i.e., no different versions). Positivity says that the instrument is not deterministic, in the sense that every unit has some chance of receiving each level of the instrument, regardless of covariates. Unconfoundedness says that the instrument is essentially randomized once we condition on covariates, i.e., that it is unrelated to potential outcomes and treatments under different instrument values. The exclusion restriction says that the instrument only affects outcomes through treatment. Assumptions 4.2–4.5 can hold by design in trials where the instrument is externally randomized by investigators, but in observational studies these assumptions are typically untestable and require justification based on subject matter.

Finally we also employ the following regularity conditions on the latent threshold distribution and local instrumental variable curve.

Assumption 4.6 (Instrumentation) $\inf_t p(t \mid \mathbf{v}) > 0$.

Assumption 4.7 (Continuity) *T is continuously distributed and $\gamma(t, \mathbf{v})$ is continuous in t.*

Instrumentation means there are some units who would be encouraged to take treatment when the instrument passes $Z = t$ (for now we leave the set over which the infimum is taken ambiguous). The following theorem indicates that the local instrumental variable curve can be identified with observed data, under the above assumptions.

Theorem 4.1 *Suppose Assumption $1'$ holds. Let $\mathcal{T} \subset supp(Z)$ denote a compact set on which we wish to identify $\gamma(t, \mathbf{v})$. If Assumptions 4.2–4.5 hold for all $z \in \mathcal{T}$ and Assumptions 4.6–4.7 hold for all $t \in \mathcal{T}$, then the local instrumental variable curve is identified for any $t \in \mathcal{T}$ by*

$$\gamma(t, \mathbf{v}) = \left. \frac{\frac{\partial}{\partial z}\mathbb{E}\{\mathbb{E}(Y \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}}{\frac{\partial}{\partial z}\mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}} \right|_{z=t}. \tag{4.2}$$

A proof of Theorem 4.1 is given in the Supplementary Materials; the logic follows as in more standard settings where Z is discrete and $\mathbf{V} = \mathbf{X}$. Importantly, the local instrumental variable curve can only be identified on subsets of $\text{supp}(Z)$; thus as in the binary instrument setting, we cannot identify effects for never-takers or always-takers with $T = \pm\infty$.

From this point forward, when we write $\gamma(t, \mathbf{v})$ we mean the observed data expression in (4.2), which represents the causal effect given in (4.1) under Assumptions 4.2–4.7 as described in Theorem 4.1. Of course, if the conditions of Theorem 4.1 do not hold then the observed data expression in (4.2) may represent something other than the aforementioned causal effect. For example, if only Assumptions 2–4 hold, then we can only think of the instrument as an unconfounded continuous exposure (or dose), and $\gamma(t, \mathbf{v})$ would represent the ratio of derivatives of the dose-response curves $\mathbb{E}(Y^z \mid \mathbf{V} = \mathbf{v})$ and $\mathbb{E}(A^z \mid \mathbf{V} = \mathbf{v})$.

4.4. Main Results

In this section we develop semiparametric theory for models of the local instrumental variable curve defined in (4.2), use this theory to develop novel doubly robust estimators, describe their asymptotic properties, and finally present cross-validation methods for model selection in high-dimensional settings.

4.4.1. Semiparametric Theory

Suppose we have a parametric model for the local instrumental variable curve, which we write as $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ for some finite-dimensional $\boldsymbol{\psi} \in \mathbb{R}^q$. Importantly, we do not assume this model is necessarily correct, and instead follow Neugebauer and van der Laan (2007), Rosenblum and van der Laan (2010), and others in using a working model approach, by formulating our target estimand as the projection of the true curve $\gamma(t, \mathbf{v})$ onto the posed working model. Specifically, we use the weighted least squares projection given by

$$\boldsymbol{\psi}_0 = \underset{\boldsymbol{\psi} \in \mathbb{R}^q}{\arg\min} \; \mathbb{E}\Big[w(T, \mathbf{V}) \{\gamma(T, \mathbf{V}) - \gamma(T, \mathbf{V}; \boldsymbol{\psi})\}^2 \Big], \tag{4.3}$$

where $w(t, \mathbf{v})$ is some user-specified weight function. Whether to use $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ only for projections or to assume it is the true model is a bias-variance trade-off. If the model happens to be correct, then the projection approach will result in appropriate but generally not fully efficient estimators. However, if the posited model is not correct, then the projection approach is still validly defined as a best-fitting wrong model, while a model-based approach would technically no longer be applicable and can be more difficult to interpret. The projection approach thus formalizes how models are often viewed as approximations in practice.

Note that the above projection depends on the distribution of the latent threshold T. Although the threshold is not observed directly, its distribution is identified in the observed data (under Assumptions 4.1–4.7) by

$$p(t \mid \mathbf{v}) = \begin{cases} \mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = z_{\min}) \mid \mathbf{V} = \mathbf{v}\} & \text{for } t = -\infty \\ \frac{\partial}{\partial z}\mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}|_{z=t} & \text{for } t \in \text{supp}(Z) \\ \mathbb{E}[\mathbb{E}\{(1 - A) \mid \mathbf{X}, Z = z_{\max}\} \mid \mathbf{V} = \mathbf{v}] & \text{for } t = \infty \end{cases} \cdot \tag{4.4}$$

Importantly, the expression for $p(t \mid \mathbf{v})$ in the $t \in \text{supp}(Z)$ case equals the denominator of $\gamma(t, \mathbf{v})$ given in Theorem 4.1.

After characterizing the parameter of interest in terms of observed data as in (4.3), based on the expression in (4.2), it is possible to estimate it using any number of approaches, such as parametric or nonparametric maximum likelihood, or Bayesian methods. In our setting, however, semiparametric approaches have a number of important advantages. First, they can incorporate information about the instrument mechanism, which may be better understood or easier to model than the outcome and treatment mechanisms (which is what a likelihood-based approach would rely on modeling). Second, they allow for double robustness, which means consistent estimation of $\boldsymbol{\psi}$ is possible as long as either the instrument mechanism or the treatment/outcome mechanisms are correctly modeled (not necessarily all three, so either the instrument or the treatment and outcome models can be misspecified). And third, semiparametric doubly robust approaches allow for fast root-n rates of convergence for the parameter of interest $\boldsymbol{\psi}$ even when nuisance functions are estimated at slower rates, e.g., using flexible data-adaptive or machine learning methods. This phenomenon is referred to as orthogonality (or adaptivity) by Chernozhukov et al. (2015), and makes such estimators less sensitive to the curse of dimensionality.

A crucial aspect of developing semiparametric theory and corresponding estimators for a given problem involves characterizing the possible influence functions, and in particular finding the efficient influence function. Many details on semiparametric theory are available elsewhere (Bickel et al., 1993; van der Laan and Robins, 2003; Tsiatis, 2006; Kennedy, in press), so we give only a brief review here. Any regular asymptotically linear estimator minus its target parameter can be expressed as the empirical average of its so-called influence function plus an $o_p(1/\sqrt{n})$ error term. Viewed as elements of a Hilbert space of mean-zero finite-variance functions equipped with covariance norm, the influence functions under a given model lie in the orthogonal complement of the nuisance tangent space. The efficient influence function can then be defined as the influence function with smallest variance, the projection of any influence function onto the tangent space of scores, or as a particular pathwise derivative. The efficient influence function is especially important in practice because its variance is the semiparametric efficiency bound (thus providing a benchmark for efficient estimation), and because it can be used to construct estimators that are doubly robust and potentially semiparametric efficient.

The major challenge in deriving semiparametric theory for the projection parameter in (4.3) is its complexity; namely, it is a weighted projection of a ratio of derivatives of regression functions that are partially marginalized. The next theorem gives the efficient influence function for this parameter.

Theorem 4.2 *Suppose the weight function $w(t, \mathbf{v})$ is continuously differentiable in t and satisfies $w(t, \mathbf{v}) = 0$ for $t \notin int(\mathcal{T})$, with the set $\mathcal{T} \subset supp(Z)$ defined as in Theorem 4.1.*

Also assume that partial derivatives (with respect to $\boldsymbol{\psi}$ and t) of the working model $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ exist and are continuous. Then, under a nonparametric model, the efficient influence function for $\boldsymbol{\psi}$ defined in (4.2) *and* (4.3) *is proportional to*

$$\begin{aligned}\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \boldsymbol{\eta}) = \mathbf{g}_1(Z, \mathbf{V}; \boldsymbol{\psi}) \left\{ \frac{A - \mathbb{E}(A \mid \mathbf{X}, Z)}{p(Z \mid \mathbf{X})} \right\} - \mathbf{g}_2(Z, \mathbf{V}; \boldsymbol{\psi}) \left\{ \frac{Y - \mathbb{E}(Y \mid \mathbf{X}, Z)}{p(Z \mid \mathbf{X})} \right\} \\ + \int_{\mathcal{T}} \Big\{ \mathbf{g}_1(t, \mathbf{V}; \boldsymbol{\psi}) \mathbb{E}(A \mid \mathbf{X}, Z = t) - \mathbf{g}_2(t, \mathbf{V}; \boldsymbol{\psi}) \mathbb{E}(Y \mid \mathbf{X}, Z = t) \Big\} \, dt,\end{aligned} \tag{4.5}$$

where $\boldsymbol{\eta} = (\pi, \lambda, \mu)$ denotes the nuisance functions defined in Section 4.3.1, and $\mathbf{g}_1$ and $\mathbf{g}_2$ are the $(q \times 1)$ vectors

$$\begin{aligned}\mathbf{g}_1(z, \mathbf{v}; \boldsymbol{\psi}) &= \frac{\partial}{\partial t} \left\{ \frac{\partial}{\partial \boldsymbol{\psi}^*} \gamma(t, \mathbf{v}; \boldsymbol{\psi}^*) \Big|_{\boldsymbol{\psi}^* = \boldsymbol{\psi}} w(t, \mathbf{v}) \gamma(t, \mathbf{v}; \boldsymbol{\psi}) \right\} \Big|_{t=z} \\ \mathbf{g}_2(z, \mathbf{v}; \boldsymbol{\psi}) &= \frac{\partial}{\partial t} \left\{ \frac{\partial}{\partial \boldsymbol{\psi}^*} \gamma(t, \mathbf{v}; \boldsymbol{\psi}^*) \Big|_{\boldsymbol{\psi}^* = \boldsymbol{\psi}} w(t, \mathbf{v}) \right\} \Big|_{t=z}.\end{aligned}$$

A proof of Theorem 4.2 is given in the Appendix, and proceeds by showing that the function $\boldsymbol{\varphi}$ is the canonical gradient of the pathwise derivative of $\boldsymbol{\psi}$. Importantly, our derivation of the efficient influence function uses integration by parts to transfer derivatives of the partially marginalized treatment/outcome regression functions to the known model $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ and weight function $w(t, \mathbf{v})$. This means the efficient influence function can be evaluated without analytical differentiation of the regression functions, which, as discussed in detail in the next subsection, makes it much more practical for constructing and implementing estimators. The condition on the user-specified weight function, i.e., that it vanishes outside the interior of the set $\mathcal{T} \subset \text{supp}(Z)$, is required since the local instrumental variable curve $\gamma(t, \mathbf{v})$ is not identified outside of $\mathcal{T}$ as discussed in Theorem 4.1. Roughly speaking, the efficient influence function $\boldsymbol{\varphi}$ can be viewed as consisting of inverse-probability-weighted terms (the added term in the first line in (4.5)) plus an augmentation term (the second line and subtracted terms in the first line in (4.5)). This follows the general structure of influence functions in more common causal inference and missing data problems, although the form of the functions $\mathbf{g}_1$ and $\mathbf{g}_2$ makes the expression less standard.

4.4.2. Proposed Method

Once we have derived the efficient influence function, we can use it to construct estimators that have numerous advantageous properties. A standard approach is to solve an estimating equation based on an estimated version of the efficient influence function; specifically we can use $\boldsymbol{\varphi}$ as an estimating function, with unknown nuisance functions replaced with estimates.

Thus our proposed estimator for a given working model $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ is given by $\hat{\boldsymbol{\psi}}$, defined as the solution in $\boldsymbol{\psi}$ to the estimating equation

$$\mathbb{P}_n\{\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \hat{\boldsymbol{\eta}})\} = \mathbf{0}, \tag{4.6}$$

where $\hat{\boldsymbol{\eta}} = (\hat{\pi}, \hat{\lambda}, \hat{\mu})$ are estimated versions of the three nuisance functions. Another option for constructing estimators based on influence functions is targeted minimum loss-based methodology (van der Laan and Rubin, 2006; van der Laan and Rose, 2011), which has advantages since it yields estimators that respect the bounds of the parameter space. However, our proposed estimating equation approach will also respect any such bounds, as long as the chosen working model does; it is also relatively straightforward to implement.

First consider the simple case where $\mathbf{V} = \emptyset$ (i.e., effect modification is not of interest), and the local instrumental variable curve $\gamma(t)$ is projected onto a constant $\gamma(t; \boldsymbol{\psi}) = \psi$. In this case the target estimand ψ is a simple weighted average of $\gamma(t)$, of the form

$$\psi = \int_{\mathcal{T}} w^*(t)\gamma(t)\ dt$$

with weight $w^*(t) = w(t)p(t) / \int_{\mathcal{T}} w(t)p(t)\ dt$. The quantity ψ can also be viewed as the mean treatment effect (among compliers) in a population where the density of the latent threshold T among compliers equals $w^*(t)$. Let $\boldsymbol{\eta} = (\pi, \lambda, \mu)$ denote the nuisance functions as introduced in Section 4.3.1, with corresponding estimators $\hat{\boldsymbol{\eta}} = (\hat{\pi}, \hat{\lambda}, \hat{\mu})$. Then solving

(4.6) leads to the ratio estimator

$$\hat{\psi} = \frac{\int_{\mathcal{T}} w'(t)\hat{m}(t) \ dt + \mathbb{P}_n \left\{ w'(Z) \frac{Y - \hat{\mu}(\mathbf{X}, Z)}{\hat{\pi}(Z|\mathbf{X})} \right\}}{\int_{\mathcal{T}} w'(t)\hat{\ell}(t) \ dt + \mathbb{P}_n \left\{ w'(Z) \frac{A - \hat{\lambda}(\mathbf{X}, Z)}{\hat{\pi}(Z|\mathbf{X})} \right\}} \tag{4.7}$$

where $\hat{m}(t) = \mathbb{P}_n\{\hat{\mu}(\mathbf{X}, t)\}$ and $\hat{\ell}(t) = \mathbb{P}_n\{\hat{\lambda}(\mathbf{X}, t)\}$ are the marginalized regression functions as in Section 4.3.1. Thus $\hat{\psi}$ is an adjusted version of the regression-based plug-in estimator $\int_{\mathcal{T}} w'(t)\hat{m}(t) \ dt / \int_{\mathcal{T}} w'(t)\hat{\ell}(t) \ dt$, where adding inverse-probability-weighted terms to the numerator and denominator is the adjustment required to obtain double robustness.

More standard instrumental variable estimators are often computed with a two-stage least squares approach, where in the first stage the treatment variable is regressed on the instrument (and covariates) and then in the second stage the outcome is regressed on the predicted values from the first stage (and covariates). In fact, the weighted average estimator in (4.7) can also be constructed with a modified version of such a two-stage least squares approach; this may make it more amenable to practical use. Specifically, the following modified two-stage least squares procedure can be used to compute the weighted average estimator (using pseudo- instrument, treatment, and outcome $w'(Z)$, A^*, and Y^* respectively):

1. Regress $A^* = \frac{A - \hat{\lambda}(\mathbf{X}, Z)}{\pi(Z|\mathbf{X})} + \frac{\mathbb{1}\{w'(Z) \neq 0\}}{w'(Z)} \int_{\mathcal{T}} w'(t)\hat{\lambda}(\mathbf{X}, t) \ dt$ on $w'(Z)$ without an intercept, and obtain predicted values $\hat{A}^*$.

2. Regress $Y^* = \frac{Y - \hat{\mu}(\mathbf{X}, Z)}{\pi(Z|\mathbf{X})} + \frac{\mathbb{1}\{w'(Z) \neq 0\}}{w'(Z)} \int_{\mathcal{T}} w'(t)\hat{\mu}(\mathbf{X}, t) \ dt$ on $\hat{A}^*$, without an intercept.

Then the coefficient in front of $\hat{A}^*$ in the second stage equals $\hat{\psi}$ from (4.7).

Closed-form estimators are also available even when effect modification is of interest, as long as we project onto linear models of the form $\gamma(t, \mathbf{v}; \boldsymbol{\psi}) = \mathbf{h}(t, \mathbf{v})^{\mathrm{T}}\boldsymbol{\psi}$, where $\mathbf{h} : \mathcal{T} \times \mathrm{supp}(\mathbf{V}) \to \mathbb{R}^q$ is a known mapping. Specifically, in such cases the estimator $\hat{\boldsymbol{\psi}}$ defined as the solution to (4.6) is given by

$$\hat{\boldsymbol{\psi}} = \mathbb{P}_n \left[\mathbf{g}_1^*(Z, \mathbf{V}) \left\{ \frac{A - \hat{\lambda}(\mathbf{X}, Z)}{\hat{\pi}(Z \mid \mathbf{X})} \right\} + \int_{\mathcal{T}} \mathbf{g}_1^*(t, \mathbf{V}) \hat{\lambda}(\mathbf{X}, Z) \ dt \right]^{-1}$$
$$\times \mathbb{P}_n \left[\mathbf{g}_2(Z, \mathbf{V}) \left\{ \frac{Y - \hat{\mu}(\mathbf{X}, Z)}{\hat{\pi}(Z \mid \mathbf{X})} \right\} + \int_{\mathcal{T}} \mathbf{g}_2(t, \mathbf{V}) \hat{\mu}(\mathbf{X}, Z) \ dt \right]$$

where $\mathbf{g}_1^*(z, \mathbf{v}) = \frac{\partial}{\partial t}\{\mathbf{h}(t, \mathbf{v}) w(t, \mathbf{v}) \mathbf{h}(t, \mathbf{v})^{\mathrm{T}}\}|_{t=z}$, and $\mathbf{g}_2(z, \mathbf{v}) = \frac{\partial}{\partial t}\{\mathbf{h}(t, \mathbf{v}) w(t, \mathbf{v})\}|_{t=z}$ is as defined in Theorem 4.2. Closed-form expressions will typically not be available for estimators in general non-linear models; however, since such estimators are still defined as estimating equation-based Z-estimators, they can be computed with standard software (for example, one could use the `optim` function in R). Variance estimation and confidence interval construction will be discussed in the next section.

4.4.3. Asymptotic Theory

In this section we discuss the asymptotic properties of our proposed estimation approach. In particular we show that under very weak conditions our estimator is doubly robust and consistent, and that if the nuisance functions are estimated well enough it is asymptotically normal and efficient. Further, asymptotic normality and efficiency are possible even after machine learning-based covariate adjustment. (Our results equally apply to estimators that only solve the efficient influnce function estimating equation asymptotically, up to order $o_p(1/\sqrt{n})$, such as targeted minimum loss-based estimators.)

Theorem 4.3 *Assume that:*

1. $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}) \xrightarrow{p} (\boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}})$, *where* $\overline{\boldsymbol{\eta}} = (\overline{\pi}, \overline{\lambda}, \overline{\mu})$ *with either* $\overline{\pi} = \pi_0$ *or* $(\overline{\lambda}, \overline{\mu}) = (\lambda_0, \mu_0)$.

2. *The sequence of functions* $\hat{\boldsymbol{\varphi}}_n = \boldsymbol{\varphi}(\cdot; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}})$ *and its limit* $\boldsymbol{\varphi}_0 = \boldsymbol{\varphi}(\cdot; \boldsymbol{\psi}_{\mathbf{0}}, \overline{\boldsymbol{\eta}})$ *are contained in a Donsker class with* $||\hat{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0|| = o_p(1)$.

3. *The map* $\boldsymbol{\psi} \to \mathbb{E}\{\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \boldsymbol{\eta})\}$ *is differentiable at* $\boldsymbol{\psi}_0$ *uniformly in* $\boldsymbol{\eta}$ *(around* $\overline{\boldsymbol{\eta}}$*), with invertible derivative matrix* $\mathbf{D}(\boldsymbol{\psi}_0, \boldsymbol{\eta}) \to \mathbf{D}(\boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}}) \equiv \mathbf{D}_0$.

Then the proposed estimator is consistent with rate of convergence

$$||\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0|| = O_p\left\{1/\sqrt{n} + ||\hat{\pi} - \pi_0||\Big(||\hat{\lambda} - \lambda_0|| + ||\hat{\mu} - \mu_0||\Big)\right\}.$$

Suppose further that:

4. $||\hat{\pi} - \pi_0||(||\hat{\lambda} - \lambda_0|| + ||\hat{\mu} - \mu_0||) = o_p(1/\sqrt{n})$.

Then the proposed estimator is asymptotically normal with

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \rightsquigarrow N\Big(\mathbf{0}, \mathbb{E}[\{\mathbf{D}_0^{-1}\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}_0, \boldsymbol{\eta}_0)\}^{\otimes 2}]\Big),$$

and thus semiparametric efficient.

A proof of Theorem 4.3 is given in the Appendix; it follows from standard Z-estimator theory and empirical process results (van der Vaart and Wellner, 1996; van der Vaart, 2002). The first condition indicates the double robustness of our approach, since some of the nuisance estimators $\hat{\boldsymbol{\eta}} = (\hat{\pi}, \hat{\lambda}, \hat{\mu})$ can be misspecified. Specifically, as long as either $\hat{\pi}$ or $(\hat{\lambda}, \hat{\mu})$ is consistent, then the proposed estimator $\hat{\boldsymbol{\psi}}$ will be as well. This gives analysts two chances at consistency, and is particularly important in the instrumental variable setting since it can be easier to model the single instrument density π rather than the two treatment and outcome regression functions (λ, μ), as required in previous approaches.

Conditions 2–3 of Theorem 4.3 are standard regularity conditions for M- and Z-estimators (van der Vaart and Wellner, 1996; van der Vaart, 2000, 2002). Condition 2 puts a mild restriction on the flexibility of the nuisance estimators (and their limits), but Donsker classes cover many complex functions and thus allow $\hat{\boldsymbol{\eta}}$ to be constructed with potentially very flexible estimators. For example, parametric Lipschitz functions are Donsker, but so also are many more complicated function types such as infinite-dimensional smooth functions with bounded partial derivatives, VC classes, Sobolev classes, and functions with bounded uniform sectional variation, as well as convex combinations and Lipschitz transformations of any these classes. More discussion and examples can be found in Sections 2.6–2.7 of van der Vaart and Wellner (1996) and Examples 19.6–19.12 of van der Vaart (2000), as well as in Kennedy (in press). Condition 2 is important because it means we do not have to rely on restrictive parametric models to estimate the potentially complicated and high-dimensional nuisance functions (π, λ, μ), and can instead use more flexible machine learning and data-adaptive methods. Condition 2 can also be weakened in various ways; for example, the Donsker condition really only needs to hold in a shrinking neighborhood of $(\boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}})$, or with high probability as $n \to \infty$. Alternatively we could formulate Condition 2 in terms of weaker entropy or bracketing conditions, or even use sample-splitting to do away with complexity conditions entirely, in the same spirit as Zheng and van der Laan (2010). The differentiability in Condition 3 is standard and required to use a delta method-type result (note that this condition does not require the influence function to be differentiable itself, only its expectation).

Under Conditions 1–3 of Theorem 4.3, the proposed estimator is consistent with rate of convergence given by $1/\sqrt{n} + ||\hat{\pi} - \pi_0||(||\hat{\lambda} - \lambda_0|| + ||\hat{\mu} - \mu_0||)$. Again the double robustness is apparent since we will have consistency, i.e., $||\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0|| = o_p(1)$, as long as either the instrument density is consistently estimated, $||\hat{\pi} - \pi_0|| = o_p(1)$, or the treatment/outcome regressions are, $(||\hat{\lambda} - \lambda_0|| + ||\hat{\mu} - \mu_0||) = o_p(1)$. Note that this result is agnostic about how well the nuisance functions are estimated, since slow rates on the nuisance estimators $\hat{\boldsymbol{\eta}}$ will yield slow rates for the parameter of interest $\hat{\boldsymbol{\psi}}$. Importantly, this result also shows how double robustness is useful even apart from giving two chances at consistency; in particular, if we estimate the regression functions (λ, μ) at slow rates, double robustness gives us a chance to obtain faster rates for $\hat{\boldsymbol{\psi}}$ by consistently estimating π, and vice versa.

