Targeted Minimum Loss Based Estimation: Applications and Extensions in Causal Inference and Big Data

by

Samuel David Lendle

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark J. van der Laan, Chair
Professor Sandrine Dudoit
Professor Jasjeet S. Sekhon

Spring 2015

Targeted Minimum Loss Based Estimation: Applications and Extensions in Causal Inference and Big Data

Abstract

Targeted Minimum Loss Based Estimation: Applications and Extensions in Causal Inference and Big Data

by

Samuel David Lendle

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark J. van der Laan, Chair

Causal inference generally requires making some assumptions on a causal mechanism followed by statistical estimation. The statistical estimation problem in causal inference is often that of estimating a pathwise differentiable parameter in a semiparametric or nonparametric model. Targeted minimum loss-based estimating (TMLE) is a framework for constructing an asymptotically linear plug-in estimator for such parameters.

The natural direct effect (NDE) is a parameter that quantifies how some treatment affects some outcome directly, as opposed to indirectly through some mediator value between the treatment and outcome on the causal pathway. In Chapter 2, we introduce the NDE among the untreated and show that under some assumptions the NDE among the untreated is identifiable and equivalent to a statistical parameter as the so called average treatment effect among the untreated. We then present a locally efficient, doubly robust TMLE for the statistical target parameter and apply it to the estimation of the NDE among the untreated in simulations and of the NDE in a data set from an RCT.

Some estimators that adjust for the propensity score (PS) nonparametrically, such as PS matching or stratification by the PS, are robust to slight misspecification of the PS estimator. In particular, if the PS estimator fails to estimate the true propensity score, but still approximates some other balancing score, such methods are still consistent for average treatment effect (ATE). In Chapter 3, we extend a traditional TMLE for the ATE to have this property while still being locally efficient and doubly robust and investigate the performance of the proposed estimator in a simulation study.

Online estimators are estimators that process a relatively small piece of a data set at a time, and can be updated as more data becomes available. Typically, online estimators are used in the large scale machine learning literature, but to our knowledge, have not been used to estimate statistical parameters associated with causal parameters. In Chapter 4, we propose two online estimators for the ATE that are asymptotically efficient and doubly robust in a single pass through a data set. The first is similar to the augmented inverse probability of treatment weighting estimator in the batch setting, and the second involves an additional

targeting step inspired by TMLE, which improves performance in some cases. We investigate the performance of both in a simulation study.

To my mother and father.

Contents

List of Figures

List of Tables

Acknowledgments

I am very grateful to Mark van der Laan for his extraordinary support as my advisor throughout my time at Berkeley. I am also appreciative to Sandrine Dudoit, Maya Petersen, Alan Hubbard, and Maureen Lahiff for their contributions to my education. Special thanks goes to Sharon Norris, who saved me from more missed deadlines and screw ups than I can count, and to Burke Bundy who, in addition to his fantastic computational support, produced a bolt cutter when my combination lock broke keeping me from my laptop.

I am also thankful to Michael Hudgens and Lisa LaVange for their help during my time as a masters student at UNC and for encouraging me and supporting my decision to come to Berkeley for my PhD. My life has changed in so many ways both professionally and otherwise in ways that could not be possible had I not moved to Berkeley.

I'm lucky to have been surrounded by so many incredible students including Susan Gruber, Iván Diáz, Luca Pozzi, Alex Luedtke, Wenjing Zheng, Erin Ledell, Oleg Sofrygin, Linda Tran, Jeremy Coyle, Robin Mejia, Sara Moore, Katie Colborn and Mina Subbaraman. Their discussions and collaborations have been very valuable.

Finally, I am grateful to my father, Don Lendle, and to my mother, Margaret Harper, for their perpetual love, support and encouragement.