For example, if Condition 4 holds so that $||\hat{\pi} - \pi_0||(||\hat{\lambda} - \lambda_0|| + ||\hat{\mu} - \mu_0||) = o_p(1/\sqrt{n})$, and therefore the nuisance estimation is asymptotically negligible, then the parameter of interest $\hat{\boldsymbol{\psi}}$ is root-n consistent, asymptotically normal, and semiparametric efficient. Importantly, Condition 4 can hold even if the nuisance functions are estimated at slower than parametric root-n rates, so that efficient estimation and valid inference is possible for $\boldsymbol{\psi}$ even if we use machine learning-based covariate adjustment, via flexible estimation of the nuisance functions (π, λ, μ). For example, if the nuisance functions (π, λ, μ) are all estimated at faster than $n^{1/4}$ rates, so that $||\hat{\pi} - \pi_0|| = ||\hat{\lambda} - \lambda_0|| = ||\hat{\mu} - \mu_0|| = o_p(n^{-1/4})$, then Condition 4 holds since $o_p(n^{-1/4})o_p(n^{-1/4}) = o_p(1/\sqrt{n})$. Such rates are possible in various flexible models; for instance, under some conditions (Horowitz, 2009) generalized additive model estimators can obtain rates of the form $O_p(n^{-2/5})$, which is $o_p(n^{-1/4})$ since $R_n = O_p(n^{-2/5})$ implies $n^{1/4}R_n = n^{-3/20}n^{2/5}R_n = O_p(n^{-3/20}) = o_p(1)$. Another way Condition 4 can hold is if one of π or (λ, μ) is estimated with a correctly specified parametric model and the other is merely estimated consistently. However, outside of the randomized trial setting, it is typically uncommon to know detailed parametric structure.

If Condition 4 holds, asymptotic confidence intervals can be constructed with the bootstrap, or using a direct estimate of the asymptotic variance given in Theorem 4.3, such as

$$\mathbb{P}_n[\{\hat{\mathbf{D}}^{-1}\boldsymbol{\varphi}(\mathbf{O}; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}})\}^{\otimes 2}]$$

where $\hat{\mathbf{D}} = \mathbb{P}_n\{\partial\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \hat{\boldsymbol{\eta}})/\partial\boldsymbol{\psi}^{\mathrm{T}}\}|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$ is an estimate of the derivative matrix from Condition 3 of Theorem 4.3. If Condition 4 fails to hold but parametric models are used to estimate all three nuisance functions, then the bootstrap is still valid since in this case the contribution from nuisance estimation will be asymptotically linear (an analytic expression could also be derived, using the fact that $\hat{\boldsymbol{\psi}}$ together with the estimated nuisance parameters solve a large system of estimating equations). Inference is more complicated in a truly doubly robust but nonparametric setting, where one nuisance estimator can be misspecified but methods more flexible than parametric models are used to construct the estimates $(\hat{\pi}, \hat{\lambda}, \hat{\mu})$; an approach similar to van der Laan (2014) could potentially be developed to address this issue in our setting, but we leave this to future work.

4.4.4. Model Selection

To this point we have presumed that we have an a priori model $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ in hand, which we either believe represents the truth, or which we just want to use for projections to construct low-dimensional summaries of the truth. However, in practice such a priori models are not available; instead we often aim to learn the true form of the function $\gamma(t, \mathbf{v})$ from data. Thus in this section we propose a doubly robust cross-validation approach for model selection. Model selection is an important issue in causal inference in general, but this is especially the case for the local instrumental variable curve, since the latent threshold T is continuous; thus saturated parametric models are not possible, even when effect modification is not of interest (i.e., $\mathbf{V} = \emptyset$). Specifically, in this section we derive the efficient influence function for the risk of a given candidate estimator, and show how it can be used as a doubly robust loss function in the cross-validation framework developed by van der Laan and Dudoit (2003).

If we knew the true local instrumental variable curve, and the true distribution of the data, a natural way to evaluate the performance of a given candidate estimator $\hat{\gamma}_k$ would be to compute the risk as mean squared error

$$R^*(\hat{\gamma}_k) = \int_{\mathcal{V}} \int_{\mathcal{T}} w(t, \mathbf{v}) \Big\{ \gamma(t, \mathbf{v}) - \hat{\gamma}_k(t, \mathbf{v}) \Big\}^2 \, dP(t, \mathbf{v}).$$

On the other hand, if our only goal was to compare or rank a set of candidate estimators $\{\hat{\gamma}_k : k \in \mathcal{K}\}$, we could equivalently use the pseudo-risk

$$R(\hat{\gamma}_k) = \int_{\mathcal{V}} \int_{\mathcal{T}} w(t, \mathbf{v}) \Big\{ \hat{\gamma}_k(t, \mathbf{v})^2 - 2\gamma(t, \mathbf{v})\hat{\gamma}_k(t, \mathbf{v}) \Big\} \, dP(t, \mathbf{v}), \tag{4.8}$$

since $R(\hat{\gamma}_k) = R^*(\hat{\gamma}_k) - \mathbb{E}\{w(T, \mathbf{V})\gamma(T, \mathbf{V})^2\}$ is simply a shifted version of the mean squared error $R^*(\hat{\gamma}_k)$, where the shift does not depend on the candidate estimator $\hat{\gamma}_k$, and so is irrelevant in evaluating its performance. This is the same phenomenon that has been discussed before in, for example, standard nonparametric regression settings (Wasserman, 2006). However, in these more standard settings it is possible to estimate risk unbiasedly without worrying about nuisance function estimation; in contrast, in our setting the risk parameter $R(\hat{\gamma}_k)$ depends on complex nuisance functions through its dependence on the curve $\gamma(t, \mathbf{v})$ and the distribution of the latent threshold T. Thus estimation of the risk $R(\hat{\gamma}_k)$ will itself require nuisance estimation, and in fact we can treat $R(\hat{\gamma}_k)$ as a parameter in its own right, for which we can develop semiparametric theory and estimation procedures. Thus in the next theorem we give the efficient influence function for the risk $R(\gamma_k)$ for a given fixed candidate γ_k, and go on to show how to use this efficient influence function as a doubly robust loss function for cross-validation-based model selection.

Theorem 4.4 *Consider the same setting and assumptions as in Theorem 4.2. Under a nonparametric model, the efficient influence function for the risk $R(\gamma_k)$ defined in* (4.8) *for a fixed candidate γ_k is given by*

$$\begin{aligned} L(\mathbf{O}; \gamma_k, \boldsymbol{\eta}) = f_1(Z, \mathbf{V}; \gamma_k) \left\{ \frac{Y - \mathbb{E}(Y \mid \mathbf{X}, Z)}{p(Z \mid \mathbf{X})} \right\} - f_2(Z, \mathbf{V}; \gamma_k) \left\{ \frac{A - \mathbb{E}(A \mid \mathbf{X}, Z)}{p(Z \mid \mathbf{X})} \right\} \quad (4.9) \\ + \int_{\mathcal{T}} \Big\{ f_1(t, \mathbf{V}; \gamma_k) \mathbb{E}(Y \mid \mathbf{X}, Z = t) - f_2(t, \mathbf{V}; \gamma_k) \mathbb{E}(A \mid \mathbf{X}, Z = t) \Big\} \, dt \end{aligned}$$

where $\boldsymbol{\eta} = (\pi, \lambda, \mu)$ are the nuisance functions from before, and f_1 and f_2 are defined as

$$f_1(z, \mathbf{v}; \boldsymbol{\psi}) = 2 \frac{\partial}{\partial t} \Big\{ w(t, \mathbf{v}) \gamma_k(t, \mathbf{v}) \Big\} \Big|_{t=z}, \quad f_2(z, \mathbf{v}; \boldsymbol{\psi}) = \frac{\partial}{\partial t} \Big\{ w(t, \mathbf{v}) \gamma_k(t, \mathbf{v})^2 \Big\} \Big|_{t=z}.$$

A proof of Theorem 4.4 is given in the Appendix, and follows similar logic as in the proof of Theorem 4.2. We also show that the efficient influence function $L(\mathbf{O}; \gamma_k, \boldsymbol{\eta})$ is a doubly robust loss function for the risk $R(\gamma_k)$ in the sense that $\mathbb{E}\{L(\mathbf{O}; \gamma_k, \overline{\boldsymbol{\eta}})\} = R(\gamma_k)$ for nuisance function $\overline{\boldsymbol{\eta}} = (\overline{\pi}, \overline{\lambda}, \overline{\mu})$ as long as either $\overline{\pi} = \pi_0$ or $(\overline{\lambda}, \overline{\mu}) = (\lambda_0, \mu_0)$, and not necessarily both. Thus we can use $L(\mathbf{O}; \gamma_k, \boldsymbol{\eta})$ as a doubly robust estimating function, similar to how we used $\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \boldsymbol{\eta})$ in previous sections. However, an additional complication is that we typically do not have an independent sample from which we can generate candidate estimators $\hat{\gamma}_k$, and so need to generate them from the same sample in which we estimate risk. Thus we can use sample-splitting via cross-validation to prevent over-fitting.

In particular, we propose using the loss function in (4.9) for doubly robust model selection following the general approach of van der Laan and Dudoit (2003). First we need to introduce some new notation. Let $\mathbf{S} = (S_1, ..., S_n)$ denote a random variable independent of the sample that splits the data into training ($S_i = 0$) and test ($S_i = 1$) sets. Various cross-validation schemes are covered by different choices of the distribution of $\mathbf{S}$. For example standard v-fold cross-validation arises by allowing the split variable $\mathbf{S}$ to take v different values $\{\mathbf{S}_1, ..., \mathbf{S}_v\}$, each with equal probability $1/v$, where $\sum_i S_{iv} = n/v$ for all v and $\sum_v S_{iv} = 1$ for all i, so that test sets are all of size n/v and each unit is only used in one test set. Now that we have notation for splitting the data into training and test sets, we can define $\mathbb{P}^0_{\mathbf{s}}$ and $\mathbb{P}^1_{\mathbf{s}}$ as the sub-empirical distributions for the training data $\{i : S_i = 0\}$ and test data $\{i : S_i = 1\}$, respectively, for a given split $\mathbf{S} = \mathbf{s}$. Therefore, for example, $\hat{\boldsymbol{\eta}}(\mathbb{P}^0_{\mathbf{s}})$ denotes the nuisance function estimates based only on the training set data, and $\hat{\gamma}_k(\mathbb{P}^0_{\mathbf{s}})$ denotes the local instrumental variable curve estimate based only on the training set data (which also depends on the nuisance function estimates constructed from the training data).

The cross-validation selection approach of van der Laan and Dudoit (2003) is very similar to standard cross-validation, but incorporates extra steps for nuisance function estimation; it proceeds as follows. For a given split $\mathbf{s}$ and a given candidate estimator $\hat{\gamma}_k$, we first estimate the nuisance functions with the training data to obtain $\hat{\boldsymbol{\eta}}(\mathbb{P}^0_{\mathbf{s}})$, and then estimate the local instrumental variable curve with the training data to obtain $\hat{\gamma}_k(\mathbb{P}^0_{\mathbf{s}})$. At this point we can evaluate the loss function L for any observation $\mathbf{O}_i$ based on these training estimates, and thus we do so on the test data $\mathbb{P}^1_{\mathbf{s}}$ and compute the average, given by

$$\hat{R}_{\mathbf{s}}(\hat{\gamma}_k) = \int L\Big\{\mathbf{o}; \hat{\gamma}_k(\mathbb{P}^0_{\mathbf{s}}), \hat{\boldsymbol{\eta}}(\mathbb{P}^0_{\mathbf{s}})\Big\} \, d\mathbb{P}^1_{\mathbf{s}}(\mathbf{o}),$$

which we call the estimated risk for candidate k at the current split $\mathbf{s}$. We repeat the above process for each split, average the split-specific risk estimates to get an overall risk estimate for candidate k, defined as $\hat{R}(\hat{\gamma}_k) = \mathbb{E}_{\mathbf{S}}\{\hat{R}_{\mathbf{S}}(\hat{\gamma}_k)\}$, and finally we repeat for each candidate

$k \in \mathcal{K}$ and pick the one $\hat{k}$ that yields the smallest overall risk estimate $\hat{k} = \arg\min_{k \in \mathcal{K}} \hat{R}(\hat{\gamma}_k)$. Thus the cross-validation selector can be written as

$$\hat{k} = \underset{k \in \mathcal{K}}{\arg\min} \ \mathbb{E}_{\mathbf{S}} \int L\Big\{\mathbf{o}; \hat{\gamma}_k(\mathbb{P}^0_{\mathbf{S}}), \hat{\boldsymbol{\eta}}(\mathbb{P}^0_{\mathbf{S}})\Big\} \ d\mathbb{P}^1_{\mathbf{S}}(\mathbf{o}). \tag{4.10}$$

van der Laan and Dudoit (2003) gave conditions under which the risk $\hat{R}(\hat{\gamma}_{\hat{k}})$ of the above cross-validation selector is asymptotically equivalent to that of an oracle selector

$$\tilde{k} = \underset{k \in \mathcal{K}}{\arg\min} \ \mathbb{E}_{\mathbf{S}} \int L\Big\{\mathbf{o}; \hat{\gamma}_k(\mathbb{P}^0_{\mathbf{S}}), \overline{\boldsymbol{\eta}}\Big\} \ dP(\mathbf{o}),$$

and also derived corresponding finite-sample bounds. We refer to van der Laan and Dudoit (2003) for further details, including a precise statement of conditions and results, along with proofs.

4.5. Illustration

In this section we apply the proposed methodology to estimate the effects on infant mortality of delivery at hospitals with high- versus low-level neonatal intensive care units. Following Lorch et al. (2012) and others, we define high-level units as those that are designated as level III by the American Academy of Pediatrics, and that deliver at least 50 low birthweight infants on average per year. Level III units have high technical capacity, providing for example subspecialist teams, advanced imaging, and the ability for sustained mechanical assisted ventilation. On the other hand, level I-II units are only designed to provide basic care to lower-risk infants. The question of whether and how care at high-level units might impact infant mortality is important for numerous reasons, from both patient and policy perspectives. From the policy perspective, for example, if high-level units can reduce infant mortality compared to low-level units, particularly among high-risk infants, then regionalization policies that send high-risk infants to high-level units might be worthwhile to pursue.

To help discern the potential benefits of delivery at hospitals with high-level units, Lorch et al. (2012) collected data on all 192,078 premature births in Pennsylvania between 1995 and 2006. Available covariate information included data about the infant, such as birthweight and gestational age, as well as about the delivering mother, such as age, race, and measures of socioeconomic status and comorbidities. More detailed information about the data can be found elsewhere in Baiocchi et al. (2010) and Lorch et al. (2012). Importantly, however, the data are missing some covariate information that might be useful in explaining the process by which a mother delivers at a hospital with a high- versus low-level unit. Specifically, detailed clinical confounders such as comorbidity severity and laboratory results are not available, thus making analyses relying on 'no unmeasured confounding'-type assumptions suspect. Luckily, though, Baiocchi et al. (2010) and Lorch et al. (2012) identified a potential instrumental variable, defined as the excess travel time (in minutes) it takes a mother to get to the nearest high- versus low-level intensive care unit. This is a plausible instrument since it affects where mothers deliver (larger values mean mothers have to travel longer to get to high-level units), it plausibly does not independently affect infant mortality, and it likely is not associated with unmeasured confounders that also affect mortality (at least conditional on measured factors like socioeconomic status). Again more details about the instrument can be found in Baiocchi et al. (2010) and Lorch et al. (2012). Figure 6 shows loess fits of the unadjusted relationship between the instrument and treatment (which is strong), and between the instrument and outcome (which is less strong); the points also give some indication of the marginal distribution of the instrument.

We conducted two sets of analyses based on the methodology proposed in previous sections. First we estimated the local instrumental variable curve only conditional on the threshold value (so that $\mathbf{V} = \emptyset$), and used the proposed cross-validation approach to select among spline models. Second we estimated how effects vary with birthweight and gestational age, which are two important potential effect modifiers.

In both analyses it is first necessary to estimate the nuisance functions, which we did

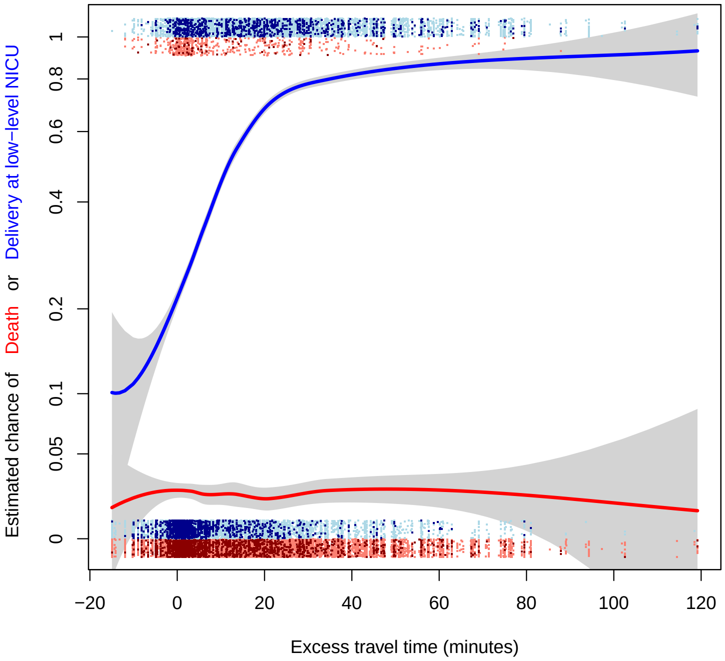

Figure 6: Relationship between instrument Z (excess travel time) and treatment A (delivery at low-level unit), and instrument and outcome Y (infant mortality).

using generalized additive models. To estimate the instrument density π, we used a model previously used by Kennedy et al. (in press), in which the density only depends on covariates through the mean and variance functions but is otherwise flexible. Specifically this model assumes $Z = \pi_1(\mathbf{X}) + \pi_2(\mathbf{X})\epsilon$, where ϵ satisfies $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$ and $\mathbb{E}(\epsilon^2 \mid \mathbf{X}) = 1$, the density f_ϵ of ϵ is unspecified but smooth, and (π_1, π_2) follow generalized additive models with identity and log links, respectively. Thus under this model the conditional density of the instrument is given by $\pi(z \mid \mathbf{x}) = f_\epsilon[\{z - \pi_1(\mathbf{x})\}/\pi_2(\mathbf{x})]$. We chose to use generalized additive models for $\pi_j(\mathbf{x})$ because of their computational speed, but other general regression methods including Super Learner (van der Laan et al., 2007) could be used instead.

In our first analysis we estimated the local instrumental variable curve $\gamma(t)$ using a density-weighted projection based on the marginal density of the instrument, so that $w(t) = \hat{p}(z = t)$ for $\hat{p}$ a usual kernel density estimator. We used natural cubic splines for $\gamma_k(t; \boldsymbol{\psi}_k)$ with degrees of freedom k selected via cross-validation with two folds. Table 4 gives the doubly robust pseudo-risk estimate $\hat{R}(\hat{\gamma}_k)$ for degrees of freedom $k = 1, 2, 3$, after scaling by 10^6 for easier comparison.

Table 4: Cross-validated model selection results.

df k	Estimated risk $\hat{R}(\hat{\gamma}_k)$
1	-12.6
2	-13.7
3	-7.3
4	1454.4

The linear model with $k = 2$ gave the smallest risk, although the risk was similar to that of the constant effect model with $k = 1$, and both models yield very similar estimates. For example, for the linear model the effect estimates range from 9.0 to 8.9 deaths per 1000 births for excess travel times ranging from 0 to 100, and at level 0.05 we cannot reject the hypothesis that the slope parameter equals zero ($p = 0.98$). Table 5 gives estimates and 95% confidence intervals (based on the bootstrap) for three estimators based on the constant effect model $\gamma(t; \boldsymbol{\psi}) = \psi$; specifically the inverse-probability-weighted estimator only relies on estimating the conditional instrument density π (i.e., it plugs in sample averages of A and Y for $\hat{\lambda}$ and $\hat{\mu}$), the regression-based estimator only relies on estimating the treatment and outcome regressions (λ, μ) (i.e., it plugs in ∞ for $\hat{\pi}$), and the doubly robust estimator is the proposed approach detailed in Section 4.4.2.

Table 5: Effect estimates and 95% confidence intervals.

Method	Est (95% CI)
Inverse-probability-weighted	-4.8 (-17.2, 7.6)
Regression-based	9.2 (6.3, 12.1)
Doubly robust	8.9 (5.4, 12.5)

Based on the proposed doubly robust estimator, the constant effect model indicates a mortality benefit (risk difference) of 8.9 fewer deaths per 1000 births due to high-level unit care (95% CI: 5.4, 12.5), among compliers who could be encouraged by travel time to go to a low-level unit. For comparison, this estimate contrasts with the unadjusted risk difference of -18.6 (-20.0, -17.2), which makes high-level units appear to be harming infants, and a doubly robust no-unmeasured-confounding-based estimate of -0.6 (-2.8, 1.6) for the average treatment effect, which does not give any evidence of benefit. The regression-based estimator was similar to the doubly robust estimator. There may be evidence that the model for the conditional density is misspecified (although this estimator is also imprecise), since the inverse-probability-weighted estimator differed from both the regression-based and doubly robust estimators.

In our second analysis (exploring effect modification by birthweight and gestational age), we projected onto a model in which effects do not vary with the latent threshold (based on the results of our first analysis) but can vary with normal versus low birthweight (2500+ grams versus <2500 grams) and early versus very early gestational age (36–37 weeks versus ≤35 weeks). Therefore in this analysis we set $\gamma(t, v; \boldsymbol{\psi}) = \sum_j \psi_j \mathbb{1}(v = j)$ where $j \in \{1, 2, 3, 4\}$ indexes the four groups. Results are given in Table 6.

Table 6: Effect estimates (95% confidence intervals) by birthweight and gestational age.

	Gestational age	
Birthweight	≤ 35 wks	36–37 wks
Low (<2500 g)	32.1 (27.7, 36.5)	1.5 (-2.3, 5.4)
Normal (2500+ g)	4.0 (0.5, 7.6)	1.9 (-1.0, 4.7)

The largest effect of high-level care was for the highest-risk infants with low birthweight and very early gestational age; in particular, for this group, care at high-level units was estimated to yield 32.1 fewer deaths per 1000 births (95% CI: 27.7, 36.5). Effects in the other three groups were relatively similar to each other, ranging from 1.5 to 4.0 fewer deaths per 1000 births. Regardless of birthweight, both groups with very early gestational age had effect estimates that were statistically significantly different from zero (at level 0.05), while estimated effects for both groups with early gestational age were not statistically significantly different from zero.

As shown in Table 7, results were similar but more pronounced when we used more extreme cutoffs for birthweight (2000+ grams versus <2000 grams) and gestational age (35–37 weeks versus ≤34 weeks).

Table 7: Effect estimates (95% confidence intervals) by birthweight and gestational age.

	Gestational age	
Birthweight	$\leq$ 34 wks	35–37 wks
<2000 g	58.5 (52.7, 64.3)	6.1 (2.2, 10.0)
2000+ g	10.1 (2.7, 17.4)	2.4 (-0.8, 5.6)

4.6. Discussion

In this paper we developed novel semiparametric theory and estimation procedures for a marginal version of the local instrumental variable curve, which represents the effect among local compliers who would be encouraged to take treatment at a given threshold value of the instrument but not below. Importantly, in contrast to available methods for estimating the fully conditional local instrumental variable curve, our methods have the following advantages: they do not require parametric assumptions (but can still yield parametric root-n rates of convergence), incorporate information about the instrument mechanism, are doubly robust (i.e., still yield consistent estimates under misspecification of either the instrument or treatment/outcome processes), and allow for estimating effect modification. We described the asymptotic properties of our methods under weak empirical process conditions, and also proposed a doubly robust cross-validation approach for model selection. Finally we used the proposed methods to study the effects of care at high-level neonatal intensive care units on infant mortality, including how such effects are modified by infants' birthweight and gestational age.

There are a number of opportunities for future work based on this research. First, it would be of interest to determine the efficient choice of the weight function $w(t, \mathbf{v})$ for the case where the working model $\gamma(t, \mathbf{v}; \boldsymbol{\psi})$ is believed to be the true model. Second, it will be very useful to develop computationally efficient software for implementing the proposed methods for general non-linear working models. The methods are computationally demanding due to the need to calculate multiple derivatives and integrals, especially in cases involving complex effect modification. A third area of future work is in the application studying the effects of high-level neonatal intensive care, where it would be useful to implement more flexible covariate adjustment (e.g., via Super Learner) and explore more complex effect modification models.

APPENDIX TO CHAPTER 2

A.1. Brief Literature Review

There are many important papers on semiparametric estimation of the effect on the treated in a simple random sampling setting. Here we give a brief description of this literature. Rubin (1977) and Heckman and Robb (1985) were two of the earlier papers to discuss effects on the treated in some detail. Later, Heckman et al. (1997) and Heckman et al. (1998) considered kernel-based matching approaches for estimation, including using an estimated propensity score. Hahn (1998) derived the efficient influence function for the effect on the treated (under a nonparametric model and a model in which the propensity score is known), and developed semiparametric efficient estimators that rely on nonparametric estimation of the propensity score and outcome regression functions. Dehejia and Wahba (1999) used stratification and matching on the propensity score with the data from LaLonde (1986), and found that the estimates were more similar to a randomized trial benchmark than those based on regression. Hirano and Imbens (2001) considered doubly robust estimation based on the efficient influence function, and used the approach to estimate effects of right heart catheterization. Hirano et al. (2003) discussed a potentially efficient estimator that only relies on estimation of the propensity score. Imbens (2004) gave a broad overview of semiparametric methods for estimating treatment effects. Abadie and Imbens (2006) derived asymptotic theory for matching estimators that use a fixed number of matches, and showed in Abadie and Imbens (2008) that the standard bootstrap is generally not valid for such estimators. More recently, Chen et al. (2008) generalized much of the above work to settings involving non-linear, possibly non-smooth, over-identified moment conditions, and considered a general context in which results can be applied to missing data and measurement error problems as well as causal inference. Kline (2011) showed that an early estimator of the effect on the treated, proposed by Oaxaca (1973) and Blinder (1973), actually fits in the doubly robust framework proposed by Robins et al. (1994) and further developed elsewhere, and gave a re-analysis of the LaLonde (1986) data. Zhang et al. (2012) considered quantile effects on the treated.

A.2. Effect Modification

In many studies interest centers not just on marginal treatment effects, but also on how effects can change with covariates. The average effect on the treated conditional on putative effect modifiers $V \subseteq L$ is given by $\gamma(v) = \mathbb{E}(Y^1 - Y^0 \mid V = v, A = 1)$, and is identified under Assumptions 1-3 in the main text by the expression

$$\gamma(v) = \int y \; p(y \mid v, a = 1) \; d\eta(y) - \int \mu(l, 0) p(\overline{v} \mid v, a = 1) \; d\nu(l).$$

Thus the conditional effect $\gamma(v)$ is identified in any study design that identifies $p(y \mid l, a)$ and $p(\overline{v} \mid v, a = 1)$, including matched cohort studies.

In the next section we derive the efficient influence function for $\gamma(v)$ under a nonparametric model with distribution Q. However, in practice V might include variables with many levels or continuous components, so that specifying a saturated model is impossible. In such cases we might want to assume a parsimonious model $\gamma(v; \psi)$ for $\gamma(v)$, which is indexed by finite-dimensional parameter $\psi \in \mathbb{R}^p$. One approach in this setting is to develop estimators under the assumption that this model is exactly correctly specified. See Lei (2011) for nice work in this setting, which can yield important efficiency advantages when the effect modification can be modeled well. An alternative approach is to only assume $\gamma(v; \psi)$ is a possibly misspecified working model and define the target parameter of interest as a projection of $\gamma(v)$ onto the working model (Neugebauer and van der Laan, 2007). We take the latter approach, defining ψ as the minimizer of the distance $\int \{\gamma(v) - \gamma(v; \psi)\}^2 p(v \mid a = 1) \; d\nu(v)$. If the working model is incorrect, this parameter is still validly defined as a projection. Further, if the working model happens to be correct, the efficient influence function for ψ derived under the working model assumption will still be valid under the assumption that the model is correct, just not necessarily efficient.

Theorem A.1 *Let $\gamma(v;\psi)$ be a working model for $\gamma(v)$ with $\psi \in \mathbb{R}^p$, so that ψ is defined as the projection $\arg\min_{\psi^* \in \mathbb{R}^p} \int \{\gamma(v) - \gamma(v;\psi^*)\}^2 p(v \mid a = 1)\ d\nu(v)$. The efficient influence function for ψ under a nonparametric model with distribution Q is then $-\mathbb{E}[\{\partial\gamma(\mu,\xi;\psi)/\partial\psi\}^{\otimes 2} \mid A = 1]^{-1}\varphi^*(\mu,\xi;\psi)$, where $\varphi^*(\mu,\xi;\psi)$ is defined as*

$$\frac{\partial\gamma(V;\psi)}{\partial\psi}\left[\frac{A}{q(a=1)}\Big\{Y - \mu(L,0) - \gamma(V;\psi)\Big\} - \frac{1-A}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\Big\{Y - \mu(L,0)\Big\}\right].$$

A.3. Proofs of identification and semiparametric theory

Here we prove results for $\gamma(v)$, since results for $\mathbb{E}(Y^1 - Y^0 \mid A = 1)$ follow by taking $V = \emptyset$. We write expectations of g under F as either $\mathbb{E}_F(g)$ or $Fg = \int g\ dF$, but use $\mathbb{E}_P = \mathbb{E}$.

Proof A.1 (Identification) *It follows that $\mathbb{E}(Y^1 \mid V = v, A = 1) = \int y\ p(y \mid v, a = 1)\ d\eta(y)$ from the consistency assumption alone. Then*

$$\begin{aligned}\mathbb{E}(Y^0 \mid V = v, A = 1) &= \int \mathbb{E}(Y^0 \mid L = l, A = 1)\ p(\overline{v} \mid v, a = 1)\ d\nu(\overline{v}) \\ &= \int \mathbb{E}(Y^0 \mid L = l, A = 0)\ p(\overline{v} \mid v, a = 1)\ d\nu(\overline{v}) = \int \mu(l,0)\ p(\overline{v} \mid v, a = 1)\ d\nu(\overline{v})\end{aligned}$$

where the first equality follows by iterated expectation, the second by ignorability, and the third by consistency. Positivity is required so as to prevent conditioning on null sets.

Proof A.2 (Theorems 2.1 and A.1) *Let $q(z;\epsilon)$ be a parametric submodel with parameter $\epsilon \in \mathbb{R}$ and $q(z;0) = q(z)$. Recalling the identifying expression of $\gamma(v)$, we write*

$$\gamma(v;\epsilon) = \int\int y\Big\{p(y \mid l, a = 1;\epsilon) - p(y \mid l, a = 0;\epsilon)\Big\}p(\overline{v} \mid v, a = 1;\epsilon)\ d\eta(y)d\nu(\overline{v}).$$

By definition the efficient influence function under Q is the unique function $\varphi(Z)$ that satisfies $\partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0} = \mathbb{E}_Q\{\varphi(Z)S_\epsilon(Z)\}$, where $S_\epsilon(Z)$ is defined as $\partial \log q(z;\epsilon)/\partial\epsilon|_{\epsilon=0}$

with

$$\frac{\partial \log q(z;\epsilon)}{\partial \epsilon}\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\Big\{ \log p(y \mid l, a;\epsilon) + \log p(\overline{v} \mid v, a;\epsilon) + \log p(v \mid a=1;\epsilon) + \log q(a;\epsilon)\Big\}\Big|_{\epsilon=0}.$$

We denote the four terms on the right as $S_y(y,l,a)$*,* $S_{\overline{v}}(\overline{v},v,a)$*,* $S_v(v)$*, and* $S_a(a)$*. Then it is straightforward to show that* $\partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0}$ *equals*

$$\mathbb{E}\bigg(Y\Big\{S_y(Y,L,1)+S_{\overline{v}}(\overline{V},V,1)\Big\}-\mathbb{E}\Big[Y\Big\{S_y(Y,L,0)+S_{\overline{v}}(\overline{V},V,1)\Big\}\;\Big|\; L, A=0\Big]\;\Big|\; V=v, A=1\bigg).$$

Denote the putative efficient influence function as

$$\varphi(Z) = \frac{I(V=v)}{p(v \mid a=1)}\left[\frac{A}{q(a=1)}\Big\{Y-\mu(L,0)-\gamma(v)\Big\} - \frac{1-A}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\Big\{Y-\mu(L,0)\Big\}\right].$$

Then one can also verify that $\mathbb{E}_Q\{\varphi(Z)S_\epsilon(Z)\} = \partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0} - h(v)$ *where* $h(v)$ *equals*

$$\begin{aligned}
&\mathbb{E}\left[\frac{q(a=0)}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\Big\{Y-\mu(L,0)\Big\}\Big\{S_{\overline{v}}(\overline{V},V,0)+S_v(V)+S_a(0)\Big\}\;\Big|\; V=v, A=0\right]\\
&\quad+\mathbb{E}\left[\Big\{\mu(L,0)+\gamma(v)\Big\}S_y(Y,L,1)\;\Big|\; V=v, A=1\right]+\mathbb{E}\left\{\gamma(v)S_{\overline{v}}(\overline{V},V,1)\;\Big|\; V=v, A=1\right\}\\
&\quad-\mathbb{E}\left[\frac{q(a=0)}{q(a=1)}\left\{\frac{\xi(L)}{1-\xi(L)}\right\}\mu(L,0)S_y(Y,L,0)\;\Big|\; V=v, A=0\right]\\
&\quad-\mathbb{E}\left[\Big\{Y-\mu(L,0)-\gamma(v)\Big\}\Big\{S_v(V)+S_a(1)\Big\}\;\Big|\; V=v, A=1\right].
\end{aligned}$$

However, the first line above is zero by iterated expectations since $\mathbb{E}(Y \mid L, A=0)-\mu(L,0) = 0$*. The second and third lines are zero by iterated expectations and standard properties of conditional score functions, in particular that* $\mathbb{E}\{S_y(Y,L,A) \mid L,A\} = \mathbb{E}\{S_{\overline{v}}(\overline{V},V,A) \mid V,A\} = 0$*. Similarly the fourth line is zero by the definition of* $\gamma(v)$*. Therefore* $\mathbb{E}_Q\{\varphi(Z)S_\epsilon(Z)\} = \partial\gamma(v;\epsilon)/\partial\epsilon|_{\epsilon=0}$ *and it follows that* $\varphi(Z)$ *is the efficient influence function.*

A.4. Proofs of double robustness and asymptotics

Proof A.3 (Double robustness) *Here we show that $\mathbb{E}_Q\{\varphi^*(\mu,\xi;\psi_0)\} = 0$ if $\tilde{\mu} = \mu$ or $\tilde{\xi} = \xi$ (not necessarily both). Note that we can write this expectation as*

$$\begin{aligned}
&\mathbb{E}_Q\bigg(\frac{\partial\gamma(V;\psi)}{\partial\psi}\bigg[\frac{A}{q(a=1)}\Big\{Y-\tilde{\mu}(L,0)-\gamma(V;\psi)\Big\}-\frac{1-A}{q(a=1)}\left\{\frac{\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{Y-\tilde{\mu}(L,0)\Big\}\bigg]\bigg)\\
&=\mathbb{E}_Q\bigg(\frac{\partial\gamma(V;\psi)}{\partial\psi}\bigg[\frac{\xi(L)}{q(a=1)}\{\mu(L,1)-\tilde{\mu}(L,0)-\gamma(V;\psi)\}\\
&\qquad\qquad-\frac{1-\xi(L)}{q(a=1)}\left\{\frac{\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\{\mu(L,0)-\tilde{\mu}(L,0)\}\bigg]\bigg)\\
&=\frac{1}{q(a=1)}\mathbb{E}_Q\bigg(\frac{\partial\gamma(V;\psi)}{\partial\psi}\bigg[\xi(L)\Big\{\mu(L,1)-\mu(L,0)-\gamma(V;\psi)\Big\}\\
&\qquad\qquad-\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\bigg]\bigg)\\
&=\int\frac{\partial\gamma(v;\psi)}{\partial\psi}\Big\{\mu(l,1)-\mu(l,0)-\gamma(v;\psi)\Big\}p(l\mid a=1)\ d\nu(l)\\
&\qquad\qquad-\frac{1}{q(a=1)}\mathbb{E}_Q\bigg[\frac{\partial\gamma(V;\psi)}{\partial\psi}\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\bigg]\\
&=\int\frac{\partial\gamma(v;\psi)}{\partial\psi}\Big\{\gamma(v)-\gamma(v;\psi)\Big\}p(v\mid a=1)\ d\nu(v)\\
&\qquad\qquad-\frac{1}{q(a=1)}\mathbb{E}_Q\bigg[\frac{\partial\gamma(V;\psi)}{\partial\psi}\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\bigg]\\
&=\frac{1}{q(a=1)}\mathbb{E}_Q\left[\frac{\partial\gamma(V;\psi)}{\partial\psi}\Big\{\mu(L,0)-\tilde{\mu}(L,0)\Big\}\left\{\frac{\xi(L)-\tilde{\xi}(L)}{1-\tilde{\xi}(L)}\right\}\right].
\end{aligned}$$

The first equality follows by definition, the second by iterated expectation, the third by adding and subtracting $\mu(L,0)$ and rearranging, the fourth since

$$\int\frac{\xi(l)}{q(a=1)}g(l)q(l)\ d\nu(l)=\int g(l)\frac{q(a=1\mid l)q(l)}{q(a=1)}\ d\nu(l)=\int g(l)q(l\mid a=1)\ d\nu(l)$$

and $q(l\mid a=1)=p(l\mid a=1)$, the fifth by iterated expectation, and the last by the fact that $\int\partial\gamma(v;\psi)/\partial\psi\{\gamma(v)-\gamma(v;\psi)\}p(v\mid a=1)\ d\nu(v)=0$ by definition when $\psi=\arg\min_{\psi^*\in\mathbb{R}^p}\int\{\gamma(v)-\gamma(v;\psi^*)\}^2p(v\mid a=1)\ d\nu(v)$.

Now the result follows since the term after the last equality reduces to zero whenever either $\tilde{\mu} = \mu$ or $\tilde{\xi} = \xi$.

Proof A.4 (Asymptotic normality) *Define $\hat{\psi}$ as the solution to $\mathbb{Q}_n \varphi(\psi, \hat{\eta}) = 0$ where $\eta = (\mu, \xi)$, let $||f||^2 = \int f^2 dQ$ denote the squared $L_2(Q)$ norm, and assume*

1. *$\hat{\psi} - \psi_0 = o_p(1)$, $||\hat{\mu} - \tilde{\mu}|| = o_p(1)$, and $||\hat{\xi} - \tilde{\xi}|| = o_p(1)$ with either $\tilde{\mu} = \mu_0$ or $\tilde{\xi} = \xi_0$.*
2. *$\varphi(\psi, \hat{\eta})$ lies in a Donsker class with probability one as $n \to \infty$.*
3. *The map $\psi \to Q\varphi(\psi, \eta)$ is differentiable at ψ_0 uniformly in η, with derivative $D_{\psi,\eta}$.*
4. *$\varphi(\psi, \eta)$ is continuous in $L_2(Q)$ at $(\psi_0, \tilde{\eta})$.*

We will show that if $\tilde{\eta} = \eta_0$ and $||\hat{\mu} - \mu_0|| \cdot ||\hat{\xi} - \xi_0|| = o_p(n^{-1/2})$ then $\hat{\psi}$ is root-n consistent, asymptotically normal, and efficient. If $\eta \in \mathbb{R}^d$, the map $\eta \to Q\varphi(\psi, \eta)$ is differentiable with nonsingular derivative $\Delta_{\psi,\eta}$, and $\hat{\eta}$ has influence function $\phi(\eta)$ so that $\hat{\eta} - \tilde{\eta} = \mathbb{Q}_n \phi(\tilde{\eta}) + o_p(n^{-1/2})$, then $\hat{\psi}$ is root-n consistent and asymptotically normal, even if $\tilde{\eta} \neq \eta_0$ (i.e., even if one of $\tilde{\mu} \neq \mu_0$ or $\tilde{\xi} \neq \xi_0$).

By Theorem 5.31 from van der Vaart (2000) (also see van der Vaart (2002)), under Assumptions 1-4 above we have

$$\hat{\psi} - \psi_0 = -D^{-1}_{\psi_0,\tilde{\eta}} \, Q\varphi(\psi_0, \hat{\eta}) - D^{-1}_{\psi_0,\tilde{\eta}} \, \mathbb{Q}_n \varphi(\psi_0, \tilde{\eta}) + o_p\Big(n^{-1/2} + ||Q\varphi(\psi_0, \hat{\eta})||\Big).$$

Further from the double robustness result on the previous page we have

$$Q\varphi(\psi_0, \hat{\eta}) = \frac{1}{q(a=1)} Q\left[\frac{\partial \gamma(V; \psi_0)}{\partial \psi} \Big\{\mu_0(L, 0) - \hat{\mu}(L, 0)\Big\} \left\{\frac{\xi_0(L) - \hat{\xi}(L)}{1 - \hat{\xi}(L)}\right\}\right].$$

First assume $\tilde{\eta} = \eta_0$ and $||\hat{\mu} - \mu_0|| \cdot ||\hat{\xi} - \xi_0|| = o_p(n^{-1/2})$. Since $Q(fg) \leq ||f|| \cdot ||g||$ by the

Cauchy-Schwarz inequality, for some constant C it follows that

$$Q\varphi(\psi_0, \hat{\eta}) \leq C||\hat{\mu} - \mu_0|| \cdot ||\hat{\xi} - \xi_0||,$$

and the right-hand side is $o_p(n^{-1/2})$ by assumption. Therefore

$$\hat{\psi} - \psi_0 = -D^{-1}_{\psi_0,\eta_0} \mathbb{Q}_n \varphi(\psi_0, \eta_0) + o_p(n^{-1/2})$$

so that $\hat{\psi}$ is root-n consistent, asymptotically normal, and efficient.

Now assume $\eta \in \mathbb{R}^d$, the map $\eta \to Q\varphi(\psi, \eta)$ is differentiable, and $\hat{\eta}$ has influence function $\phi(\eta)$. Then by the Delta method we have

$$Q\varphi(\psi_0, \hat{\eta}) = Q\varphi(\psi_0, \hat{\eta}) - Q\varphi(\psi_0, \tilde{\eta}) = \Delta_{\psi_0,\tilde{\eta}} \ \mathbb{Q}_n \phi(\tilde{\eta}) + o_p(n^{-1/2}).$$

Therefore $\hat{\psi} - \psi_0 = -D^{-1}_{\psi_0,\tilde{\eta}} \ \Delta_{\psi_0,\tilde{\eta}} \ \mathbb{Q}_n \phi(\tilde{\eta}) - D^{-1}_{\psi_0,\tilde{\eta}} \ \mathbb{Q}_n \varphi(\psi_0, \tilde{\eta}) + o_p(n^{-1/2} + O_p(n^{-1/2}))$, and this implies that

$$\hat{\psi} - \psi_0 = -D^{-1}_{\psi_0,\tilde{\eta}} \ \mathbb{Q}_n \Big\{ \Delta_{\psi_0,\tilde{\eta}} \ \phi(\tilde{\eta}) + \varphi(\psi_0, \tilde{\eta}) \Big\} + o_p(n^{-1/2}),$$

so that $\hat{\psi}$ is root-n consistent and asymptotically normal.

A.5. Proofs for Section 2.5

Proof A.5 (Efficiency bound) *Using notation from the main text, we have*

$$\begin{aligned} &var_Q \left[\frac{A}{q(a=1)} \Big\{ Y - \mu(L,0) - \psi \Big\} - \frac{1-A}{q(a=1)} \left\{ \frac{\xi(L)}{1 - \xi(L)} \right\} \Big\{ Y - \mu(L,0) \Big\} \right] \\ &\quad = \frac{1}{q(a=1)^2} \mathbb{E}_Q \Big[\xi(L)\sigma^2(L,1) + \{\mu(L,1) - \mu(L,0) - \psi\}^2 \xi(L) + \frac{\xi(L)^2}{1-\xi(L)} \sigma^2(L,0) \Big] \\ &\quad = \frac{1}{q(a=1)} \mathbb{E} \Big[\sigma^2(L,1) + \{\mu(L,1) - \mu(L,0) - \psi\}^2 + \frac{\xi(L)}{1-\xi(L)} \sigma^2(L,0) \ \Big| \ A = 1 \Big] \\ &\quad = \frac{\Omega + \Sigma_1}{q(a=1)} + \frac{1}{q(a=1)} \mathbb{E} \left[\frac{\pi(L)}{1-\pi(L)} \frac{q(a=1)}{q(a=0)} \frac{p(a=0)}{p(a=1)} \frac{p(W \mid a=0)}{p(W \mid a=1)} \sigma^2(L,0) \middle| A = 1 \right], \end{aligned}$$

and the result follows since $\Sigma_0^* = \mathbb{E}\left[\frac{\pi(L)}{1-\pi(L)}\frac{p(W|a=0)}{p(W|a=1)}\sigma^2(L,0) \mid A=1\right]$.

Proof A.6 (Condition for $B_Q < B_P$) *Using the expressions for* B_Q *and* B_P*, we have*

$$\begin{aligned} B_Q < B_P &\iff \frac{\Omega+\Sigma_1}{q(a=1)} + \frac{p(a=0)}{p(a=1)}\frac{\Sigma_0^*}{q(a=0)} < \frac{\Omega+\Sigma_1+\Sigma_0}{p(a=1)} \\ &\iff \Sigma_0^* < \frac{q(a=0)}{p(a=0)}p(a=1)\left\{\frac{\Omega+\Sigma_1+\Sigma_0}{p(a=1)} - \frac{\Omega+\Sigma_1}{q(a=1)}\right\} \\ &\iff \Sigma_0^* < \frac{q(a=0)}{p(a=0)}\left\{\Sigma_0 - \frac{p(a=1)-q(a=1)}{q(a=1)}\Big(\Omega+\Sigma_1\Big)\right\}. \end{aligned}$$

This gives the desired result.

Proof A.7 (Theorem 2.2) *If there is no matching then* $\Sigma_0^* = \Sigma_0$*, and therefore*

$$\begin{aligned} B_Q < B_P &\iff \Sigma_0 < \frac{q(a=0)}{p(a=0)}\left\{\Sigma_0 - \frac{p(a=1)-q(a=1)}{q(a=1)}\Big(\Omega+\Sigma_1\Big)\right\} \\ &\iff \Big\{p(a=1)-q(a=1)\Big\}\frac{q(a=0)}{q(a=1)} < \Big\{p(a=1)-q(a=1)\Big\}\frac{\Sigma_0}{\Omega+\Sigma_1}. \end{aligned}$$

If $p(a=1) > q(a=1)$ *then the above is equivalent to* $q(a=1) > (\Omega+\Sigma_1)/(\Omega+\Sigma_1+\Sigma_0)$*, while if* $p(a=1) < q(a=1)$ *then it is equivalent to* $q(a=1) < (\Omega+\Sigma_1)/(\Omega+\Sigma_1+\Sigma_0)$.

Proof A.8 (Theorem 2.3) *Since* $q(a=1) = 1/(k+1)$ *we can write the efficiency bound as*

$$B_Q = (k+1)(\Omega+\Sigma_1) + \left(\frac{k+1}{k}\right)\frac{p(a=0)}{p(a=1)}\Sigma_0^* = kc_1 + \frac{c_2}{k} + (c_1+c_2),$$

where $c_1 = \Omega+\Sigma_1$ *and* $c_2 = \{p(a=0)/p(a=1)\}\Sigma_0^*$*. We want to find the value of* k *that minimizes this expression. The derivative with respect to* k *is* $c_1 - (c_2/k^2)$*, which when solved for* k *yields* $k^* = (c_2/c_1)^{1/2}$*. This is guaranteed to be a minimum since the second derivative at this value is* $2c_2/(k^*)^3$*, and both* c_1 *and* c_2 *(and thus also* k^**) are necessarily positive. Therefore* $k_{opt} = (c_2/c_1)^{1/2}$.

A.6. Additional Simulation Results

In Table 3 we give additional simulation results comparing matched cohort sampling (with 1:1 and 3:1 matching) versus random sampling, using the same simulation model as described in the main text. For this simulation setting we have $p(a = 1) \approx 0.203$ and

$$\Omega = 0, \ \Sigma_1 = 1, \ \text{and} \ \Sigma_0^* \approx 0.495,$$

so that the optimal number of matched controls is approximately 1.4. Therefore 1:1 matching should be more efficient than 3:1 matching and random sampling, at least for the doubly robust estimator under correct model specification; in fact this is exactly what we see.

For our simulations, in general, the estimators applied in 1:1 matched cohort samples were more efficient than in 3:1 matched cohort samples, which were more efficient than in random samples of the same size. However this relation did not always hold when models were misspecified, or for inverse-probability-weighted estimators even under correct model specification. This is to be expected based on theory, since under model misspecification and with inefficient estimators there are generally no theoretical efficiency guarantees.