Some of the material presented here has been published elsewhere and is joint work with co-authors. Chapter 2 was co-authored with Meenakshi Subbaraman and Mark van der Laan and is published as "Identification and efficient estimation of the natural direct effect among the untreated" in *Biometrics* (Lendle, Subbaraman, and Mark J van der Laan, 2013). Chapter 3 was co-authored with Bruce Fireman and Mark van der Laan and is published as "Balancing score adjusted targeted minimum loss-based estimation" in the *Journal of Causal Inference* (Lendle, Fireman, and Mark J van der Laan, 2015). Chapter 4 was co-authored with Mark van der Laan and is under review for publication in the *Journal of Machine Learning Research*.

Chapter 1

Introduction

In causal inference, we are interested in estimating some parameter in a causal model. For example, we may be interested in the average difference between some outcome among members of a population had they all received some treatment of interest versus had they all received an alternate treatment or control. This is the so called average treatment effect (ATE). In order to estimate a causal parameter from an observed data set, we generally have to make some assumptions that allow us to write the causal parameter as a statistical parameter of the distribution of the observed data (J. Pearl, 2009; M. J. van der Laan and S. Rose, 2011).

Because we usually know very little about the true data generating distribution, we posit a semiparametric statistical model. Targeted minimum loss-based estimation (TMLE) is a general framework for constructing an asymptotically linear plug-in estimator for a pathwise differentiable parameter in a semiparametric statistical model with additional properties such as efficiency and double robustness in some cases (M. J. van der Laan and S. Rose, 2011; Mark J. van der Laan, 2010; Mark J van der Laan, 2010; Mark J. van der Laan and Daniel Rubin, 2006).

The natural direct effect (NDE) is a parameter that quantifies how some treatment affects some outcome directly, as opposed to indirectly through some mediator value between the treatment and outcome on the causal pathway (Judea Pearl, 2001; J.M. Robins and Greenland, 1992). In particular, it is the average change in an outcome had every member of a population received treatment versus a control while some mediator is held at the level it would have been had each member received that control. In Chapter 2, we define the NDE more formally, and introduce the NDE among the untreated as the same quantity but only averaged over the untreated group. We show that under some assumptions the NDE among the untreated is identifiable and equivalent to a statistical parameter as the so called average treatment effect among the untreated. The target statistical parameter is also equivalent to the statistical parameter for an NDE when there are no baseline covariates (M. J. van der Laan and S. Rose, 2011, Chapter 8), or when treatment is independent of baseline covariates like in a randomized control trial (RCT). We then present a locally efficient, doubly robust TMLE for the statistical target parameter and apply it to the estimation of the NDE among

the untreated in simulations and of the NDE in a data set from an RCT.

A balancing score as defined by Rosenbaum and D.B. Rubin (1983) is a function of baseline covariates such that treatment and baseline covariates are independent conditional on that function. Balancing scores play an important role in estimating the ATE and other causal parameters. The propensity score (PS), the probability of receiving treatment given baseline covariates, is perhaps the most well known example of a balancing score. Some estimators that adjust for the propensity score nonparametrically, such as PS matching or stratification by the PS, are robust to slight misspecification of the PS estimator. In particular, if the PS estimator fails to estimate the true propensity score, but still approximates some other balancing score, such methods are still consistent. We say that estimators that have this sort of robustness have the "balancing score property". Though these conditions may not be met often, other estimators that use the PS are not robust to this issue such as inverse probability of treatment weighting estimators and the traditional TMLE for the ATE. In Chapter 3, we extend a traditional TMLE for the ATE to have the balancing score property while still being locally efficient and doubly robust. We investigate the performance of the proposed estimator in a simulation study.