A.7. Additional Illustration Details

In the efficiency analysis given in the main text (but not the main analysis estimating the effect of hysterectomy), we assumed for simplicity that the distribution of the matching covariates was the same for the treated and controls, i.e., $p(w \mid a) = p(w)$. This simplifying assumption allowed us to focus the discussion on how the efficiency bounds compare when varying the marginal proportion treated $p(a = 1)$. However, since most studies match on variables that are thought to be strong confounders, typically we would expect $p(w \mid a = 0)$ to be far from equal to $p(w \mid a = 1)$. Therefore in practice it would often be preferable to specify different values of $p(w \mid a = 0)$, as was done for $p(a = 1)$ in the main text, and see how the bounds under Q and P change relative to each other. Also note that although the

Table 8: Percent bias, scaled empirical standard errors, and confidence interval coverage based on 500 simulated datasets

		Correct model							
	Sampling	Neither		Treatment		Outcome		Both	
n	Estimator	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov	Bias (SE)	Cov
10^2	*MCS (1:1)*								
	IPW	-20 (129)	98	6 (172)	97	-20 (129)	98	6 (172)	97
	Reg	-54 (29)	68	-54 (29)	68	0 (2.3)	95	0 (2.3)	95
	DR	-41 (36)	76	-7 (35)	94	0 (2.5)	94	0 (2.7)	92
	MCS (3:1)								
	IPW	-25 (139)	99	13 (109)	99	-25 (139)	99	13 (109)	99
	Reg	-53 (37)	68	-53 (37)	68	0 (2.5)	94	0 (2.5)	94
	DR	-37 (41)	81	-5 (30)	97	0 (2.7)	95	0 (2.7)	94
	SRS								
	IPW	-20 (197)	99	13 (171)	98	-20 (197)	99	13 (171)	98
	Reg	-97 (53)	58	-97 (53)	58	0 (2.7)	95	0 (2.7)	95
	DR	-44 (81)	81	-10 (59)	96	0 (3.2)	93	0 (3.4)	93
10^3	*MCS (1:1)*								
	IPW	-27 (83)	84	-1 (95)	96	-27 (83)	84	-1 (95)	96
	Reg	-55 (30)	0	-55 (30)	0	0 (2.2)	95	0 (2.2)	95
	DR	-41 (33)	4	-1 (30)	94	0 (2.4)	95	0 (2.5)	96
	MCS (3:1)								
	IPW	-30 (58)	63	1 (62)	98	-30 (58)	63	1 (62)	98
	Reg	-56 (37)	0	-56 (37)	0	0 (2.4)	95	0 (2.4)	95
	DR	-38 (38)	8	-1 (26)	96	0 (2.5)	95	0 (2.6)	95
	SRS								
	IPW	-21 (81)	89	1 (92)	96	-21 (81)	89	1 (92)	96
	Reg	-101 (55)	0	-101 (55)	0	0 (2.7)	95	0 (2.7)	95
	DR	-43 (48)	19	-1 (49)	95	0 (2.9)	94	0 (3.0)	94
10^4	*MCS (1:1)*								
	IPW	-25 (75)	9	0 (88)	96	-25 (75)	9	0 (88)	96
	Reg	-56 (31)	0	-56 (31)	0	0 (2.1)	96	0 (2.1)	96
	DR	-41 (34)	0	0 (30)	96	0 (2.3)	95	0 (2.3)	96
	MCS (3:1)								
	IPW	-30 (58)	0	0 (61)	95	-30 (58)	0	0 (61)	95
	Reg	-56 (36)	0	-56 (36)	0	0 (2.5)	95	0 (2.5)	95
	DR	-38 (37)	0	0 (25)	94	0 (2.6)	94	0 (2.6)	95
	SRS								
	IPW	-20 (75)	23	0 (89)	94	-20 (75)	23	0 (89)	94
	Reg	-102 (58)	0	-102 (58)	0	0 (2.6)	96	0 (2.6)	96
	DR	-43 (49)	0	0 (50)	95	0 (2.7)	97	0 (2.8)	97

SE, standard error multiplied by $n^{1/2}$; IPW, inverse-probability-weighted; Reg, regression; DR, doubly robust. *MCS (k:1)* denotes a matched cohort sampling with k:1 matching on $W \in \{0, 1\}$, and *SRS* denotes simple random sampling.

bound under P is not identifiable under a matched sampling scheme, the bound under Q is identifiable. In particular, assuming correctly specified treatment and outcome models, it can be estimated with an estimate of the variance (under Q) of the doubly robust estimator.

A.8. R Code

```
require(sandwich)


#---INPUT---
#Y: string; name of outcome in data
#Yformula: glm formula for outcome,
#note: this model is fitted to the unexposed
#(i.e. those with A=0), so the formula should not contain A
#Yfamily: glm family for outcome (only used for link function)
#A: string; name of exposure (coded as 0/1) in data
#Aformula: logistic formula for exposure
#method: string; estimation method ("ML", "IPW", "DR", or "DRwt").
#"DRwt" gives DR estimation with weighted LS outcome regression
#cluster: name of cluster id variable
#data: dataset containing all variables


#---OUTPUT---
#psi: estimate of psi
#se: standard error for the estimate of psi


matched <- function(Y,Yformula,Yfamily,A,Aformula,method,cluster,data){


  #preparation
  unexposed <- which(data[,A]==0)
  data0 <- data[unexposed,]; data0star <- data; data0star[,A] <- 0
  A <- data[,A]; Y <- data[,Y]; n <- nrow(data)
  if(missing(cluster)){ ncluster <- n } else {
    ncluster <- length(unique(data[,cluster])) }

  #fit models
  if(method=="IPW" | method=="DR" | method=="DRwt"){
    Afit <- glm(formula=Aformula,family="binomial",data=data)
    w <- exp(predict(object=Afit,newdata=data,type="link"))
    #w = omega/(1-omega)
    data0$w <- w[unexposed]; nApar <- length(Afit$coef)
    LA <- model.matrix(object=Aformula,data=data) }
  if(method=="ML" | method=="DR" | method=="DRwt"){
    if(method=="DRwt")
      Yfit <- glm(formula=Yformula,family=Yfamily,data=data0,weights=w)
    else
      Yfit <- glm(formula=Yformula,family=Yfamily,data=data0)
    mu0 <- predict(object=Yfit,newdata=data0star,type="respons")
    eta0 <- predict(object=Yfit,newdata=data0star,type="link")
    nYpar <- length(Yfit$coef)
    LY <- model.matrix(object=Yformula,data=data) }

  #calculate estimate
  if(method=="ML") psi <- sum((Y-mu0)*A)/sum(A)
  if(method=="IPW") psi <- sum(Y*(A-(1-A)*w))/sum(A)
  if(method=="DR" | method=="DRwt") psi <- sum((Y-mu0)*(A-(1-A)*w))/sum(A)

  #calculate standard error
  if(method=="ML" | method=="DR" | method=="DRwt"){
    Yres <- matrix(0,nrow=n,ncol=nYpar)
    #must include those with A==1 as well here
    Yres[A==0,] <- estfun(Yfit); g <- family(Yfit)$mu.eta
    dmu.deta <- g(eta0); deta.dbeta <- LY
    dmu.dbeta <- dmu.deta*deta.dbeta
}
```

```
if(method=="IPW" | method=="DR"  | method=="DRwt"){
  Ares <- estfun(Afit)
}
if(method=="ML"){
  psires <- A*(Y-mu0-psi); res <- cbind(psires,Yres)
  psiI <- c(sum(-A),colSums(-A*dmu.dbeta))/ncluster
  YI <- cbind(matrix(rep(0,nYpar),nrow=nYpar,ncol=1),
    -solve(vcov(object=Yfit))/ncluster)
  I <- rbind(psiI,YI)
}
if(method=="IPW"){
  psires <- A*(Y-psi)-(1-A)*w*Y; res <- cbind(psires,Ares)
  psiI <- c(sum(-A),colSums(-(1-A)*Y*w*LA))/ncluster
  AI <- cbind(matrix(rep(0,nApar),nrow=nApar,ncol=1),
    -solve(vcov(object=Afit))/ncluster)
  I <- rbind(psiI,AI)
}
if(method=="DR"  | method=="DRwt"){
  psires <- A*(Y-mu0-psi)-(1-A)*w*(Y-mu0)
  res <- cbind(psires,Yres,Ares)
  psiI <- c(sum(-A),
    colSums(((1-A)*w-A)*dmu.dbeta),
    colSums(-(1-A)*(Y-mu0)*LA*w))/ncluster
  if(method=="DR")
    YI <- cbind(matrix(rep(0,nYpar),nrow=nYpar,ncol=1),
      -solve(vcov(object=Yfit))/ncluster,
      matrix(rep(0,nYpar*nApar),nrow=nYpar,ncol=nApar))
  if(method=="DRwt")
    YI <- cbind(matrix(rep(0,nYpar),nrow=nYpar,ncol=1),
      -solve(vcov(object=Yfit))/ncluster,
      t(Yres)%*%LA/ncluster)
  AI <- cbind(matrix(rep(0,nApar),nrow=nApar,ncol=1),
      matrix(rep(0,nApar*nYpar),nrow=nApar,ncol=nYpar),
      -solve(vcov(object=Afit))/ncluster)
    I <- rbind(psiI,YI,AI)
  }
  if(!missing(cluster))
    res <- aggregate(x=res,by=list(data[,cluster]),FUN=sum)[,-1]
  J <- var(res)
  se <- sqrt((solve(I)%*%J%*%t(solve(I))/ncluster)[1,1])

  #output
  out <- list(psi=psi,se=se); return(out)

}
```

APPENDIX TO CHAPTER 3

B.1. Guide to notation

$\mathbf{Z} = (\mathbf{L}, A, Y)$ = observed data arising from distribution P with density $p(\mathbf{z}) = p(y \mid \mathbf{l}, a)p(a \mid \mathbf{l})p(\mathbf{l})$ and support $\text{supp}(\mathbf{Z}) = \mathcal{Z} = \mathcal{L} \times \mathcal{A} \times \mathcal{Y}$.

$\mathbb{P}_n = \frac{1}{n}\sum_i \delta_{\mathbf{Z}_i}$ = empirical measure so that $\mathbb{P}_n(f) = \mathbb{P}_n\{f(\mathbf{Z})\} = \frac{1}{n}\sum_i f(\mathbf{z}_i)$.

$\mathbb{P}(f) = \mathbb{P}\{f(\mathbf{Z})\} = \int_{\mathcal{Z}} f(\mathbf{z})\ dP(\mathbf{z})$ = expectation for new $\mathbf{Z}$ treating f as fixed (so $\mathbb{P}(\hat{f})$ is random if $\hat{f}$ depends on sample, in which case $\mathbb{P}(\hat{f}) \neq \mathbb{E}(\hat{f})$).

$\pi(a \mid \mathbf{l}) = p(a \mid \mathbf{l}) = \frac{\partial}{\partial a} P(A \leq a \mid \mathbf{l})$ = conditional density of treatment A.

$\hat{\pi}(a \mid \mathbf{l})$ = user-specified estimator of $\pi(a \mid \mathbf{l})$, which converges to limit $\overline{\pi}(a \mid \mathbf{l})$ that may not equal true π.

$\varpi(a) = p(a) = \frac{\partial}{\partial a} P(A \leq a) = \mathbb{E}\{\pi(a \mid \mathbf{L})\} = \int_{\mathcal{L}} \pi(a \mid \mathbf{l})\ dP(\mathbf{l})$ = density of A.

$\hat{\varpi}(a) = \mathbb{P}_n\{\hat{\pi}(a \mid \mathbf{L})\} = \int_{\mathcal{L}} \hat{\pi}(a \mid \mathbf{l})\ d\mathbb{P}_n(\mathbf{l}) = \frac{1}{n}\sum_i \hat{\pi}(a \mid \mathbf{l}_i)$ = estimator of ϖ, which converges to limit $\overline{\varpi}(a)$ that may not equal true ϖ.

$\mu(\mathbf{l}, a) = \mathbb{E}(Y \mid \mathbf{L} = \mathbf{l}, A = a) = \int_{\mathcal{Y}} y\ dP(y \mid \mathbf{l}, a)$ = conditional mean outcome.

$\hat{\mu}(\mathbf{l}, a)$ = user-specified estimator of $\mu(\mathbf{l}, a)$, which converges to limit $\overline{\mu}(\mathbf{l}, a)$ that may not equal true μ.

$\hat{m}(a) = \mathbb{P}_n\{\hat{\mu}(\mathbf{L}, a)\} = \int_{\mathcal{L}} \hat{\mu}(\mathbf{l}, a)\ d\mathbb{P}_n(\mathbf{l}) = \frac{1}{n}\sum_i \hat{\mu}(\mathbf{l}_i, a)$ = regression-based plug-in estimator of $\theta(a)$, which converges to limit $\overline{m}(a)$ that may not equal true θ.

B.2. Proof of Theorem 3.1

Let $p(\mathbf{z};\epsilon)$ be a parametric submodel with parameter $\epsilon \in \mathbb{R}$ and $p(\mathbf{z};0) = p(\mathbf{z})$, for example $p(\mathbf{z};\epsilon) = \{1+\epsilon b(\mathbf{z})\}p(\mathbf{z})$ where $\mathbb{E}\{b(\mathbf{Z})\} = 0$ with $|b(\mathbf{Z})| < B$ and $|\epsilon| \leq (1/B)$ to ensure that $p(\mathbf{z};\epsilon) \geq 0$. For notational simplicity we denote $\{\partial f(\mathbf{t};\epsilon)/\partial\epsilon\}|_{\epsilon=0}$ by $f'_\epsilon(\mathbf{t};0)$ for any general function f of ϵ and other arguments $\mathbf{t}$.

By definition the efficient influence function for ψ is the unique function $\phi(\mathbf{Z})$ that satisfies $\psi'_\epsilon(0) = \mathbb{E}\{\phi(\mathbf{Z})\ell'_\epsilon(\mathbf{Z};0)\}$, where $\psi(\epsilon)$ represents the parameter of interest as a functional on the parametric submodel and $\ell(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon) = \log p(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon)$ for any partition $(\mathbf{W},\overline{\mathbf{W}}) \subseteq \mathbf{Z}$. Therefore

$$\ell'_\epsilon(\mathbf{z};\epsilon) = \ell'_\epsilon(y \mid \mathbf{l},a;\epsilon) + \ell'_\epsilon(a \mid \mathbf{l};\epsilon) + \ell'_\epsilon(\mathbf{l};\epsilon).$$

We give two important properties of such score functions $\ell'_\epsilon(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon)$ that will be used throughout this proof. First note that since $\ell(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon)$ is a log transformation of $p(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon)$, it follows that $\ell'_\epsilon(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon) = p'_\epsilon(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon)/p(\mathbf{w} \mid \overline{\mathbf{w}};\epsilon)$ because for general functions f we have $\partial \log f(\epsilon)/\partial\epsilon = \{\partial f(\epsilon)/\partial\epsilon\}/f(\epsilon)$. Similarly, as with any score function, note that $\mathbb{E}\{\ell'_\epsilon(\mathbf{W} \mid \overline{\mathbf{W}};0) \mid \overline{\mathbf{W}}\} = 0$ since

$$\int_{\mathcal{W}} \ell'_\epsilon(\mathbf{w} \mid \overline{\mathbf{w}};0)\ dP(\mathbf{w} \mid \overline{\mathbf{w}}) = \int_{\mathcal{W}} dP'_\epsilon(\mathbf{w} \mid \overline{\mathbf{w}}) = \frac{\partial}{\partial\epsilon}\int_{\mathcal{W}} dP(\mathbf{w} \mid \overline{\mathbf{w}}) = 0.$$

Our goal in this proof is to show that $\psi'_\epsilon(0) = \mathbb{E}\{\phi(\mathbf{Z})\ell'_\epsilon(\mathbf{Z};0)\}$ for the proposed influence function $\phi(\mathbf{Z}) = \xi(\mathbf{Z};\pi,\mu) - \psi + \int_{\mathcal{A}}\{\mu(\mathbf{L},a) - \int_{\mathcal{L}} \mu(\mathbf{l},a)dP(\mathbf{l})\}\varpi(a)da$ given in the main text. First we will give an expression for $\psi'_\epsilon(0)$. By definition $\psi(\epsilon) = \int_{\mathcal{A}} \theta(a;\epsilon)\varpi(a;\epsilon)\ da$, so

$$\psi'_\epsilon(0) = \int_{\mathcal{A}} \{\theta'_\epsilon(a;0)\varpi(a) + \theta(a)\varpi'_\epsilon(a;0)\}\ da = \mathbb{E}\{\theta'_\epsilon(A;0) + \theta(A)\ell'_\epsilon(A;0)\}.$$

Also since $\theta(a;\epsilon) = \int_{\mathcal{L}} \int_{\mathcal{Y}} y \ p(y \mid \mathbf{l}, a; \epsilon) p(\mathbf{l}; \epsilon) \ d\eta(y) \ d\nu(\mathbf{l})$, we have

$$\begin{aligned}
\theta'_\epsilon(a;0) &= \int_{\mathcal{L}} \int_{\mathcal{Y}} y \Big\{ p'_\epsilon(y \mid \mathbf{l}, a; 0) p(\mathbf{l}) + p(y \mid \mathbf{l}, a) p'_\epsilon(\mathbf{l}; 0) \Big\} d\eta(y) \ d\nu(\mathbf{l}) \\
&= \int_{\mathcal{L}} \int_{\mathcal{Y}} y \Big\{ \ell'_\epsilon(y \mid \mathbf{l}, a; 0) p(y \mid \mathbf{l}, a) p(\mathbf{l}) + p(y \mid \mathbf{l}, a) \ell'_\epsilon(\mathbf{l}; 0) p(\mathbf{l}) \Big\} d\eta(y) \ d\nu(\mathbf{l}) \\
&= \mathbb{E}\Big[\mathbb{E}\{ Y \ell'_\epsilon(Y \mid \mathbf{L}, A; 0) \mid \mathbf{L}, A = a \} \Big] + \mathbb{E}\Big\{ \mu(\mathbf{L}, a) \ell'_\epsilon(\mathbf{L}; 0) \Big\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\psi'_\epsilon(0) = \int_{\mathcal{A}} \Bigg(& \mathbb{E}\Big[\mathbb{E}\{ Y \ell'_\epsilon(Y \mid \mathbf{L}, A; 0) \mid \mathbf{L}, A = a \} \Big] \\
& + \mathbb{E}\Big\{ \mu(\mathbf{L}, a) \ell'_\epsilon(\mathbf{L}; 0) \Big\} + \theta(a) \ell'_\epsilon(a; 0) \Bigg) \varpi(a) \ da.
\end{aligned}$$

Now we will consider the covariance

$$\mathbb{E}\{ \phi(\mathbf{Z}) \ell'_\epsilon(\mathbf{Z}; 0) \} = \mathbb{E}\Big[\phi(\mathbf{Z}) \Big\{ \ell'_\epsilon(Y \mid \mathbf{L}, A; 0) + \ell'_\epsilon(A, \mathbf{L}; 0) \Big\} \Big],$$

which we need to show equals the earlier expression for $\psi'_\epsilon(0)$.

Recall the proposed efficient influence function given in the main text is

$$\frac{Y - \mu(\mathbf{L}, A)}{\pi(A \mid \mathbf{L})} \varpi(A) + m(A) - \psi + \int_{\mathcal{A}} \Big\{ \mu(\mathbf{L}, a) - m(a) \Big\} \varpi(a) \ da$$

where we define

$$m(a) = \int_{\mathcal{L}} \mu(\mathbf{l}, a) \ dP(\mathbf{l})$$

as the marginalized version of the regression function μ, so that $m(a) = \theta(a)$ if μ is the true regression function.

Thus $\mathbb{E}\{ \phi(\mathbf{Z}) \ell'_\epsilon(Y \mid \mathbf{L}, A; 0) \}$ equals

$$\mathbb{E}\Bigg(\Bigg[\frac{Y - \mu(\mathbf{L}, A)}{\pi(A \mid \mathbf{L}) / \varpi(A)} + \int_{\mathcal{A}} \Big\{ \mu(\mathbf{L}, a) - \theta(a) \Big\} \varpi(a) \ da + \theta(A) - \psi \Bigg] \ell'_\epsilon(Y \mid \mathbf{L}, A; 0) \Bigg)$$

$$
\begin{aligned}
&= \mathbb{E}\left\{ \frac{Y\ell_\epsilon'(Y \mid \mathbf{L}, A; 0)}{\pi(A \mid \mathbf{L})/\varpi(A)} \right\} = \mathbb{E}\left[\frac{\mathbb{E}\{Y\ell_\epsilon'(Y \mid \mathbf{L}, A; 0) \mid \mathbf{L}, A\}}{\pi(A \mid \mathbf{L})/\varpi(A)} \right] \\
&= \int_{\mathcal{A}} \mathbb{E}\Big[\mathbb{E}\{Y\ell_\epsilon'(Y \mid \mathbf{L}, A; 0) \mid \mathbf{L}, A = a\} \Big] \varpi(a) \ da
\end{aligned}
$$

where the first equality follows since $\mathbb{E}\{\ell_\epsilon'(Y \mid \mathbf{L}, A; 0) \mid \mathbf{L}, A\} = 0$, the second by iterated expectation conditioning on $\mathbf{L}$ and A, and the third by iterated expectation conditioning on $\mathbf{L}$. Now note that $\mathbb{E}\{\phi(\mathbf{Z})\ell_\epsilon'(A, \mathbf{L}; 0)\}$ equals

$$
\begin{aligned}
&\mathbb{E}\bigg[\bigg\{ \frac{Y - \mu(\mathbf{L}, A)}{\pi(A \mid \mathbf{L})/\varpi(A)} \bigg\} \ell_\epsilon'(A, \mathbf{L}; 0) + \{\theta(A) - \psi\} \Big\{ \ell_\epsilon'(\mathbf{L} \mid A; 0) + \ell_\epsilon'(A; 0) \Big\} \\
&\qquad + \int_{\mathcal{A}} \Big\{ \mu(\mathbf{L}, a) - \theta(a) \Big\} \varpi(a) \ da \ \Big\{ \ell_\epsilon'(A \mid \mathbf{L}; 0) + \ell_\epsilon'(\mathbf{L}; 0) \Big\} \bigg] \\
&= \mathbb{E}\bigg[\theta(A)\ell_\epsilon'(A; 0) + \int_{\mathcal{A}} \mu(\mathbf{L}, a)\ell_\epsilon'(\mathbf{L}; 0)\varpi(a) \ da \bigg]
\end{aligned}
$$

since by definition $\ell_\epsilon'(A, \mathbf{L}; 0) = \ell_\epsilon'(A \mid \mathbf{L}; 0) + \ell_\epsilon'(\mathbf{L}; 0) = \ell_\epsilon'(\mathbf{L} \mid A; 0) + \ell_\epsilon'(A; 0)$, and the equality used iterated expectation conditioning on $\mathbf{L}$ and A for the first term in the first line, A for the second term in the first line, and $\mathbf{L}$ for the second line. Adding the expressions $\mathbb{E}\{\phi(\mathbf{Z})\ell_\epsilon'(Y \mid \mathbf{L}, A; 0)\}$ and $\mathbb{E}\{\phi(\mathbf{Z})\ell_\epsilon'(A, \mathbf{L}; 0)\}$ gives

$$
\int_{\mathcal{A}} \bigg(\mathbb{E}\Big[\mathbb{E}\{Y\ell_\epsilon'(Y \mid \mathbf{L}, A; 0) \mid \mathbf{L}, A = a\} + \mu(\mathbf{L}, a)\ell_\epsilon'(\mathbf{L}; 0) \Big] + \theta(a)\ell_\epsilon'(a; 0) \bigg) \varpi(a) \ da,
$$

which equals $\psi_\epsilon'(0)$. Thus ϕ is the efficient influence function.

B.3. Double robustness of efficient influence function & mapping

Here we will show that $\mathbb{E}\{\phi(\mathbf{Z}; \overline{\pi}, \overline{\mu}, \psi)\} = 0$ if either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$, where $\phi(\mathbf{Z}; \overline{\pi}, \overline{\mu}, \psi)$ is the influence function defined as in the main text as

$$
\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) - \psi + \int_{\mathcal{A}} \Big\{ \overline{\mu}(\mathbf{L}, a) - \int_{\mathcal{L}} \overline{\mu}(\mathbf{l}, a) \ dP(\mathbf{l}) \Big\} \int_{\mathcal{L}} \overline{\pi}(a \mid \mathbf{l}) \ dP(\mathbf{l}) \ da,
$$

where

$$
\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) = \frac{Y - \overline{\mu}(\mathbf{L}, A)}{\overline{\pi}(A \mid \mathbf{L})} \int_{\mathcal{L}} \overline{\pi}(A \mid \mathbf{l}) \ dP(\mathbf{l}) + \int_{\mathcal{L}} \overline{\mu}(\mathbf{l}, A) \ dP(\mathbf{l}).
$$

First note that, letting $\overline{\varpi}(a) = \mathbb{E}\{\overline{\pi}(a \mid \mathbf{L})\}$ and $\overline{m}(a) = \mathbb{E}\{\overline{\mu}(\mathbf{L}, a)\}$, we have

$$\begin{aligned}
\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} &= \mathbb{E}\left\{ \frac{Y - \overline{\mu}(\mathbf{L}, A)}{\overline{\pi}(A \mid \mathbf{L})/\overline{\varpi}(A)} + \overline{m}(A) \;\middle|\; A = a \right\} \\
&= \int_{\mathcal{L}} \frac{\mu(\mathbf{l}, a) - \overline{\mu}(\mathbf{l}, a)}{\overline{\pi}(a \mid \mathbf{l})/\overline{\varpi}(a)} \; dP(\mathbf{l} \mid a) + \overline{m}(a) \\
&= \int_{\mathcal{L}} \Big\{ \mu(\mathbf{l}, a) - \overline{\mu}(\mathbf{l}, a) \Big\} \frac{\pi(a \mid \mathbf{l})/\varpi(a)}{\overline{\pi}(a \mid \mathbf{l})/\overline{\varpi}(a)} \; dP(\mathbf{l}) + \overline{m}(a) \\
&= \theta(a) + \int_{\mathcal{L}} \Big\{ \mu(\mathbf{l}, a) - \overline{\mu}(\mathbf{l}, a) \Big\} \left\{ \frac{\pi(a \mid \mathbf{l})/\varpi(a)}{\overline{\pi}(a \mid \mathbf{l})/\overline{\varpi}(a)} - 1 \right\} \; dP(\mathbf{l})
\end{aligned}$$

where the first equality follows by iterated expectation, the second follows since $p(\mathbf{l} \mid a) = p(a \mid \mathbf{l})p(\mathbf{l})/p(a)$, and the third by rearranging. The last line shows that $\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} = \theta(a)$ as long as either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$, since in either case the remainder is zero.

Therefore if $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$ we have

$$\int_{\mathcal{A}} \mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} \varpi(a) \; da - \psi = \int_{\mathcal{A}} \theta(a) \varpi(a) \; da - \psi = 0$$

so that

$$\mathbb{E}\{\phi(\mathbf{Z}; \overline{\pi}, \overline{\mu}, \psi)\} = \mathbb{E}\left[\int_{\mathcal{A}} \Big\{ \overline{\mu}(\mathbf{L}, a) - \overline{m}(a) \Big\} \overline{\varpi}(a) \; da \right].$$

But

$$\mathbb{E} \int_{\mathcal{A}} \Big\{ \overline{\mu}(\mathbf{L}, a) - \overline{m}(a) \Big\} \overline{\varpi}(a) \; da = \int_{\mathcal{A}} \Big\{ \overline{m}(a) - \overline{m}(a) \Big\} \overline{\varpi}(a) \; da = 0$$

by definition.

Therefore $\mathbb{E}\{\phi(\mathbf{Z}; \overline{\pi}, \overline{\mu}, \psi)\} = 0$ if either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$.

B.4. TMLE version of estimator

As we note in the main text, the proposed estimator

$$\hat{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\mathbf{g}_{ha}(A)^{\mathrm{T}}\}^{-1}\mathbb{P}_n\{\mathbf{g}_{ha}K_{ha}(A)\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})\}$$

is not guaranteed to respect bounds on Y, e.g., if $Y \in [0,1]$ is binary. If some observations have very small values of the denominator quantity $\hat{\pi}(A \mid \mathbf{L})/\hat{\varpi}(A)$ then the estimator could be unstable and may take values outside the range of Y. Targeted maximum likelihood or minimum loss-based estimators (TMLEs), developed by van der Laan and Rubin (2006), help combat this problem (see discussion for example in van der Laan and Rose (2011) and elsewhere). In this section we present a TMLE that should give better finite-sample performance, for example, when there are near-violations of the positivity assumption.

Our proposed TMLE can be fit as follows. First estimate the nuisance functions $\hat{\pi}$ and $\hat{\mu}$, for example with flexible machine learning (e.g., Super Learner). Then fit a logistic regression model regressing Y on 'clever covariate' vector

$$\hat{\mathbf{c}}_{ha}(\mathbf{L}, A) = \frac{\mathbf{g}_{ha}(A)K_{ha}(A)}{\hat{\pi}(A \mid \mathbf{L})/\hat{\varpi}(A)}$$

with logit$\{\hat{\mu}(\mathbf{L}, A)\}$ included as an offset (and no intercept term). This ensures

$$\mathbb{P}_n\left\{\frac{\mathbf{g}_{ha}(A)K_{ha}(A)}{\hat{\pi}(A \mid \mathbf{L})/\hat{\varpi}(A)}\left(Y - \text{expit}\Big[\text{logit}\{\hat{\mu}(\mathbf{L}, A)\} + \hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\mathbf{c}}_{ha}(\mathbf{L}, A)\Big]\right)\right\} = 0$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)$ are the parameters in the logistic regression fit. Now define

$$\hat{\mu}^*_{ha}(\mathbf{L}, A) = \text{expit}\Big[\text{logit}\{\hat{\mu}(\mathbf{L}, A)\} + \hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\mathbf{c}}_{ha}(\mathbf{L}, A)\Big].$$

Then the proposed method proceeds as before, simply replacing predicted values $\hat{\mu}(\mathbf{L}, A)$

with $\hat{\mu}^*_{ha}(\mathbf{L}, A)$. Specifically we estimate $\theta(a)$ with

$$\hat{\theta}^*_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\mathbf{g}_{ha}(A)^{\mathrm{T}}\}^{-1}\mathbb{P}_n\{\mathbf{g}_{ha}K_{ha}(A)\hat{m}^*_{ha}(A)\},$$

$$\hat{m}^*_{ha}(t) = \mathbb{P}_n\{\hat{\mu}^*_{ha}(\mathbf{L}, t)\} = \mathbb{P}_n\Big(\text{expit}\Big[\text{logit}\{\hat{\mu}(\mathbf{L}, t)\} + \hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\mathbf{c}}_{ha}(\mathbf{L}, t)\Big]\Big).$$

The above TMLE is somewhat more complicated to fit than the estimator proposed in the main text. An alternative approach that would also respect bounds on Y would be to estimate $\theta(a)$ with $\hat{\theta}_h(a) = \text{expit}\{\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_h(a)\}$ where

$$\hat{\boldsymbol{\beta}}_h(a) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\ \mathbb{P}_n\left(K_{ha}(A)\Big[\hat{\xi}(\mathbf{Z}; \hat{\pi}, \hat{\mu}) - \text{expit}\{\mathbf{g}_{ha}(A)^{\mathrm{T}}\boldsymbol{\beta}\}\Big]^2\right).$$

Another simple option would be to use the original estimator from the main text and project onto the range of possible Y values.

B.5. Stochastic equicontinuity lemmas

In this section we discuss the concept of asymptotic or stochastic equicontinuity, and give two lemmas that play a central role in subsequent proofs.

Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$. A sequence of empirical processes $\{\mathbb{G}_n V_n(f) : f \in \mathcal{F}\}$ indexed by elements f ranging over a metric space $\mathcal{F}$ (equipped with semimetric ρ) is stochastically equicontinuous (Pollard, 1984; Andrews, 1994; van der Vaart and Wellner, 1996) if for every $\varepsilon > 0$ and $\zeta > 0$ there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} P\Bigg(\sup_{\rho(f_1,f_2)<\delta} |\mathbb{G}_n V_n(f_1) - \mathbb{G}_n V_n(f_2)| > \varepsilon\Bigg) < \zeta.$$

An important consequence of stochastic equicontinuity for our purposes is that if $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ is stochastically equicontinuous then $\rho(\hat{f}, \overline{f}) = o_p(1)$ implies that $\mathbb{G}_n\{V_n(\hat{f}) - V_n(\overline{f})\} = o_p(1)$ (Pollard, 1984; Andrews, 1994).

Before presenting relevant lemmas, we first need to introduce some notation. Let F denote an envelope function for the space $\mathcal{F}$, i.e., a function with $F(\mathbf{z}) \geq |f(\mathbf{z})|$ for every $f \in \mathcal{F}$ and $\mathbf{z} \in \mathcal{Z}$. Also let $N(\varepsilon, \mathcal{F}, ||\cdot||)$ denote the covering number, i.e., the minimal number of ε-balls (using distance $||\cdot||$) needed to cover $\mathcal{F}$, and let

$$J(\delta, \mathcal{F}, L_2) = \int_0^\delta \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))}\ d\varepsilon,$$

where $L_2(Q)$ denotes the usual L_2 semimetric under distribution Q, which for any f is $||f||_{Q,2} = (\int f^2 dQ)^{1/2}$. We call $J(\infty, \mathcal{F}, L_2)$ the uniform entropy integral.

To show that a sequence of processes $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ as defined above is stochastically equicontinuous, one can use Theorem 2.11.1 from van der Vaart and Wellner (1996). (Note that in their notation $Z_n(f) = (1/\sqrt{n}) V_n(f)$.) Specifically, Theorem 2.11.1 states that stochastic equicontinuity follows from the following two Lindeberg conditions (conditions 1 and 2), with an additional restriction on the complexity of the space $\mathcal{F}$ (condition 3):

(1) $\mathbb{E}\{||V_n||_{\mathcal{F}}^2\ I(||V_n||_{\mathcal{F}} > \varepsilon\sqrt{n})\} \to 0$ for every $\varepsilon > 0$.

(2) $\sup_{\rho(f_1,f_2)<\delta_n} \mathbb{E}[\{V_n(f_1) - V_n(f_2)\}^2] \to 0$ for every sequence $\delta_n \to 0$.

(3) $\int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))}\ d\varepsilon \xrightarrow{p} 0$ for every sequence $\delta_n \to 0$.

We will give conditions under which two particular kinds of sequences of empirical processes are stochastically equicontinuous. Specifically we consider processes $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ where

$$V_n(f) = \sqrt{h}\ g_{ha}(A) K_{ha}(A) f(\mathbf{Z}),$$
$$V_n(f) = \int f(\mathbf{L}, t) g_{ha}(t) K_{ha}(t)\ dt,$$

with $g_{ha}(t)$ and $K_{ha}(t)$ defined earlier (note V_n depends on n since $h = h_n$ does).

Lemma B.1 *Consider the sequence of processes* $\{\mathbb{G}_n V_{n,j}(\cdot) : n \geq 1\}$ *with*

$$V_{n,j}(f) = \sqrt{h}\left(\frac{A-a}{h}\right)^{j-1} \frac{1}{h} K\left(\frac{A-a}{h}\right) f(\mathbf{Z}) \ , \quad j = 1, 2,$$

where $f \in \mathcal{F}$ *with envelope* $F(\mathbf{z}) = \sup_{f \in \mathcal{F}} |f(\mathbf{z})|$. *Assume the following:*

1. *The bandwidth* $h = h_n$ *satisfies* $h \to 0$ *and* $nh^3 \to \infty$ *as* $n \to \infty$.
2. *The kernel* K *is a bounded symmetric probability density with support* $[-1, 1]$.
3. A *has compact support* $\mathcal{A}$ *and continuous density* ϖ.
4. *The envelope* F *is uniformly bounded, i.e.,* $||F||_{\mathcal{Z}} \leq f_{max} < \infty$.
5. $\mathcal{F}$ *has a finite uniform entropy integral, i.e.,* $J(\delta, \mathcal{F}, L_2) < \infty$.

Then $\{\mathbb{G}_n V_{n,j}(\cdot) : n \geq 1\}$ *is stochastically equicontinuous.*

Proof B.1 *Recall that to show stochastic equicontinuity we can check conditions (1)–(3) of Theorem 2.11.1 from van der Vaart and Wellner (1996), as given earlier.*

We will show Lindeberg condition (1) using the dominated convergence theorem, which says if $X_n \overset{p}{\to} X$ *and* $|X_n| \leq Y$ *with* $\mathbb{E}(Y) < \infty$ *then* $\mathbb{E}(X_n) \to \mathbb{E}(X)$. *First note that* $||V_{n,j}||^2_{\mathcal{F}} \ I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) = o_p(1)$ *since for any* $\delta > 0$

$$\begin{aligned} &\lim_{n\to\infty} P\Big\{||V_{n,j}||^2_{\mathcal{F}} \ I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) \geq \delta\Big\} \\ &\qquad \leq \lim_{n\to\infty} P\Big(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}\Big) \\ &\qquad = \lim_{n\to\infty} P\Big\{(A-a)^{j-1} K\left(\frac{A-a}{h}\right) F(Z) > \varepsilon\sqrt{nh^{2j-1}}\Big\} \\ &\qquad \leq \lim_{n\to\infty} P\Big\{(A-a)^{j-1} ||K||_{[-1,1]} f_{max} > \varepsilon\sqrt{nh^{2j-1}}\Big\}. \end{aligned}$$

The last line above used the kernel and envelope conditions (b) and (c). The expression in the last line tends to zero as $n \to \infty$, *since* $nh \to \infty$ *and* $nh^3 \to \infty$ *by the bandwidth* *condition (a) (note that $nh \to \infty$ is implied by the fact that $h \to 0$ and $nh^3 \to \infty$), and since A has compact support by condition (c). We also have $||V_{n,j}||^2_{\mathcal{F}} I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) \leq ||V_{n,j}||^2_{\mathcal{F}}$ since $I(\cdot)$ is the indicator function, and $\mathbb{E}\{||V_{n,j}||^2_{\mathcal{F}}\} < \infty$ since*

$$\begin{aligned}\mathbb{E}\{||V_{n,j}||^2_{\mathcal{F}}\} &= \mathbb{E}\Big[\left(\frac{A-a}{h}\right)^{2(j-1)} \frac{1}{h}K\left(\frac{A-a}{h}\right)^2 F(Z)^2\Big] \\ &\leq f^2_{max}||\varpi||_{\mathcal{A}} \int \left(\frac{t-a}{h}\right)^{2(j-1)} \frac{1}{h}K\left(\frac{A-a}{h}\right)^2 \ dt \\ &= f^2_{max}||\varpi||_{\mathcal{A}} \int u^{2(j-1)} K(u)^2 \ dt < \infty.\end{aligned}$$

The second line above follows by the distribution condition (c) and the envelope condition (d), and the last line is finite by the kernel properties assumed in condition (b). Therefore since $||V_{n,j}||^2_{\mathcal{F}} \ I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) = o_p(1)$ and $||V_{n,j}||^2_{\mathcal{F}} \ I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) \leq ||V_{n,j}||^2_{\mathcal{F}}$ with $\mathbb{E}\{||V_{n,j}||^2_{\mathcal{F}}\} < \infty$, the dominated convergence theorem implies that $\mathbb{E}\{||V_{n,j}||^2_{\mathcal{F}} \ I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n})\} \to 0$ as $n \to \infty$ and thus Lindeberg condition (1) holds.

Lindeberg condition (2) holds when $\rho(\cdot)$ is the uniform norm since

$$\begin{aligned}&\sup_{\rho(f_1,f_2)<\delta_n} \mathbb{E}[\{V_{n,j}(f_1) - V_{n,j}(f_2)\}^2] \\ &\quad = \sup_{||f_1-f_2||_{\mathcal{Z}}<\delta_n} \mathbb{E}\left[\left(\frac{A-a}{h}\right)^{2(j-1)} \frac{1}{h}K\left(\frac{A-a}{h}\right)^2 \Big\{f_1(\mathbf{Z}) - f_2(\mathbf{Z})\Big\}^2\right] \\ &\quad \leq \delta_n^2 \int \left(\frac{t-a}{h}\right)^{2(j-1)} \frac{1}{h}K\left(\frac{t-a}{h}\right)^2 \varpi(t) \ dt \\ &\quad \leq \delta_n^2 \ ||\varpi||_{\mathcal{A}} \int u^{2(j-1)} K(u)^2 \ dt \ \to \ 0 \quad , \quad \text{for any } \delta_n \to 0.\end{aligned}$$

The first equality above follows by definition, the second inequality by the fact that $||f_1 - f_2||_{\mathcal{Z}} < \delta_n$, and the third by condition (c) and a change of variables. The last line tends to zero as $\delta_n \to 0$ by the kernel properties in condition (b).

Now we consider the complexity condition (3). As described in Section 2.11.1.1 (page 209) of van der Vaart and Wellner (1996), a process $(1/\sqrt{n})V_n(f)$ is measure-like if for some *(random) measure ν_{ni} we have*

$$\frac{1}{n}\Big\{V_n(f_1) - V_n(f_2)\Big\}^2 \le \int (f_1 - f_2)^2 \, d\nu_{ni} \ , \quad \textit{for every } f_1, f_2 \in \mathcal{F}.$$

van der Vaart and Wellner (1996) show in their Lemma 2.11.6 that if $\mathcal{F}$ has a finite uniform entropy integral, then measure-like processes indexed by $\mathcal{F}$ satisfy the complexity condition (3) of Theorem 2.11.1.

Note that for our process $V_{n,j}(f)$ of interest, we have

$$\frac{1}{n}\Big\{V_{n,j}(f_1) - V_{n,j}(f_2)\Big\}^2 = \Big\{f_1(\mathbf{Z}) - f_2(\mathbf{Z})\Big\}^2 \sqrt{h}\left(\frac{A-a}{h}\right)^{j-1} \frac{1}{h}K\Big(\frac{A-a}{h}\Big).$$

Therefore the processes $V_{n,j}(f)$ are measure-like for the random measure $\nu_{ni} = \sqrt{h} g_{ha} K_{ha} \delta_{\mathbf{Z}_i}$, where $\delta_{\mathbf{Z}_i}$ denotes the Dirac measure. Hence, by Lemma 2.11.6 of van der Vaart and Wellner (1996), the fact that $\mathcal{F}$ has a finite uniform entropy integral (assumed in condition (e)) implies that complexity condition (3) is satisfied.

Therefore the sequence $\{\mathbb{G}_n V_{n,j}(\cdot) : n \ge 1\}$ is stochastically equicontinuous.

As mentioned earlier, Lemma B.1 implies that if $||\hat{f} - f||_{\mathcal{Z}} = o_p(1)$ then

$$\sqrt{nh}(\mathbb{P}_n - \mathbb{P})\left[\left(\frac{A-a}{h}\right)^{j-1} \frac{1}{h}K\Big(\frac{A-a}{h}\Big)\Big\{\hat{f}(\mathbf{Z}) - f(\mathbf{Z})\Big\}\right] = o_p(1).$$

Lemma B.2 *Consider the sequence of processes $\{\mathbb{G}_n V_{n,j}(\cdot) : n \ge 1\}$ with*

$$V_{n,j}(f) = \int f(\mathbf{L}, t)\Big(\frac{t-a}{h}\Big)^{j-1}\frac{1}{h}K\Big(\frac{t-a}{h}\Big)\, dt \ , \quad j = 1, 2,$$

where $f \in \mathcal{F}$ with envelope F as in Lemma B.1. Assume conditions (b), (d), and (e) of Lemma B.1 hold. Then $\{\mathbb{G}_n V_{n,j}(\cdot) : n \ge 1\}$ is stochastically equicontinuous.

Proof B.2 *The proof of Lemma B.2 is very similar to that of Lemma B.1. We again show Lindeberg condition (1) using the dominated convergence theorem. First note $||V_{n,j}||^2_{\mathcal{F}}\, I(||V_{n,j}||_{\mathcal{F}} >$*

$\varepsilon\sqrt{n}) = o_p(1)$ since for any $\delta > 0$

$$\begin{aligned}\lim_{n\to\infty} P\Big\{||V_{n,j}||^2_{\mathcal{F}}\ I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) \geq \delta\Big\} &\leq \lim_{n\to\infty} P\Big(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}\Big)\\ &= \lim_{n\to\infty} P\Big\{\int F(\mathbf{L},t)\{(t-a)/h\}^{j-1}K\{(t-a)/h\}/h\ dt > \varepsilon\sqrt{n}\Big\}\\ &\leq \lim_{n\to\infty} I\Big\{f_{max}\int |u|^{j-1}K(u)\ dt > \varepsilon\sqrt{n}\Big\} = 0.\end{aligned}$$

The last line above used the kernel and envelope conditions (b) and (d). We also have $||V_{n,j}||^2_{\mathcal{F}}I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n}) \leq ||V_{n,j}||^2_{\mathcal{F}}$ and $\mathbb{E}\{||V_{n,j}||^2_{\mathcal{F}}\}$ equals

$$\Big\{\int F(\mathbf{L},t)\left(\frac{t-a}{h}\right)^{j-1}\frac{1}{h}K\left(\frac{t-a}{h}\right)dt\Big\}^2 \leq f^2_{max}\Big\{\int |u|^{j-1}K(u)\ du\Big\}^2,$$

which is finite again using conditions (b) and (d). Therefore Lindeberg condition (1) holds since $\mathbb{E}\{||V_{n,j}||^2_{\mathcal{F}}I(||V_{n,j}||_{\mathcal{F}} > \varepsilon\sqrt{n})\} \to 0$ by dominated convergence.

Lindeberg condition (2) holds with the uniform norm since, by definition and using the kernel condition (b), $\sup_{\rho(f_1,f_2)<\delta_n}\mathbb{E}[\{V_n(f_1)-V_n(f_2)\}^2]$ equals

$$\begin{aligned}\sup_{||f_1-f_2||_{\mathcal{Z}}<\delta_n}\mathbb{E}\bigg(\bigg[\int\Big\{f_1(\mathbf{L},t)-f_2(\mathbf{L},t)\Big\}\left(\frac{t-a}{h}\right)^{j-1}\frac{1}{h}K\left(\frac{t-a}{h}\right)dt\bigg]^2\bigg)&\\ \leq \delta_n^2\Big\{\int |u|^{j-1}K(u)\ du\Big\}^2 \to 0 \quad , \quad \textit{for any } \delta_n \to 0.&\end{aligned}$$

As in Lemma B.1, we use that $V_{n,j}$ is measure-like to check condition (3). Here

$$\begin{aligned}\frac{1}{n}\{V_{n,j}(f_1)-V_{n,j}(f_2)\}^2 &= \frac{1}{n}\bigg[\int\{f_1(\mathbf{L},t)-f_2(\mathbf{L},t)\}\Big(\frac{t-a}{h}\Big)^{j-1}\frac{1}{h}K\Big(\frac{t-a}{h}\Big)dt\bigg]^2\\ &\leq \frac{1}{n}\int\Big\{f_1(\mathbf{L},t)-f_2(\mathbf{L},t)\Big\}^2\left|\frac{t-a}{h}\right|^{2(j-1)}\frac{1}{h}K\left(\frac{t-a}{h}\right)dt\end{aligned}$$

by Jensen's inequality. Therefore the processes $V_{n,j}(f)$ are measure-like, and the fact that $\mathcal{F}$ has a finite uniform entropy integral (assumed in condition (e)) implies that complexity condition (3) is satisfied. This concludes the proof.

B.6. Proof of Theorem 3.2

Here we let $\tilde{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\xi(\mathbf{Z};\overline{\pi},\overline{\mu})\}$ denote the infeasible estimator one would use if the nuisance functions were known, with $\hat{\mathbf{D}}_{ha} = \mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\mathbf{g}_{ha}(A)^{\mathrm{T}}\}$ as in the main text. Our proposed estimator is $\hat{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})\}$. We use the decomposition

$$\hat{\theta}_h(a) - \theta(a) = \left\{\tilde{\theta}_h(a) - \theta(a)\right\} + \left\{\hat{\theta}_h(a) - \tilde{\theta}_h(a)\right\} = \left\{\tilde{\theta}_h(a) - \theta(a)\right\} + (R_{n,1} + R_{n,2})$$

where

$$\begin{aligned} R_{n,1} &= \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}(\mathbb{P}_n - \mathbb{P})\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right] \\ R_{n,2} &= \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right]. \end{aligned}$$

Our proof is divided into three parts, one for the analysis of each of the terms above.