Traditionally, the computational cost of statistical methods is not taken into account. With the size of data sets growing larger and larger, scalability of methods becomes more important, and methods which require multiple passes over a data set may not be feasible in practice. Online estimators are estimators that process a relatively small piece of a data set at a time, and can be updated as more data becomes available. Typically, online estimators are used in the large scale machine learning literature, but to our knowledge, have not been used to estimate statistical parameters associated with causal parameters. In Chapter 4, we propose two online estimators for the ATE that are asymptotically efficient and doubly robust in a single pass through a data set. We call the first an online one-step estimator, which is similar to the augmented inverse probability of treatment weighting estimator in the batch setting The second is called the online targeted one-step estimator, because it involves an additional targeting step inspired by TMLE, which improves performance in some cases. We investigate the performance of both in a simulation study.

Chapter 2

Identification and efficient estimation of the natural direct effect among the untreated

2.1 Introduction

Researchers are often interested in not only the total effect of an exposure on an outcome, but also how the exposure acts to effect the outcome by way of a mediator. For example, suppose there is a dietary intervention designed to reduce the risk of acute myocardial infarction (AMI) which also tends to result in weight loss. An investigator may be interested in the effect of diet on risk of AMI that is not due to weight loss. Specifically, she may ask "how would a patient's risk of AMI have changed due to the intervention diet if their weight had been set to whatever it would have been had the patient not been on the intervention diet?" This sort of effect is known as a natural direct effect (Judea Pearl, 2001; J.M. Robins and Greenland, 1992).

The study of natural direct effects is also known as causal mediation analysis. Direct effects are often defined in the context of the counterfactual or potential outcomes framework, which we use in this paper (Albert, 2008; Albert and Nelson, 2011; Kosuke Imai, Luke Keele, and Yamamoto, 2010; D.B. Rubin, 2004; Tchetgen Tchetgen and Ilya Shpitser, 2011).

Many methods for estimating the natural direct effect require consistent estimation of the conditional distribution of the intermediate variable conditional on treatment and baseline covariates, e.g. K. Imai, L. Keele, and Tingley (2010), Kosuke Imai, Luke Keele, and Yamamoto (2010), Petersen, Sinisi, and M. van der Laan (2006), M. van der Laan and Petersen (2008), VanderWeele (2009), and VanderWeele and Stijn Vansteelandt (2010). If the intermediate variable, like weight loss in the example above, is continuous or multivariate, this becomes difficult without relying on strong parametric assumptions. Jo et al. (2011) describe a propensity score based estimation method but it is restricted to settings with a binary mediator.

Tchetgen Tchetgen and Ilya Shpitser (2011) develop semiparametric theory for the natural direct effect and present a multiply robust estimating equation estimator for the statistical parameter, and Zheng and Mark J van der Laan (2012) develop a targeted minimum loss estimator for statistical parameter.

In this paper, we propose a new causal parameter which we call the natural direct effect among the untreated. We show that this parameter is identifiable under similar assumptions to those of the natural direct effect, and in a randomized controlled trial, it is equal to the natural direct effect. We introduce a sensitivity analysis for some of the assumptions. Additionally we present a targeted minimum loss estimator (TMLE) for the statistical parameter. We investigate the performance of the TMLE compared to other estimators in a simulation study, and demonstrate its use in a real data example. We also define and discuss the estimation of the natural direct effect among the treated as well as the indirect effect among the untreated and among the treated.

2.2 The counterfactual framework and natural direct effects

Following J.M. Robins and Greenland (1992) and Judea Pearl (2001), we define natural direct effects using the counterfactual framework. For an individual, let Z_a be the counterfactual value of the intermediate variable, or mediator, had their exposure, A, been set to a for all $a \in \mathcal{A}$, the set of all possible exposures. Similarly, let Y_{az} be the counterfactual outcome had the individual's exposure and intermediate been set to a and z, respectively, for all $(a, z) \in \mathcal{A} \times \mathcal{Z}$. These values are called counterfactual because in practice, a researcher can only observe the mediator and outcome for the exposure level that an individual was observed to have.