B.6.1. Convergence rate of $\tilde{\theta}_h(a) - \theta(a)$

Since the infeasible estimator $\tilde{\theta}_h(a)$ is a standard local linear kernel estimator with outcome $\xi(\mathbf{Z};\overline{\pi},\overline{\mu})$ and regressor A, it can be analyzed with results from the local polynomial kernel regression literature. In particular, since our Assumption 3.2 (Positivity) along with conditions (b), (c), (d) of our Theorem 3.2 imply the bandwidth condition and conditions 1(i)-1(iv) in Fan (1993), by their Theorem 1 we have $\mathbb{E}[\tilde{\theta}_h(a) - \mathbb{E}\{\xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = a\}]^2 = O(1/nh + h^4)$. Further, condition (a) of our Theorem 3.2 implies $\mathbb{E}\{\xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = a\} = \theta(a)$ by the results in Section B.3 of this Appendix. Therefore $\mathbb{E}\{\tilde{\theta}_h(a) - \theta(a)\}^2 = O(1/nh + h^4)$.

Now let $X_n = \tilde{\theta}_h(a) - \theta(a)$. The above implies that, for some $M^* > 0$, $\limsup_{n\to\infty} \mathbb{E}\{X_n^2/(1/nh + h^4)\} \leq M^*$. Therefore for any $\epsilon > 0$, if $M \geq M^*/\epsilon$,

$$\lim_{n\to\infty} P\left(\frac{X_n^2}{1/nh + h^4} \geq M\right) \leq \limsup_{n\to\infty} \frac{1}{M}\mathbb{E}\left(\frac{X_n^2}{1/nh + h^4}\right) \leq M^*/M \leq \epsilon$$

where the first equality follows by Markov's inequality, the second by the fact that $\mathbb{E}(X_n^2) = O(1/nh + h^4)$, and the third by definition of M. Since $\epsilon > 0$ was arbitrary this implies $\{\tilde{\theta}_h(a) - \theta(a)\}^2 = O_p(1/nh + h^4)$.

Now let $b_n = 1/\sqrt{nh} + h^2$ and $c_n = 1/nh + h^4$, and note that

$$P\left(\left|\frac{X_n}{b_n}\right| \geq \sqrt{M}\right) = P\left(\left|\frac{X_n^2}{c_n + 2h\sqrt{h/n}}\right| \geq M\right) \leq P\left(\left|\frac{X_n^2}{c_n}\right| \geq M\right).$$

Taking limits as $n \to \infty$ implies that

$$\left|\tilde{\theta}_h(a) - \theta(a)\right| = O_p\left(\frac{1}{\sqrt{nh}} + h^2\right).$$

B.6.2. Asymptotic negligibility of $R_{n,1}$

Now we will show that

$$R_{n,1} = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}(\mathbb{P}_n - \mathbb{P})\left[\mathbf{g}_{ha}(A)K_{ha}(A)\Big\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\Big\}\right]$$

is asymptotically negligible up to order $\sqrt{nh}$, i.e., $|R_{n,1}| = o_p(1/\sqrt{nh})$.

First we will show that $\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1} = O_p(1)$. Consider the elements of the matrix $\hat{\mathbf{D}}_{ha}$. Using the continuity of ϖ from condition (d) of Theorem 3.2 in the main text, along with properties of the kernel function from condition (c), it is straightforward to show that

$$\mathbb{E}\Big([\mathbb{P}_n\{K_{ha}(A)\} - \varpi(a)]^2\Big) = O(h) + O(1/nh).$$

Hence $\mathbb{E}([\mathbb{P}_n\{K_{ha}(A)\} - \varpi(a)]^2) = o(1)$, since $h \to 0$ and $nh \to \infty$ by condition (b), and therefore $\mathbb{P}_n\{K_{ha}(A)\} \xrightarrow{p} \varpi(a)$ by Markov's inequality. This is a standard result in classical kernel estimation problems. By the same logic we similarly have

$$\mathbb{P}_n\{K_{ha}(A)(A-a)/h\} \xrightarrow{p} 0,$$
$$\mathbb{P}_n[K_{ha}(A)\{(A-a)/h\}^2] \xrightarrow{p} \varpi(a)\int u^2K(u)\,du.$$

Therefore $\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1} \xrightarrow{p} \begin{pmatrix} 1 & 0 \end{pmatrix} \mathrm{diag}\{\varpi(a), \varpi(a)\nu_2\}^{-1} = \begin{pmatrix} \varpi(a)^{-1} & 0 \end{pmatrix}$, where $\mathrm{diag}(c_1, c_2)$ is a (2×2) diagonal matrix with elements c_1 and c_2 on the diagonal, $\nu_2 = \int u^2K(u)\,du$, and $\varpi(a) \neq 0$ because of Assumption 3.2 (Positivity). Thus we have shown that $\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1} = \begin{pmatrix} \varpi(a)^{-1} & 0 \end{pmatrix} + o_p(1) = O_p(1)$.

Now we will analyze the term

$$(\mathbb{P}_n - \mathbb{P})\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right],$$

which we will show is $o_p(1/\sqrt{nh})$. This is equivalent to showing

$$\mathbb{G}_n\left[\sqrt{h}\ \mathbf{g}_{ha}(A)K_{ha}(A)\hat{\xi}(\mathbf{Z})\right] = \mathbb{G}_n\left[\sqrt{h}\ \mathbf{g}_{ha}(A)K_{ha}(A)\overline{\xi}(\mathbf{Z})\Big\}\right] + o_p(1),$$

where we define $\hat{\xi}(\mathbf{Z}) = \hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ and $\overline{\xi}(\mathbf{Z}) = \xi(\mathbf{Z};\overline{\pi},\overline{\mu})$. Note that, as discussed in the previous section on stochastic equicontinuity, if $||\hat{\xi} - \overline{\xi}||_{\mathcal{Z}} = o_p(1)$ then the above result follows if the sequence of empirical processes $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ is stochastically equicontinuous, where we define $V_n(\xi) = \sqrt{h}\mathbf{g}_{ha}(A)K_{ha}(A)\xi(\mathbf{Z})$ with $\xi \in \Xi$ for some metric space Ξ. Thus first we will show that $||\hat{\xi} - \overline{\xi}||_{\mathcal{Z}} = \sup_{\mathbf{z}\in\mathcal{Z}} |\hat{\xi}(\mathbf{z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{z};\overline{\pi},\overline{\mu})| = o_p(1)$. Then we will check the conditions given in Lemma B.1 of the previous section, which ensure that $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ defined above is stochastically equicontinuous.

First note that after some rearranging we can write

$$\begin{aligned}\hat{\xi}(\mathbf{z}) - \xi(\mathbf{z}) &= \frac{y - \hat{\mu}(\mathbf{l}, a)}{\hat{\pi}(a \mid \mathbf{l})}\hat{\varpi}(a) + \hat{m}(a) - \frac{y - \overline{\mu}(\mathbf{l}, a)}{\overline{\pi}(a \mid \mathbf{l})}\overline{\varpi}(a) - \overline{m}(a) \\ &= \frac{y - \overline{\mu}(\mathbf{l}, a)}{\overline{\pi}(a \mid \mathbf{l})}\frac{\hat{\varpi}(a)}{\hat{\pi}(a \mid \mathbf{l})}\left\{\overline{\pi}(a \mid \mathbf{l}) - \hat{\pi}(a \mid \mathbf{l})\right\} + \frac{\hat{\varpi}(a)}{\hat{\pi}(a \mid \mathbf{l})}\left\{\overline{\mu}(\mathbf{l}, a) - \hat{\mu}(\mathbf{l}, a)\right\}\end{aligned}$$

$$+ \frac{y - \overline{\mu}(\mathbf{l}, a)}{\overline{\pi}(a \mid \mathbf{l})} \Big\{ \hat{\varpi}(a) - \overline{\varpi}(a) \Big\} + \Big\{ \hat{m}(a) - \overline{m}(a) \Big\}.$$

Therefore, letting $\hat{\xi}(\mathbf{z}) = \xi(\mathbf{z}; \hat{\pi}, \hat{\mu})$ and similarly $\overline{\xi}(\mathbf{z}) = \xi(\mathbf{z}; \overline{\pi}, \overline{\mu})$, by the uniform boundedness assumed in condition (e) and the triangle inequality we have

$$||\hat{\xi} - \overline{\xi}||_{\mathcal{Z}} = O_p\Big(||\hat{\pi} - \overline{\pi}||_{\mathcal{Z}} + ||\hat{\mu} - \overline{\mu}||_{\mathcal{Z}} + ||\hat{\varpi} - \overline{\varpi}||_{\mathcal{A}} + ||\hat{m} - \overline{m}||_{\mathcal{A}} \Big).$$

Therefore since $||\hat{\pi} - \overline{\pi}||_{\mathcal{Z}} = o_p(1)$ and $||\hat{\mu} - \overline{\mu}||_{\mathcal{Z}} = o_p(1)$ by definition, and since $O_p(o_p(1)) = o_p(1)$, the above implies

$$||\hat{\xi} - \overline{\xi}||_{\mathcal{Z}} = O_p\Big(||\hat{\varpi} - \overline{\varpi}||_{\mathcal{A}} + ||\hat{m} - \overline{m}||_{\mathcal{A}} \Big) + o_p(1).$$

Now, since by definition $\hat{\varpi}(a) = \mathbb{P}_n\{\hat{\pi}(a \mid \mathbf{L})\}$ and $\overline{\varpi}(a) = \mathbb{E}\{\overline{\pi}(a \mid \mathbf{L})\}$, we have that

$$\begin{aligned} ||\hat{\varpi} - \overline{\varpi}||_{\mathcal{A}} &= \sup_{a \in \mathcal{A}} |\hat{\varpi}(a) - \overline{\varpi}(a)| = \sup_{a \in \mathcal{A}} \Big| \mathbb{P}_n \hat{\pi}(a \mid \mathbf{L}) - \mathbb{P} \overline{\pi}(a \mid \mathbf{L}) \Big| \\ &= \sup_{a \in \mathcal{A}} \Big| \mathbb{P}_n\{\hat{\pi}(a \mid \mathbf{L}) - \overline{\pi}(a \mid \mathbf{L})\} + (\mathbb{P}_n - \mathbb{P}) \overline{\pi}(a \mid \mathbf{L}) \Big| \\ &\leq \sup_{a \in \mathcal{A}} \Big| \mathbb{P}_n\{\hat{\pi}(a \mid \mathbf{L}) - \overline{\pi}(a \mid \mathbf{L})\} \Big| + \sup_{a \in \mathcal{A}} \Big| (\mathbb{P}_n - \mathbb{P}) \overline{\pi}(a \mid \mathbf{L}) \Big| \\ &\leq ||\hat{\pi} - \overline{\pi}||_{\mathcal{Z}} + \sup_{a \in \mathcal{A}} \Big| (\mathbb{P}_n - \mathbb{P}) \overline{\pi}(a \mid \mathbf{L}) \Big|, \end{aligned}$$

where the last two lines used the triangle inequality. By definition the first term on the right hand side of the last line is $o_p(1)$, and by the uniform entropy assumption in condition (e) the second term is also $o_p(1)$ since it implies that $\overline{\pi}$ is Glivenko-Cantelli (van der Vaart, 2000; van der Vaart and Wellner, 1996). Therefore we have $||\hat{\varpi} - \overline{\varpi}||_{\mathcal{Z}} = o_p(1)$. By exactly the same logic, using definitions and condition (e) we similarly have

$$\begin{aligned} ||\hat{m} - \overline{m}||_{\mathcal{A}} &\leq \sup_{a \in \mathcal{A}} \Big| \mathbb{P}_n\{\hat{\mu}(\mathbf{L}, a) - \overline{\mu}(\mathbf{L}, a)\} \Big| + \sup_{a \in \mathcal{A}} \Big| (\mathbb{P}_n - \mathbb{P}) \overline{\mu}(\mathbf{L}, a) \Big| \\ &\leq ||\hat{\mu} - \overline{\mu}||_{\mathcal{Z}} + \sup_{a \in \mathcal{A}} \Big| (\mathbb{P}_n - \mathbb{P}) \overline{\mu}(\mathbf{L}, a) \Big| = o_p(1). \end{aligned}$$

Therefore $||\hat{\xi} - \overline{\xi}||_{\mathcal{Z}} = \sup_{\mathbf{z}\in\mathcal{Z}} |\hat{\xi}(\mathbf{z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{z};\overline{\pi},\overline{\mu})| = o_p(1)$.

Now we will show that the conditions given in Lemma B.1 hold, indicating that the sequence $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ defined above is stochastically equicontinuous. Conditions (a)–(c) of Lemma B.1 are given exactly in the statement of Theorem 3.2 and so hold immediately. For conditions (d) and (e) of Lemma B.1 we need to consider the space Ξ containing elements $\xi(\mathbf{z})$. The space Ξ can be constructed as a transformation of the spaces $(\mathcal{F}_\pi, \mathcal{F}_\mu, \mathcal{F}_\varpi, \mathcal{F}_m)$ containing the functions (π, μ, ϖ, m), along with the single identity function that takes $\mathbf{Z}$ as input and outputs Y. Specifically, we have

$$\Xi = (Y \oplus \mathcal{F}_\mu)\mathcal{F}_\pi^{-1}\mathcal{F}_\varpi \oplus \mathcal{F}_m$$

where Y is shorthand for the single function that outputs Y from $\mathbf{Z}$, and we define $\mathcal{F}_1 \oplus \mathcal{F}_2 = \{f_1 + f_2 : f_j \in \mathcal{F}_j\}$, $\mathcal{F}^{-1} = \{1/f : f \in \mathcal{F}\}$, and similarly $\mathcal{F}_1\mathcal{F}_2 = \{f_1 f_2 : f_j \in \mathcal{F}_j\}$, for arbitrary function classes $\mathcal{F}$ and $\mathcal{F}_j$ containing functions f and f_j respectively. For more discussion of such constructions of higher-level function spaces based on lower-level building blocks, we refer the reader to Pollard (1990) (Section 5), Andrews (1994) (Section 4.1), van der Vaart and Wellner (1996) (Section 2.10), and van der Vaart (2000) (Examples 19.18–19.20); for use in a related example and more discussion see van der Vaart and van der Laan (2006) (Section 5).

By condition (e) of Theorem 3.2, the classes $(\mathcal{F}_\pi, \mathcal{F}_\mu, \mathcal{F}_\varpi, \mathcal{F}_m)$ are uniformly bounded (i.e., their minimal envelopes are bounded above by some constant). Similarly the class $\mathcal{F}_\pi^{-1}$ is also uniformly bounded by the second part of condition (e). Therefore the constructed class Ξ is bounded as well, so that condition (d) of Lemma B.1 holds.

Condition (e) of Lemma B.1 can be verified by using permanence or stability properties of the uniform entropy integral (Andrews, 1994; van der Vaart and Wellner, 1996; van der Vaart and van der Laan, 2006). Specifically, by condition (e) of Theorem 3.2, the classes $(\mathcal{F}_\pi, \mathcal{F}_\mu, \mathcal{F}_\varpi, \mathcal{F}_m)$ all have a finite uniform entropy integral (as does the single function Y, or any finite set of functions). Therefore by Theorem 3 of Andrews (1994), since $\mathcal{F}_\pi^{-1}$ is appropriately bounded with finite envelope, it follows that the class Ξ also has a finite uniform entropy integral. Thus condition (e) of Lemma B.1 holds. For results similar to Theorem 3 of Andrews (1994), also see Theorem 2.10.20 of van der Vaart and Wellner (1996), and Lemma 5.1 and subsequent examples of van der Vaart and van der Laan (2006).

Thus since the conditions of Lemma B.1 hold, the sequence $\{\mathbb{G}_n V_n(\cdot) : n \geq 1\}$ with $V_n(\xi) = \sqrt{h}\mathbf{g}_{ha}(A)K_{ha}(A)\xi(\mathbf{Z})$ is stochastically equicontinuous, and since $||\hat{\xi} - \overline{\xi}||_{\mathcal{Z}} = \sup_{\mathbf{z}\in\mathcal{Z}} |\hat{\xi}(\mathbf{z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{z};\overline{\pi},\overline{\mu})| = o_p(1)$, it therefore follows that

$$(\mathbb{P}_n - \mathbb{P})\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right] = o_p(1/\sqrt{nh}).$$

Combined with the fact that $\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1} = O_p(1)$, this implies that $R_{n,1} = o_p(1/\sqrt{nh})$ and so is asymptotically negligible.

B.6.3. Convergence rate of $R_{n,2}$

In this section we will derive the convergence rate of

$$R_{n,2} = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right],$$

which will depend on how well the nuisance functions π and μ are estimated.

In the previous subsection we showed that $\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1} = O_p(1)$ using conditions (b), (c), and (d) of Theorem 3.2, along with Assumption 3.2 (Positivity). Therefore we will consider the term $\mathbb{P}[\mathbf{g}_{ha}(A)K_{ha}(A)\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\}]$, which is a vector with j^{th} element ($j = 1,2$) equal to

$$\int_{\mathcal{A}} g_{ha,j}(t)K_{ha}(t)\ \mathbb{P}\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = t\right\}\varpi(t)\ dt,$$

where $g_{ha,j}(t) = \{(t-a)/h\}^{j-1}$ as before. Note that

$$\begin{aligned}
\mathbb{P}\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = t\} &= \mathbb{P}\left\{ \frac{Y - \hat{\mu}(\mathbf{L},A)}{\hat{\pi}(A \mid \mathbf{L})/\hat{\varpi}(A)} \;\Big|\; A = t \right\} + \hat{m}(t) - \theta(t) \\
&= \mathbb{P}\left[\Big\{ \mu(\mathbf{L},t) - \hat{\mu}(\mathbf{L},t) \Big\} \left\{ \frac{\pi(t \mid \mathbf{L})/\varpi(t)}{\hat{\pi}(t \mid \mathbf{L})/\hat{\varpi}(t)} \right\} \right] + \hat{m}(t) - \theta(t) \\
&= \frac{\hat{\varpi}(t)}{\varpi(t)} \, \mathbb{P}\left[\Big\{ \mu(\mathbf{L},t) - \hat{\mu}(\mathbf{L},t) \Big\} \left\{ \frac{\pi(t \mid \mathbf{L}) - \hat{\pi}(t \mid \mathbf{L})}{\hat{\pi}(t \mid \mathbf{L})} \right\} \right] \\
&\quad + \frac{1}{\varpi(t)} \, \mathbb{P}\Big\{ \hat{\pi}(t \mid \mathbf{L}) - \pi(t \mid \mathbf{L}) \Big\} \mathbb{P}\Big\{ \mu(\mathbf{L},t) - \hat{\mu}(\mathbf{L},t) \Big\} \\
&\quad + \frac{\mathbb{P}\{\mu(\mathbf{L},t) - \hat{\mu}(\mathbf{L},t)\}}{\varpi(t)} (\mathbb{P}_n - \mathbb{P})\{\hat{\pi}(t \mid \mathbf{L})\} + (\mathbb{P}_n - \mathbb{P})\{\hat{\mu}(\mathbf{L},t)\}.
\end{aligned}$$

The first equality above follows since $\mathbb{E}\{\xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = t\} = \theta(t)$ because either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$ (as shown in Section B.3), the second by iterated expectation and the fact that $p(\mathbf{l} \mid a) = \{\pi(a \mid \mathbf{l})/\varpi(a)\}p(\mathbf{l})$, and the third by rearranging terms and the definitions $\hat{\varpi}(t) = \mathbb{P}_n\{\hat{\pi}(t \mid \mathbf{L})\}$ and $\hat{m}(t) = \mathbb{P}_n\{\hat{\mu}(\mathbf{L},t)\}$.

Therefore using the Cauchy-Schwarz inequality ($\mathbb{P}(fg) \leq ||f|| \; ||g||$), the triangle inequality, Assumption 3.2 (Positivity), and the uniform boundedness assumed in condition (e), we have

$$\begin{aligned}
&\left| \mathbb{P}\Big[g_{ha,j}(A) K_{ha}(A) \Big\{ \hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \Big\} \Big] \right| \\
&\quad = O_p\bigg(\left| \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \; ||\hat{\pi}(t \mid \mathbf{L}) - \pi(t \mid \mathbf{L})|| \; ||\hat{\mu}(\mathbf{L},t) - \mu(\mathbf{L},t)|| \; dt \right| \\
&\qquad + \left| (\mathbb{P}_n - \mathbb{P}) \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \; \hat{\pi}(t \mid \mathbf{L}) \; dt \right| \\
&\qquad + \left| (\mathbb{P}_n - \mathbb{P}) \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \; \hat{\mu}(\mathbf{L},t) \; dt \right| \bigg).
\end{aligned}$$

The last two terms above can be controlled by Lemma B.2 in this Appendix. Specifically, this lemma can be applied since its condition (b) corresponds exactly to condition (b) of Theorem 3.2, and since its conditions (d) and (e) are implied by condition (e) of Theorem 3.2. Therefore since $||\hat{\pi} - \overline{\pi}||_{\mathcal{Z}} = o_p(1)$ and $||\hat{\mu} - \overline{\mu}||_{\mathcal{Z}} = o_p(1)$ by definition, the stochastic equicontinuity result of Lemma B.2 implies that

$$(\mathbb{P}_n - \mathbb{P}) \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \Big\{ \hat{\pi}(t \mid \mathbf{L}) - \overline{\pi}(t \mid \mathbf{L}) \Big\} \; dt = o_p(1/\sqrt{n}),$$

and similarly replacing π with μ. Therefore by the central limit theorem we have

$$(\mathbb{P}_n - \mathbb{P}) \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \; \hat{\pi}(t \mid \mathbf{L}) \; dt = O_p(1/\sqrt{n}),$$

and similarly replacing π with μ. Thus the last two terms in the inequality on the previous page are asymptotically negligible up to order $\sqrt{nh}$ since

$$X_n = O_p(1/\sqrt{n}) \implies \sqrt{n} X_n = O_p(1) \implies \sqrt{nh} X_n = O_p(1) o_p(1) = o_p(1).$$

Therefore since $O_p(o_p(1/\sqrt{nh})) = o_p(1/\sqrt{nh})$, we have

$$\begin{aligned} &\left| \mathbb{P}\Big[g_{ha,j}(A) K_{ha}(A) \Big\{ \hat{\xi}(\mathbf{Z}; \hat{\pi}, \hat{\mu}) - \xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \Big\} \Big] \right| \\ &\qquad = O_p\bigg(\; \left| \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \; \phi_\pi(t) \; \phi_\mu(t) \; dt \right| \; \bigg) + o_p(1/\sqrt{nh}) \end{aligned}$$

where $\phi_\pi(t) = ||\hat{\pi}(t \mid \mathbf{L}) - \pi(t \mid \mathbf{L})||$ and $\phi_\mu(t) = ||\hat{\mu}(\mathbf{L}, t) - \mu(\mathbf{L}, t)||$.

Now let $||K||_{[-1,1]} = K_{max}$. Since $K(u) \leq K_{max} I(|u| \leq 1)$, we have

$$\begin{aligned} \int_{\mathcal{A}} g_{ha,j}(t) K_{ha}(t) \; \phi_\pi(t) \phi_\mu(t) \; dt &= \int_{\mathcal{A}} \left(\frac{t-a}{h} \right)^{j-1} \frac{1}{h} K\left(\frac{t-a}{h} \right) \phi_\pi(t) \phi_\mu(t) \; dt \\ &\leq K_{max} \left\{ \sup_{t:|t-a| \leq h} \phi_\pi(t) \right\} \left\{ \sup_{t:|t-a| \leq h} \phi_\mu(t) \right\} \int_{-1}^{1} |u|^{j-1} \; du. \end{aligned}$$

In the main text we define $r_n(a)$ and $s_n(a)$ so that $\sup_{t:|t-a| \leq h} \phi_\pi(t) = O_p(r_n(a))$ and $\sup_{t:|t-a| \leq h} \phi_\mu(t) = O_p(s_n(a))$. Therefore

$$\left| \mathbb{P}\Big[g_{ha,j}(A) K_{ha}(A) \Big\{ \hat{\xi}(\mathbf{Z}; \hat{\pi}, \hat{\mu}) - \xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \Big\} \Big] \right| = O_p\Big(r_n(a) s_n(a) \Big).$$

Combining the above with the results from previous subsections yields the desired rate from the statement of Theorem 3.2,

$$\left|\hat{\theta}_h(a) - \theta(a)\right| = O_p\left(\frac{1}{\sqrt{nh}} + h^2 + r_n(a)s_n(a)\right).$$

B.7. Proof of Theorem 3.3

As in Theorem 3.2, we again use the decomposition

$$\hat{\theta}_h(a) - \theta(a) = \left\{\tilde{\theta}_h(a) - \theta(a)\right\} + \left\{\hat{\theta}_h(a) - \tilde{\theta}_h(a)\right\} = \left\{\tilde{\theta}_h(a) - \theta(a)\right\} + (R_{n,1} + R_{n,2})$$

where $\tilde{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})\}$ is our proposed estimator, $\tilde{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\xi(\mathbf{Z};\overline{\pi},\overline{\mu})\}$ is the infeasible estimator with known nuisance functions, $\hat{\mathbf{D}}_{ha} = \mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\mathbf{g}_{ha}(A)^{\mathrm{T}}\}$, and

$$\begin{aligned} R_{n,1} &= \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}(\mathbb{P}_n - \mathbb{P})\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right] \\ R_{n,2} &= \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\mathbf{D}}_{ha}^{-1}\mathbb{P}\left[\mathbf{g}_{ha}(A)K_{ha}(A)\left\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \xi(\mathbf{Z};\overline{\pi},\overline{\mu})\right\}\right]. \end{aligned}$$

We consider each term separately, as in the proof of Theorem 3.2.

B.7.1. Asymptotic normality of $\tilde{\theta}_h(a) - \theta(a)$

After scaling, the first term $\tilde{\theta}_h(a) - \theta(a)$ above is asymptotically normal by Theorem 1 from Fan et al. (1994), since $\tilde{\theta}_h(a)$ is a standard local linear kernel estimator with outcome $\xi(\mathbf{Z};\overline{\pi},\overline{\mu})$ and regressor A, and since $\mathbb{E}\{\xi(\mathbf{Z};\overline{\pi},\overline{\mu}) \mid A = a\} = \theta(a)$ by condition (a) (i.e., either $\overline{\pi} = \pi$ or $\overline{\mu} = \mu$) as shown in Section B.3 of this Appendix. Similar proofs for the asymptotic normality of local linear kernel estimators can be found elsewhere as well (Fan, 1992; Fan et al., 1995; Masry and Fan, 1997; Li and Racine, 2007). Specifically, under conditions (b), (c), and (d) of Theorem 3.2 stated in the main text, the proof given by Fan et al. (1994) shows that, for $b_h(a) = \theta''(a)(h^2/2) \int u^2 K(u)\ du$, we have

$$\sqrt{nh}\Big\{\tilde{\theta}_h(a) - \theta(a) - b_h(a)\Big\} \rightsquigarrow N\left(0,\ \frac{\sigma^2(a) \int K(u)^2\ du}{\varpi(a)}\right)$$

where, using the fact that $\mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} = \theta(a)$ and rearranging,

$$\begin{aligned}
\sigma^2(a) &\equiv \text{var}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\} \\
&= \mathbb{E}\Big(\Big[\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) - \mathbb{E}\{\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu}) \mid A = a\}\Big]^2 \ \Big|\ A = a\Big) \\
&= \mathbb{E}\left[\left\{\frac{Y - \overline{\mu}(\mathbf{L}, A)}{\overline{\pi}(A \mid \mathbf{L})/\overline{\varpi}(A)} + \overline{m}(A) - \theta(A)\right\}^2 \ \Bigg|\ A = a\right] \\
&= \mathbb{E}\left[\left\{\frac{Y - \overline{\mu}(\mathbf{L}, A)}{\overline{\pi}(A \mid \mathbf{L})/\overline{\varpi}(A)}\right\}^2 \ \Bigg|\ A = a\right] - \{\theta(a) - \overline{m}(a)\}^2 \\
&= \mathbb{E}\left[\frac{\tau^2(\mathbf{L}, a) + \{\mu(\mathbf{L}, a) - \overline{\mu}(\mathbf{L}, a)\}^2}{\{\overline{\pi}(a \mid \mathbf{L})/\overline{\varpi}(a)\}^2/\{\pi(a \mid \mathbf{L})/\varpi(a)\}}\right] - \Big\{\theta(a) - \overline{m}(a)\Big\}^2.
\end{aligned}$$

B.7.2. Asymptotic negligibility of $R_{n,1}$

We showed $R_{n,1} = o_p(1/\sqrt{nh})$ in the earlier proof of Theorem 3.2.

B.7.3. Asymptotic negligibility of $R_{n,2}$

In the proof of Theorem 3.2 in this Appendix, we showed that $R_{n,2} = O_p(r_n(a)s_n(a))$, where $r_n(a)$ and $s_n(a)$ are the local rates of convergence for the nuisance estimators $\hat{\pi}$ and $\hat{\mu}$, as defined in the main text. By condition (f) of Theorem 3.3, we have $r_n(a)s_n(a) = o_p(1/\sqrt{nh})$ so that $R_{n,2} = O_p(o_p(1/\sqrt{nh})) = o_p(1/\sqrt{nh})$, and thus $R_{n,2}$ is asymptotically negligible up to order $\sqrt{nh}$.

Therefore the proposed estimator $\hat{\theta}_h(a)$ is asymptotically equivalent to the infeasible estimator $\tilde{\theta}_h(a)$. This yields the result from Theorem 3.3 in the main text.

B.8. Uniform consistency

In this section we sketch some conditions under which our estimator is not only consistent pointwise but also uniformly in the sense that $\sup_{a \in \mathcal{A}} |\hat{\theta}_h(a) - \theta(a)| = o_p(1)$, and give a rate of convergence. However we leave a full treatment of this result to future work, in which we will also explore weak convergence of $\hat{\theta}_h(a)$ to some Gaussian process. This will be useful for testing and inference.

We use the same decomposition as in earlier proofs of Theorems 3.2 and 3.3,

$$\hat{\theta}_h(a) - \theta(a) = \left\{ \tilde{\theta}_h(a) - \theta(a) \right\} + R_{n,1}(a) + R_{n,2}(a)$$

with $R_{n,1}(a) = R_{n,1}$ and $R_{n,2}(a) = R_{n,2}$ defined as before. From Masry (1996) and Hansen (2008) (among others), under standard smoothness/bandwidth conditions we have

$$\sup_{a \in \mathcal{A}} |\tilde{\theta}_h(a) - \theta(a)| = O_p\left(\sqrt{\frac{\log n}{nh}} + h^2 \right).$$

Further, if the empirical process $V_n(a) = \sqrt{nh/\log n} R_{n,1}(a)$ is stochastically equicontinuous, then since $\sqrt{nh}|R_{n,1}(a)| = o_p(1)$ for any $a \in \mathcal{A}$ we have

$$\sup_{a \in \mathcal{A}} |R_{n,1}(a)| = o_p\left(\sqrt{\log n / nh} \right),$$

and so is asymptotically negligible. Finally the same logic as in Section B.6.3 yields

$$\sup_{a \in \mathcal{A}} |R_{n,2}(a)| = O_p\left(\sup_{a \in \mathcal{A}} ||\hat{\pi}(a \mid \mathbf{L}) - \pi(a \mid \mathbf{L})|| \cdot ||\hat{\mu}(\mathbf{L}, a) - \mu(\mathbf{L}, a)|| \right),$$

so that for $\sup_{a \in \mathcal{A}} ||\hat{\pi}(a \mid \mathbf{L}) - \pi(a \mid \mathbf{L})|| = O_p(r_n^*)$ and similarly for $\hat{\mu}$ and s_n^* we have

$$\sup_{a \in \mathcal{A}} |\hat{\theta}_h(a) - \theta(a)| = O_p\left(\sqrt{\frac{\log n}{nh}} + h^2 + r_n^* s_n^* \right).$$

B.9. Sample R code

```
### INPUT: l is an n*p matrix, a and y are vectors of length n
###  l = matrix of covariates
###  a = vector of treatment values
###  y = vector of observed outcomes

# set up evaluation points & matrices for predictions
a.min <- min(a); a.max <- max(a)
a.vals <- seq(a.min,a.max,length.out=100)
la.new <- rbind(cbind(l,a), cbind( l[rep(1:n,length(a.vals)),],
  a=rep(a.vals,rep(n,length(a.vals))) ))
l.new <- la.new[,-dim(la.new)[2]]

# fit super learner (other methods could be used here instead)
sl.lib <- c("SL.earth","SL.gam","SL.gbm","SL.glm","SL.glmnet")
pimod <- SuperLearner(Y=a, X=l, SL.library=sl.lib, newX=l.new)
pimod.vals <- pimod$SL.predict; sq.res <- (a-pimod.vals)^2
pi2mod <- SuperLearner(Y=sq.res,X=l, SL.library=sl.lib, newX=l.new)
pi2mod.vals <- pi2mod$SL.predict
mumod <- SuperLearner(Y=y, X=cbind(l,a), SL.library=sl.lib,
  newX=la.new,family=binomial); muhat.vals <- mumod$SL.predict

# construct estimated pi/varpi and mu/m values
approx.fn <- function(x,y,z){ predict(smooth.spline(x,y),x=x2)$y }
a.std <- (la.new$a-pimod.vals)/sqrt(pi2mod.vals)
pihat.vals <- approx.fn(density(a.std[1:n])$x, density(a.std[1:n])$y,
  a.std); pihat <- pihat.vals[1:n]
pihat.mat <- matrix(pihat.vals[-(1:n)], nrow=n,ncol=length(a.vals))
varpihat <- approx.fn(a.vals, apply(pihat.mat,2,mean), a)
varpihat.mat <- matrix(rep(apply(pihat.mat,2,mean),n), byrow=T,nrow=n)
muhat <- muhat.vals[1:n]
muhat.mat <- matrix(muhat.vals[-(1:n)], nrow=n,ncol=length(a.vals))
mhat <- approx.fn(a.vals, apply(muhat.mat,2,mean), a)
mhat.mat <- matrix( rep(apply(muhat.mat,2,mean),n), byrow=T,nrow=n)
```

```
# form adjusted/pseudo outcome xi
pseudo.out <- (y-muhat)/(pihat/varpihat) + mhat

# leave-one-out cross-validation to select bandwidth
library(KernSmooth); kern <- function(x){ dnorm(x) }
w.fn <- function(bw){ w.avals <- NULL; for (a.val in a.vals){
  a.std <- (a-a.val)/bw; kern.std <- kern(a.std)/bw
  w.avals <- c(w.avals, mean(a.std^2*kern.std)*(kern(0)/bw) /
    (mean(kern.std)*mean(a.std^2*kern.std)-mean(a.std*kern.std)^2))
}; return(w.avals/n) }
hatvals <- function(bw){ approx(a.vals,w.fn(bw),xout=a)$y }
cts.eff <- function(out,bw){ approx(locpoly(a,out,bw),xout=a)$y }
# note: choice of bandwidth range depends on specific problem
h.opt <- optimize( function(h){ hats <- hatvals(h);
    mean( ((pseudo.out - cts.eff(pseudo.out,bw=h))/(1-hats))^2) },
  c(0.01,50), tol=0.01)$minimum

# estimate effect curve with optimal bandwidth
est <- approx(locpoly(a,pseudo.out,bandwidth=h.opt),xout=a.vals)$y

# estimate sandwich-style pointwise confidence band
se <- NULL; for (a.val in a.vals){
a.std <- (a-a.val)/h.opt; kern.std <- (kern(a.std)/h.opt)/h.opt
beta <- coef(lm(pseudo.out ~ a.std, weights=kern.std))
Dh <- matrix( c(mean(kern.std), mean(kern.std*a.std),
  mean(kern.std*a.std), mean(kern.std*a.std^2)), nrow=2)
kern.mat <- matrix(rep(kern((a.vals-a.val)/h)/h,n), byrow=T,nrow=n)
g2 <- matrix( rep((a.vals-a.val)/h, n), byrow=T, nrow=n)
intfn1.mat <-  kern.mat*(muhat.mat - mhat.mat)*varpihat.mat
intfn2.mat <-  g2*kern.mat*(muhat.mat - mhat.mat)*varpihat.mat
int1 <- apply(matrix(rep((a.vals[-1]-a.vals[-length(a.vals)])/2,n),
  byrow=T,nrow=n)*intfn1.mat[,-1]+intfn1.mat[,-length(a.vals)],1,sum)
int2 <- apply(matrix(rep((a.vals[-1]-a.vals[-length(a.vals)])/2,n),
  byrow=T,nrow=n)*intfn2.mat[,-1]+intfn2.mat[,-length(a.vals)],1,sum)
sigma <- cov(t(solve(Dh) %*%
  rbind( wt*(out-beta[1]-beta[2]*a.std) + int1,
  a.std*wt*(out-beta[1]-beta[2]*a.std) + int2 )))
se <- c(se, sqrt(sigma[1,1])) }
ci.ll <- est-1.96*se/sqrt(n); ci.ul <- est+1.96*se/sqrt(n)
```

APPENDIX TO CHAPTER 4

C.1. Proof of Theorem 4.1

First note that

$$\mathbb{E}(Y \mid \mathbf{X}, Z = z) = \mathbb{E}\{A^z Y^1 + (1 - A^z) Y^0 \mid \mathbf{X}, Z = z\} = \mathbb{E}\{A^z(Y^1 - Y^0) \mid \mathbf{X}\} + \mathbb{E}(Y^0 \mid \mathbf{X})$$

and similarly

$$\mathbb{E}(A \mid \mathbf{X}, Z = z) = \mathbb{E}(A^z \mid \mathbf{X}, Z = z) = \mathbb{E}(A^z \mid \mathbf{X}),$$

where the first equalities follow by Assumption 4.2 (consistency) and the second by Assumptions 4.4 (unconfoundedness of Z) and 4.5 (exclusion restriction) and rearranging. Assumption 4.3 (positivity) allows us to write conditional expectations given $\mathbf{X}$ and Z. Therefore

$$\mathbb{E}(Y \mid \mathbf{X}, Z = z + \delta) - \mathbb{E}(Y \mid \mathbf{X}, Z = z) = \mathbb{E}\{(A^{z+\delta} - A^z)(Y^1 - Y^0) \mid \mathbf{X}\}$$

and

$$\mathbb{E}(A \mid \mathbf{X}, Z = z + \delta) - \mathbb{E}(A \mid \mathbf{X}, Z = z) = \mathbb{E}(A^{z+\delta} - A^z \mid \mathbf{X}),$$

so that

$$\begin{aligned}\mathbb{E}\{\mathbb{E}(Y \mid \mathbf{X}, Z = z + \delta) - \mathbb{E}(Y \mid \mathbf{X}, Z = z) \mid \mathbf{V}\} &= \mathbb{E}\{(A^{z+\delta} - A^z)(Y^1 - Y^0) \mid \mathbf{V}\} \\ &= \mathbb{E}(Y^1 - Y^0 \mid \mathbf{V}, A^{z+\delta} > A^z) P(A^{z+\delta} > A^z \mid \mathbf{V}) \\ &= \mathbb{E}(Y^1 - Y^0 \mid \mathbf{V}, z < T \leq z + \delta) P(z < T \leq z + \delta \mid \mathbf{V})\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = z + \delta) - \mathbb{E}(A \mid \mathbf{X}, Z = z) \mid \mathbf{V}\} &= \mathbb{E}(A^{z+\delta} - A^z \mid \mathbf{V}\} \\ &= P(A^{z+\delta} > A^z \mid \mathbf{V}) = P(z < T \leq z + \delta \mid \mathbf{V}),\end{aligned}$$

where the first equalities follow by iterated expectation, the second by Assumption $1'$ (monotonicity), which implies $A^{z+\delta} - A^z = \mathbb{1}(A^{z+\delta} > A^z)$, and the third by definition of the latent threshold T.

Therefore, letting $\gamma(\mathbf{v}, t) = \mathbb{E}(Y^1 - Y^0 \mid A^1 > A^0, \mathbf{V} = \mathbf{v}, T = t)$ we have

$$\begin{aligned}
&\lim_{\delta \to 0} \delta^{-1} \mathbb{E}\{\mathbb{E}(Y \mid \mathbf{X}, Z = t + \delta) - \mathbb{E}(Y \mid \mathbf{X}, Z = t) \mid \mathbf{V}\} \\
&\quad = \lim_{\delta \to 0} \delta^{-1} \mathbb{E}(Y^1 - Y^0 \mid \mathbf{V}, t \leq T \leq t + \delta) P(t \leq T \leq t + \delta \mid \mathbf{V}) \\
&\quad = \gamma(t, \mathbf{V}) \ \lim_{\delta \to 0} \delta^{-1} \{P(T \leq t + \delta \mid \mathbf{V}) - P(T \leq t \mid \mathbf{V})\} \\
&\quad = \gamma(t, \mathbf{V}) \ p(T = t \mid \mathbf{V})
\end{aligned}$$

and similarly

$$\begin{aligned}
&\lim_{\delta \to 0} \delta^{-1} \mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = t + \delta) - \mathbb{E}(A \mid \mathbf{X}, Z = t) \mid \mathbf{V}\} = \lim_{\delta \to 0} \delta^{-1} P(t \leq T \leq t + \delta \mid \mathbf{V}) \\
&\quad = \lim_{\delta \to 0} \delta^{-1} \{P(T \leq t + \delta \mid \mathbf{V}) - P(T \leq t \mid \mathbf{V})\} \\
&\quad = p(T = t \mid \mathbf{V}),
\end{aligned}$$

where the first and third equalities follow by Assumption 4.7, i.e., by the fact that T is continuously distributed, with $\frac{\partial}{\partial t} P(T \leq t \mid \mathbf{V}) = p(T = t \mid \mathbf{V})$, and the second follows by the continuity of $\gamma(\mathbf{v}, t)$ in t.

Therefore

$$\gamma(t, \mathbf{v}) = \left. \frac{\frac{\partial}{\partial z} \mathbb{E}\{\mathbb{E}(Y \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}}{\frac{\partial}{\partial z} \mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}} \right|_{z=t}$$

since the denominator is bounded away from zero by Assumption 4.6 (instrumentation).

C.2. Proof of Theorem 4.2

In this section we use subscripts to index quantities that depend on the distribution P; a zero subscript denotes a quantity evaluated at the true distribution $P = P_0$. Thus for example $\mathbb{E}_P$ denotes expectations under P and $\mathbb{E}_0$ denotes expectations under the truth $P = P_0$; similarly $\boldsymbol{\psi}_P$ denotes the parameter $\boldsymbol{\psi} = \boldsymbol{\psi}(P)$ as a map $\boldsymbol{\psi} : P \mapsto \mathbb{R}^q$ and $\boldsymbol{\psi}_0$ denotes its true value evaluated at P_0.

Here we will show that $\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}_P, \boldsymbol{\eta}_P) = \boldsymbol{\varphi}_P(\mathbf{O})$ is the efficient influence function by showing that it is the canonical gradient of the pathwise derivative of $\boldsymbol{\psi}_P$, i.e., that $\boldsymbol{\varphi}_P$ satisfies

$$\frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon} \Big|_{\epsilon=0} = \mathbb{E}_0\{\boldsymbol{\varphi}_0(\mathbf{O}) s_0(\mathbf{O})\}$$

where $\boldsymbol{\psi}_\epsilon = \boldsymbol{\psi}(P_\epsilon)$ denotes the parameter $\boldsymbol{\psi}$ evaluated at any regular parametric submodel $\{P_\epsilon : \epsilon\}$ passing through P_0 at $\epsilon = 0$, and $s_\epsilon(\mathbf{o}_1 \mid \mathbf{o}_2) = \frac{\partial}{\partial \epsilon^*} \log dP^*_\epsilon(\mathbf{o}_1 \mid \mathbf{o}_2)|_{\epsilon^*=\epsilon}$ denotes the parametric submodel score for any partition $(\mathbf{O}_1, \mathbf{O}_2) \subseteq \mathbf{O}$.

By definition we have

$$\boldsymbol{\psi}_P = \underset{\boldsymbol{\psi} \in \mathbb{R}^q}{\arg\min} \int_{\mathcal{V}} \int_{\mathcal{T}} w(t, \mathbf{v}) \Big\{ \gamma_P(t, \mathbf{v}) - \gamma(t, \mathbf{v}; \boldsymbol{\psi}) \Big\}^2 p(T = t \mid \mathbf{v}) \; dt \; dP(\mathbf{v})$$

and thus

$$\int_{\mathcal{V}} \int_{\mathcal{T}} \frac{\partial \gamma(t, \mathbf{v}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}_P} w(t, \mathbf{v}) \Big\{ \gamma_P(t, \mathbf{v}) - \gamma(t, \mathbf{v}; \boldsymbol{\psi}_P) \Big\} p(T = t \mid \mathbf{v}) \; dt \; dP(\mathbf{v}) = 0.$$

Letting $m_P(z, \mathbf{v}) = \mathbb{E}_P\{\mathbb{E}_P(Y \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}$ and $m'_P(t, \mathbf{v}) = \frac{\partial}{\partial z} m_P(z, \mathbf{v})|_{z=t}$, and similarly $\ell_P(z, \mathbf{v}) = \mathbb{E}_P\{\mathbb{E}_P(A \mid \mathbf{X}, Z = z) \mid \mathbf{V} = \mathbf{v}\}$ and $\ell'_P(t, \mathbf{v}) = \frac{\partial}{\partial z} \ell_P(z, \mathbf{v})|_{z=t}$, then under the identifying assumptions in the main text we have

$$\gamma_P(t, \mathbf{v}) = \frac{m'_P(t, \mathbf{v})}{\ell'_P(t, \mathbf{v})} \quad \text{and} \quad p(t \mid \mathbf{v}) = \ell'_P(t, \mathbf{v}).$$

Therefore the restriction above is equivalent to

$$0 = \int_{\mathcal{V}} \int_{\mathcal{T}} \frac{\partial \gamma(t, \mathbf{v}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}_P} w(t, \mathbf{v}) \Big\{ m'_P(t, \mathbf{v}) - \gamma(t, \mathbf{v}; \boldsymbol{\psi}_P) \ell'_P(t, \mathbf{v}) \Big\} \; dt \; dP(\mathbf{v})$$

$$= \int_{\mathcal{V}} \int_{\mathcal{T}} \Big\{ \mathbf{g}_2(t, \mathbf{v}; \boldsymbol{\psi}_P)\ m'_P(t, \mathbf{v}) - \mathbf{g}_1(t, \mathbf{v}; \boldsymbol{\psi}_P)\ \ell'_P(t, \mathbf{v}) \Big\}\ dt\ dP(\mathbf{v})$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are q-vectors (with known functional form not depending on P) defined as

$$\mathbf{g}_1(t, \mathbf{v}; \boldsymbol{\psi}) = \mathbf{g}_2(t, \mathbf{v}; \boldsymbol{\psi})\gamma(t, \mathbf{v}; \boldsymbol{\psi}) \ \text{ and } \ \mathbf{g}_2(t, \mathbf{v}; \boldsymbol{\psi}) = \frac{\partial \gamma(t, \mathbf{v}; \boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^*} \Big|_{\boldsymbol{\psi}^* = \boldsymbol{\psi}} w(t, \mathbf{v}).$$

And since the weight satisfies $w(t, \mathbf{v}) = 0$ for $t \notin \text{int}(\mathcal{T})$, integration by parts gives

$$\int_{\mathcal{V}} \int_{\mathcal{T}} \Big\{ \mathbf{g}'_1(t, \mathbf{v}; \boldsymbol{\psi}_P)\ \ell_P(t, \mathbf{v}) - \mathbf{g}'_2(t, \mathbf{v}; \boldsymbol{\psi}_P)\ m_P(t, \mathbf{v}) \Big\}\ dt\ dP(\mathbf{v}) = 0,$$

where $\mathbf{g}'_j(t, \mathbf{v}; \boldsymbol{\psi}) = \partial \mathbf{g}_j(z, \mathbf{v}; \boldsymbol{\psi}) / \partial z|_{z=t}$.

Evaluating the above at $P = P_\epsilon$ gives

$$\int_{\mathcal{V}} \int_{\mathcal{T}} \Big\{ \mathbf{g}'_1(t, \mathbf{v}; \boldsymbol{\psi}_\epsilon)\ell_\epsilon(t, \mathbf{v}) - \mathbf{g}'_2(t, \mathbf{v}; \boldsymbol{\psi}_\epsilon)m_\epsilon(t, \mathbf{v}) \Big\}\ dt\ dP_\epsilon(\mathbf{v}) = 0,$$

and differentiating with respect to ϵ and evaluating at the truth $\epsilon = 0$ (using the chain rule) gives

$$\begin{aligned} 0 = \int_{\mathcal{V}} \int_{\mathcal{T}} \Bigg\{ & \frac{\partial \mathbf{g}'_1(t, \mathbf{v}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi} = \boldsymbol{\psi}_0} \frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon} \Big|_{\epsilon=0} \ell_0(t, \mathbf{v}) + \mathbf{g}'_1(t, \mathbf{v}; \boldsymbol{\psi}_0) \frac{\partial \ell_\epsilon(t, \mathbf{v})}{\partial \epsilon} \Big|_{\epsilon=0} \\ & - \frac{\partial \mathbf{g}'_2(t, \mathbf{v}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \Big|_{\boldsymbol{\psi} = \boldsymbol{\psi}_0} \frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon} \Big|_{\epsilon=0} m_0(t, \mathbf{v}) - \mathbf{g}'_2(t, \mathbf{v}; \boldsymbol{\psi}_0) \frac{\partial m_\epsilon(t, \mathbf{v})}{\partial \epsilon} \Big|_{\epsilon=0} \Bigg\} dt\ dP_0(\mathbf{v}) \\ + \int_{\mathcal{V}} \int_{\mathcal{T}} & \Big\{ \mathbf{g}'_1(t, \mathbf{v}; \boldsymbol{\psi}_0)\ell_0(t, \mathbf{v}) - \mathbf{g}'_2(t, \mathbf{v}; \boldsymbol{\psi}_0)m_0(t, \mathbf{v}) \Big\} s_0(\mathbf{v})\ dt\ dP_0(\mathbf{v}). \end{aligned}$$

Rearranging, this implies that

$$\begin{aligned} \frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon} \Big|_{\epsilon=0} = \mathbf{C}_0^{-1} \int_{\mathcal{V}} \int_{\mathcal{T}} \Bigg[& \mathbf{g}'_1(t, \mathbf{v}; \boldsymbol{\psi}_0) \left\{ \frac{\partial \ell_\epsilon(t, \mathbf{v})}{\partial \epsilon} \Big|_{\epsilon=0} + \ell_0(t, \mathbf{v}) s_0(\mathbf{v}) \right\} \\ & - \mathbf{g}'_2(t, \mathbf{v}; \boldsymbol{\psi}_0) \left\{ \frac{\partial m_\epsilon(t, \mathbf{v})}{\partial \epsilon} \Big|_{\epsilon=0} + m_0(t, \mathbf{v}) s_0(\mathbf{v}) \right\} \Bigg] dt\ dP_0(\mathbf{v}) \end{aligned}$$

with

$$\mathbf{C}_P = -\int_{\mathcal{V}}\int_{\mathcal{T}}\left\{\frac{\partial \mathbf{g}_1'(t,\mathbf{v};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}_P} \ell_P(t,\mathbf{v}) - \frac{\partial \mathbf{g}_2'(t,\mathbf{v};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}_P} m_P(t,\mathbf{v})\right\} dt\; dP(\mathbf{v}),$$

and

$$\begin{aligned}\frac{\partial \ell_\epsilon(z,\mathbf{v})}{\partial \epsilon}\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\mathbb{E}_\epsilon\{\mathbb{E}_\epsilon(A \mid \mathbf{X}, Z=z) \mid \mathbf{V}=\mathbf{v}\}|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\int_{\mathcal{W}}\sum_{a\in\{0,1\}} a\; p_\epsilon(a \mid \mathbf{x}, z)\; dP_\epsilon(\mathbf{w} \mid \mathbf{v})\Big|_{\epsilon=0} \\ &= \int_{\mathcal{W}}\sum_{a\in\{0,1\}} a\left\{s_0(a \mid \mathbf{x}, z) + s_0(\mathbf{w} \mid \mathbf{v})\right\} p_0(a \mid \mathbf{x}, z)\; dP_0(\mathbf{w} \mid \mathbf{v}) \\ &= \mathbb{E}_0\Big(\mathbb{E}_0\Big[A\{s_0(A \mid \mathbf{X}, Z) + s_0(\mathbf{W} \mid \mathbf{V})\}\;\Big|\; \mathbf{X}, Z=z\Big]\;\Big|\; \mathbf{V}=\mathbf{v}\Big),\end{aligned}$$

and by the same logic

$$\frac{\partial m_\epsilon(z,\mathbf{v})}{\partial \epsilon}\Big|_{\epsilon=0} = \mathbb{E}_0\Big(\mathbb{E}_0\Big[Y\{s_0(Y \mid \mathbf{X}, Z) + s_0(\mathbf{W} \mid \mathbf{V})\}\;\Big|\; \mathbf{X}, Z=z\Big]\;\Big|\; \mathbf{V}=\mathbf{v}\Big).$$

Now we turn to $\mathbb{E}_0\{\boldsymbol{\varphi}_0(\mathbf{O})s_0(\mathbf{O})\}$. The putative efficient influence function $\boldsymbol{\varphi}_P$ from the main text is given by

$$\begin{aligned}\boldsymbol{\varphi}_P(\mathbf{O}) = \mathbf{C}_P^{-1}\Bigg[&\mathbf{g}_1'(Z,\mathbf{V};\boldsymbol{\psi}_P)\left\{\frac{A - \mathbb{E}_P(A \mid \mathbf{X}, Z)}{p(Z \mid \mathbf{X})}\right\} - \mathbf{g}_2'(Z,\mathbf{V};\boldsymbol{\psi}_P)\left\{\frac{Y - \mathbb{E}_P(Y \mid \mathbf{X}, Z)}{p(Z \mid \mathbf{X})}\right\} \\ &+ \int_{\mathcal{T}}\left\{\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi}_P)\mathbb{E}_P(A \mid \mathbf{X}, Z=t) - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi}_P)\mathbb{E}_P(Y \mid \mathbf{X}, Z=t)\right\} dt\Bigg],\end{aligned}$$

and $s_0(\mathbf{O})$ is the parametric submodel score, which can be decomposed as

$$s_0(\mathbf{O}) = s_0(Y, A \mid \mathbf{X}, Z) + s_0(Z \mid \mathbf{X}) + s_0(\mathbf{W} \mid \mathbf{V}) + s_0(\mathbf{V}).$$

Therefore

$$\begin{aligned}&\mathbf{C}_0\mathbb{E}_0\Big[\boldsymbol{\varphi}_0(\mathbf{O})\{s_0(Y, A \mid \mathbf{X}, Z) + s_0(Z \mid \mathbf{X}) + s_0(\mathbf{W} \mid \mathbf{V}) + s_0(\mathbf{V})\}\Big] \\ &\quad = \mathbb{E}_0\Bigg[\mathbf{g}_1'(Z,\mathbf{V};\boldsymbol{\psi}_0)\left\{\frac{A s_0(A \mid \mathbf{X}, Z)}{p_0(Z \mid \mathbf{X})}\right\} - \mathbf{g}_2'(Z,\mathbf{V};\boldsymbol{\psi}_0)\left\{\frac{Y s_0(Y \mid \mathbf{X}, Z)}{p_0(Z \mid \mathbf{X})}\right\}\end{aligned}$$

$$+ \int_{\mathcal{T}} \Big\{ \mathbf{g}_1'(t, \mathbf{V}; \boldsymbol{\psi}_0)\mathbb{E}_0(A \mid \mathbf{X}, Z = t) - \mathbf{g}_2'(t, \mathbf{V}; \boldsymbol{\psi}_0)\mathbb{E}_0(Y \mid \mathbf{X}, Z = t) \Big\} \, dt$$
$$\times \Big\{ s_0(\mathbf{W} \mid \mathbf{V}) + s_0(\mathbf{V}) \Big\} \Big]$$
$$= \int_{\mathcal{V}} \int_{\mathcal{T}} \bigg(\mathbf{g}_1'(t, \mathbf{v}; \boldsymbol{\psi}_0)\mathbb{E}_0\Big(\mathbb{E}_0\Big[A\{s_0(A \mid \mathbf{X}, Z) + s_0(\mathbf{W} \mid \mathbf{V})\} \,\Big|\, \mathbf{X}, Z = t \Big] \,\Big|\, \mathbf{V} = \mathbf{v} \Big)$$
$$- \mathbf{g}_2'(t, \mathbf{v}; \boldsymbol{\psi}_0)\mathbb{E}_0\Big(\mathbb{E}_0\Big[Y\{s_0(Y \mid \mathbf{X}, Z) + s_0(\mathbf{W} \mid \mathbf{V})\} \,\Big|\, \mathbf{X}, Z = t \Big] \,\Big|\, \mathbf{V} = \mathbf{v} \Big)$$
$$+ \Big\{ \mathbf{g}_1'(t, \mathbf{v}; \boldsymbol{\psi}_0)\mathbb{E}_0(A \mid \mathbf{X}, Z = t) - \mathbf{g}_2'(t, \mathbf{v}; \boldsymbol{\psi}_0)\mathbb{E}_0(Y \mid \mathbf{X}, Z = t) \Big\} s_0(\mathbf{v}) \bigg) \, dt \, dP_0(\mathbf{v})$$
$$= \int_{\mathcal{V}} \int_{\mathcal{T}} \bigg[\mathbf{g}_1'(t, \mathbf{v}; \boldsymbol{\psi}_0) \bigg\{ \frac{\partial \ell_\epsilon(z, \mathbf{v})}{\partial \epsilon} \Big|_{\epsilon=0} + \ell_0(t, \mathbf{v}) s_0(\mathbf{v}) \bigg\}$$
$$- \mathbf{g}_2'(t, \mathbf{v}; \boldsymbol{\psi}_0) \bigg\{ \frac{\partial m_\epsilon(t, \mathbf{v})}{\partial \epsilon} \Big|_{\epsilon=0} + m_0(t, \mathbf{v}) s_0(\mathbf{v}) \bigg\} \bigg] dt \, dP_0(\mathbf{v}) = \mathbf{C}_0 \frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon} \Big|_{\epsilon=0}$$

where the first equality follows by iterated expectation and the fact that $\mathbb{E}_0\{s_0(\mathbf{O}_1 \mid \mathbf{O}_2) \mid \mathbf{O}_2\} = 0$ for any $(\mathbf{O}_1, \mathbf{O}_2) \subseteq \mathbf{O}$, the second follows by iterated expectation, the third follows by iterated expectation and by definition of ℓ_0 and m_0 (along with the earlier results for their derivatives with respect to ϵ), and the fourth follows by the expression derived earlier for $\partial \boldsymbol{\psi}_\epsilon / \partial \epsilon |_{\epsilon=0}$.

Therefore, as long as $\mathbf{C}_0$ is invertible, we have $\partial \boldsymbol{\psi}_\epsilon / \partial \epsilon |_{\epsilon=0} = \mathbb{E}_0\{\boldsymbol{\varphi}_0(\mathbf{O}) s_0(\mathbf{O})\}$ and thus $\boldsymbol{\varphi}_P(\mathbf{O})$ is the efficient influence function.

C.3. Double robustness of efficient influence function $\boldsymbol{\varphi}$

Here we will show that $\mathbb{E}\{\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \overline{\pi}, \overline{\lambda}, \overline{\mu})\} = 0$ as long as either $\overline{\pi} = \pi_0$ or $(\overline{\lambda}, \overline{\mu}) = (\lambda_0, \mu_0)$. In this section expectations $\mathbb{E} = \mathbb{E}_0$ and parameters $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ are evaluated under P_0, but we drop the subscript for notational convenience.

First note that

$$\mathbf{C}_0 \mathbb{E}\{\boldsymbol{\varphi}(\mathbf{O}; \boldsymbol{\psi}, \overline{\pi}, \overline{\lambda}, \overline{\mu})\} = \bigg[\mathbf{g}_1'(Z, \mathbf{V}; \boldsymbol{\psi}) \bigg\{ \frac{A - \overline{\lambda}(\mathbf{X}, Z)}{\overline{\pi}(Z \mid \mathbf{X})} \bigg\} - \mathbf{g}_2'(Z, \mathbf{V}; \boldsymbol{\psi}) \bigg\{ \frac{Y - \overline{\mu}(\mathbf{X}, Z)}{\overline{\pi}(Z \mid \mathbf{X})} \bigg\}$$
$$+ \int_{\mathcal{T}} \Big\{ \mathbf{g}_1'(t, \mathbf{V}; \boldsymbol{\psi})\overline{\lambda}(\mathbf{X}, t) - \mathbf{g}_2'(t, \mathbf{V}; \boldsymbol{\psi})\overline{\mu}(\mathbf{X}, t) \Big\} \, dt \bigg]$$

$$
\begin{aligned}
&= \mathbb{E}\Bigg[\mathbf{g}_1'(Z,\mathbf{V};\boldsymbol{\psi})\left\{\frac{\lambda_0(\mathbf{X},Z)-\overline{\lambda}(\mathbf{X},Z)}{\overline{\pi}(Z\mid\mathbf{X})}\right\} - \mathbf{g}_2'(Z,\mathbf{V};\boldsymbol{\psi})\left\{\frac{\mu_0(\mathbf{X},Z)-\overline{\mu}(\mathbf{X},Z)}{\overline{\pi}(Z\mid\mathbf{X})}\right\} \\
&\qquad + \int_{\mathcal{T}}\Big\{\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\overline{\lambda}(\mathbf{X},t) - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\overline{\mu}(\mathbf{X},t)\Big\}\ dt\Bigg] \\
&= \mathbb{E}\int_{\mathcal{Z}}\Bigg[\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\Big\{\lambda_0(\mathbf{X},t)-\overline{\lambda}(\mathbf{X},t)\Big\} - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\Big\{\mu_0(\mathbf{X},t)-\overline{\mu}(\mathbf{X},t)\Big\}\Bigg]\frac{\pi_0(t\mid\mathbf{X})}{\overline{\pi}(t\mid\mathbf{X})}\ dt \\
&\qquad + \int_{\mathcal{T}}\Big\{\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\overline{\lambda}(\mathbf{X},t) - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\overline{\mu}(\mathbf{X},t)\Big\}\ dt\Bigg] \\
&= \mathbb{E}\int_{\mathcal{T}}\Bigg[\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\Big\{\lambda_0(\mathbf{X},t)-\overline{\lambda}(\mathbf{X},t)\Big\} - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\Big\{\mu_0(\mathbf{X},t)-\overline{\mu}(\mathbf{X},t)\Big\}\Bigg]\left\{\frac{\pi_0(t\mid\mathbf{X})}{\overline{\pi}(t\mid\mathbf{X})}-1\right\}dt \\
&\qquad + \int_{\mathcal{T}}\Big\{\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\lambda_0(\mathbf{X},t) - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\mu_0(\mathbf{X},t)\Big\}\ dt\Bigg]
\end{aligned}
$$

where the first equality is true by definition, the second and third holds by iterated expectation given $(\mathbf{X}, Z)$ and $\mathbf{X}$, respectively, and the last follows after rearranging and since $\mathbf{g}_1' = \mathbf{g}_2' = 0$ for $t \notin \text{int}(\mathcal{T})$.

Therefore if $\overline{\pi} = \pi_0$ or $(\overline{\lambda}, \overline{\mu}) = (\lambda_0, \mu_0)$ then $\mathbf{C}_0\mathbb{E}\{\boldsymbol{\varphi}(\mathbf{O};\boldsymbol{\psi},\overline{\pi},\overline{\lambda},\overline{\mu})\}$ equals

$$
\begin{aligned}
&\int_{\mathcal{T}}\mathbb{E}\Big\{\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\lambda_0(\mathbf{X},t) - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\mu_0(\mathbf{X},t)\Big\}\ dt \\
&\qquad = \int_{\mathcal{V}}\int_{\mathcal{T}}\mathbf{g}_1'(t,\mathbf{v};\boldsymbol{\psi})\ \ell_0(t),\mathbf{v}) - \mathbf{g}_2'(t,\mathbf{v};\boldsymbol{\psi})\ m_0(t,\mathbf{v})\Big\}\ dt\ dP(\mathbf{v}) = 0
\end{aligned}
$$

where the first equality follows by iterated expectation given $\mathbf{V}$ (and by the definitions of ℓ and m from the previous section), and the second follows from the restriction in the previous section (after using integration by parts).

C.4. Proof of Theorem 4.3

Theorem 4.3 follows from Theorem 5.31 of van der Vaart (2000), together with the fact that $\mathbb{P}\{\boldsymbol{\varphi}(\mathbf{O};\boldsymbol{\psi},\hat{\boldsymbol{\eta}})\}$ equals

$$
\begin{aligned}
&\mathbb{P}\int_{\mathcal{T}}\Bigg[\mathbf{g}_1'(t,\mathbf{V};\boldsymbol{\psi})\Big\{\lambda_0(\mathbf{X},t)-\hat{\lambda}(\mathbf{X},t)\Big\} - \mathbf{g}_2'(t,\mathbf{V};\boldsymbol{\psi})\Big\{\mu_0(\mathbf{X},t)-\hat{\mu}(\mathbf{X},t)\Big\}\Bigg]\left\{\frac{\pi_0(t\mid\mathbf{X})}{\hat{\pi}(t\mid\mathbf{X})}-1\right\}dt \\
&\qquad = \mathbb{P}\left[\Big\{\mathbf{g}_1'(\lambda_0-\hat{\lambda}) - \mathbf{g}_2'(\mu_0-\hat{\mu})\Big\}(\pi_0-\hat{\pi})\Big/\hat{\pi}\pi_0\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq C||\mathbf{g}_1'(\lambda_0 - \hat{\lambda}) - \mathbf{g}_2'(\mu_0 - \hat{\mu})|| \cdot ||\pi_0 - \hat{\pi}|| \\
&= O_p\left\{\Big(||\lambda_0 - \hat{\lambda}|| + ||\mu_0 - \hat{\mu}||\Big)||\pi_0 - \hat{\pi}||\right\}
\end{aligned}
$$

where the inequality follows by Cauchy-Schwarz ($\mathbb{P}(fg) \leq ||f||\ ||g||$) and boundedness of $1/\hat{\pi}\pi_0$, and the last equality by the triangle inequality and boundedness of $\mathbf{g}_1'$ and $\mathbf{g}_2'$.

C.5. Proof of Theorem 4.4 and double robustness of efficient influence function L

After replacing $\mathbf{g}_j$ with f_j, these proofs follow the same logic as the proofs given in previous sections of Theorem 4.2 and of double robustness of the efficient influence function $\boldsymbol{\varphi}$, respectively, and so are omitted.

BIBLIOGRAPHY

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

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

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

J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59(1):19–35, 1972.

D. W. Andrews. Empirical process methods in econometrics. *Handbook of Econometrics*, 4:2247–2294, 1994.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

M. Baiocchi, D. S. Small, S. Lorch, and P. R. Rosenbaum. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296, 2010.

M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

A. Basu, J. J. Heckman, S. Navarro-Lozano, and S. Urzua. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics*, 16(11):1133–1157, 2007.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

A. S. Blinder. Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455, 1973.

N. E. Breslow, J. M. Robins, and J. A. Wellner. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6(3):447–455, 2000.

P. Carneiro and S. Lee. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208, 2009.

P. Carneiro, J. J. Heckman, and E. J. Vytlacil. Estimating marginal returns to education. *National Bureau of Economic Research Working Paper Series*, Paper 16474:1–32, 2010.

X. Chen, H. Hong, and A. Tarozzi. Semiparametric efficiency in gmm models of nonclassical measurement errors, missing data and treatment effects. Technical report, Cowles Foundation Discussion Paper, 2008.

V. Chernozhukov, C. Hansen, and M. Spindler. Valid post-selection and post-regularization inference: an elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015.

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

I. Díaz and M. J. van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

I. Díaz and M. J. van der Laan. Targeted data adaptive estimation of the causal dose-response curve. *Journal of Causal Inference*, 1(2):171–192, 2013.

J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004, 1992.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, pages 196–216, 1993.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*, volume 66. CRC Press, 1996.

J. Fan, T.-C. Hu, and Y. K. Truong. Robust non-parametric function estimation. *Scandinavian Journal of Statistics*, 21(4):433–446, 1994.

J. Fan, N. E. Heckman, and M. P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.

A. F. Galvao and L. Wang. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 2015.

R. D. Gill and J. M. Robins. Causal inference for complex longitudinal data: the continuous case. *The Annals of Statistics*, 29(6):1785–1811, 2001.

M. E. Glickman and S.-L. T. Normand. The derivation of a latent threshold instrumental variables model. *Statistica Sinica*, pages 517–544, 2000.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2002.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(03):726–748, 2008.

W. Härdle, P. Hall, and J. S. Marron. How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, 83 (401):86–95, 1988.

J. J. Heckman. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, 32(3):441–462, 1997.

J. J. Heckman and R. Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1):239–267, 1985.

J. J. Heckman and P. E. Todd. A note on adapting propensity score matching and selection models to choice based samples. *The Econometrics Journal*, 12(s1):S230–S234, 2009.

J. J. Heckman and E. J. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96(8):4730–4734, 1999.

J. J. Heckman and E. J. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

J. J. Heckman and E. J. Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics*, 6:4779–4874, 2007.

J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654, 1997.

J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

M. A. Hernán and J. M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

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

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

K. Hirano and G. W. Imbens. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, volume 226164, pages 73–84. New York: Wiley, 2004.

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

J. L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics.* Springer, 2009.

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

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

G. W. Imbens. Instrumental variables: an econometricians perspective. *Statistical Science*, 29(3):323–358, 2014.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

E. Ingelsson, C. Lundholm, A. L. Johansson, and D. Altman. Hysterectomy and risk of cardiovascular disease: a population-based cohort study. *European Heart Journal*, 32(6): 745–750, 2011.

N. P. Jewell. Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72(1):11–21, 1985.

N. P. Jewell. *Statistics for Epidemiology.* London: CRC Press, 2003.

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

E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. *In: Statistical Causal Inferences and Their Applications in Public Health Research*, in press.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. in press.

P. Kline. Oaxaca-blinder as a reweighting estimator. *The American Economic Review*, 101 (3):532–537, 2011.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

Q. Lei. *Improved double-robust estimation of missing data and causal inference models and efficient estimation of the average treatment effect on the treated.* PhD thesis, Harvard University, 2011.

Q. Li and J. S. Racine. Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2):485–512, 2004.

Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, 2007.

S. A. Lorch, M. Baiocchi, C. E. Ahlberg, and D. S. Small. The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics*, 130(2):270–278, 2012.

E. Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996.

E. Masry and J. Fan. Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, 24(2):165–179, 1997.

M. D. McHugh, J. Berez, and D. S. Small. Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Health Affairs*, 32(10): 1740–1747, 2013.

R. Neugebauer and M. J. van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

R. Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973.

E. L. Ogburn, A. Rotnitzky, and J. M. Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B*, 77(2): 373–396, 2015.

D. Pollard. *Convergence of Stochastic Processes.* Springer, 1984.

D. Pollard. Empirical processes: theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR, 1990.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

J. M. Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS*, pages 113–159, 1989.

J. M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.

J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer, 2000.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

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

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

M. A. Rosenblum and M. J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6 (2), 2010.

K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology.* Philadelphia: Lippincott Williams & Wilkins, 2008.

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

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

D. B. Rubin and M. J. van der Laan. A general imputation methodology for nonparametric regression with censored data. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 194, 2005.

D. B. Rubin and M. J. van der Laan. Doubly robust censoring unbiased transformations. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 208, 2006a.

D. B. Rubin and M. J. van der Laan. Extending marginal structural models through local, penalized, and additive learning. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 212, 2006b.

A. Sjölander and S. Greenland. Ignoring the matching variables in cohort studies–when is it valid and why? *Statistics in Medicine*, 32(27):4696–4708, 2013.

A. Sjölander, A. L. Johansson, C. Lundholm, D. Altman, C. Almqvist, and Y. Pawitan.

Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Statistical Science*, 27(3):395–411, 2012.

Z. Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

Z. Tan. Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association*, 105(489):157–169, 2010.

E. J. Tchetgen Tchetgen and A. Rotnitzky. Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. *Statistics in Medicine*, 30(4):335–347, 2011.

E. J. Tchetgen Tchetgen and S. Vansteelandt. Alternative identification and inference for the effect of treatment on the treated with an instrumental variable. *Harvard University Biostatistics Working Paper Series*, Paper 166, 2013.

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

M. J. van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 130, 2003.

M. J. van der Laan and J. M. Robins. Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, 93(442): 693–701, 1998.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

M. J. van der Laan and D. B. Rubin. Targeted maximum likelihood learning. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 212, 2006.

M. J. van der Laan and Z. Yu. Comments on inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:910–917, 2001.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

M. J. van der Laan, M. Petersen, and W. Zheng. Estimating the effect of a community-based intervention with two communities. *Journal of Causal Inference*, 1(1):83–106, 2013.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge: Cambridge University Press, 2000.

A. W. van der Vaart. Semiparametric statistics. In *Lectures on Probability Theory and Statistics*, pages 331–457. New York: Springer, 2002.

A. W. van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 29(4):679–686, 2014.

A. W. van der Vaart and M. J. van der Laan. Estimating a survival distribution with current status data and high-dimensional covariates. *The International Journal of Biostatistics*, 2(1):1–40, 2006.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996.

E. J. Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

L. Wang, A. Rotnitzky, and X. Lin. Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, 105(491):1135–1146, 2010.

L. Wasserman. *All of Nonparametric Statistics.* Springer, 2006.

J. M. Wooldridge. *Econometric analysis of cross section and panel data.* MIT Press, 2010.

Z. Zhang, Z. Chen, J. F. Troendle, and J. Zhang. Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706, 2012.

W. Zheng and M. J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 273:1–58, 2010.