Without loss of generality, let exposure $A = 0$ be the reference or untreated level. The individual natural direct effect is defined as $Y_{aZ_0} - Y_{0Z_0}$. The natural direct effect is also known as the "pure direct effect" (J.M. Robins and Greenland, 1992). This is interpreted as the effect of exposure a relative to the reference level on the outcome not through the mediator. This quantity is different than the individual controlled direct effect, $Y_{az} - Y_{0z}$, where the mediator is set to some specific level z, not necessarily equal to Z_0. Goetgeluk, S. Vansteelandt, and Goetghebeur (2008) and S. Vansteelandt (2009) discuss estimation of controlled direct effects.

2.3 Identifiability

Similarly to M. van der Laan and Petersen (2008), we assume there exists a random variable $X := \{W, A, Z_a, Y_{az} : a \in \mathcal{A}, z \in \mathcal{Z}\}$. In addition, we assume $O := \{W, A, Z = Z_A, Y = Y_{AZ}\}$ is a missing data structure on X where A is the observed exposure, and W represents a possibly multivariate baseline covariate. As implied by the definition of O, we also assume

consistency, that Z is the counterfactual mediator under the observed exposure, and Y is the counterfactual outcome under the observed exposure and mediator.

Let $\mathcal{M}$ be the set of possible probability distributions P for O, and call the true distribution of O P_0. The set $\mathcal{M}$ is called the statistical model. For sake of presentation suppose O is a discrete random variable, so P represents a probability. To allow for continuous random variables, we can assume $\mathcal{M}$ is dominated by a common measure and define densities with respect to that measure. The likelihood of O can be factorized as

$$P(O) = P(W)P(A \mid W)P(Z \mid A, W)P(Y \mid A, Z, W).$$

A causal parameter is a mapping from the full data model into the real numbers, $\Psi^F : \mathcal{M}^F \to \mathbb{R}^k$, where $\mathcal{M}^F$ is the set of all possible data generating distributions of X, known as the causal model or full data model. Let $F_{X_0} \in \mathcal{M}^F$ be the true distribution of X. In order to have any hope of estimating the causal parameter $\Psi^F(F_{X0})$ of interest, we must be able to write it as a functional of only the distribution of the observed data O. That is, we need make some assumptions on $\mathcal{M}^F$ to be able to find some Ψ such that $\Psi^F(F_X) = \Psi(P(F_X))$ for all $F_X \in \mathcal{M}^F$ where $P(F_X)$ is the distribution of O implied by F_X.

Assumption 1 (Randomization)**.**

$$(A, Z) \perp Y_{az} \mid W$$

and

$$A \perp Z_a \mid W$$

Assumption 1 can be interpreted as assuming that the exposure and mediator share no common causes with the outcome and that the exposure shares no common causes with the mediator that are not measured in the set of baseline covariates.

Assumption 2 (Positivity)**.** *For $a \in \mathcal{A}$, $P_0(A = a \mid Z = z, W = w) > 0$ for all (z, w) where $P_0(Z = z, W = w \mid A = 0) > 0$.*

The positivity assumption is also known as experimental treatment assignment (ETA) assumption, and can be interpreted as assuming for every strata of W and Z that can occur when $A = 0$, treatment level a has a non-zero probability of occurring. This assumption is required for the existence of the statistical parameter associated with the causal parameter of interest.

Assumption 3.

$$E(Y_{az} - Y_{0z} \mid Z_0 = z, W) = E(Y_{az} - Y_{0z} \mid W)$$

Assumption 3 means that conditional on baseline covariates, the expected direct effect with an intermediate fixed at level z does not depend on the what the counterfactual mediator value would have been under treatment 0.

Consider the causal parameter

$$\Psi^F(F_X) = DEU(a) = E\Big\{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A = 0\Big\}, \tag{2.1}$$

a generalized natural direct effect among the untreated population.

Theorem 1. *(i) Under the randomization assumption (Assumption 1) and the positivity assumption (Assumption 2), $DEU(a)$ is identifiable. (ii) Additionally under Assumption 3, $DEU(a)$ equals the causal parameter $E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0)$.*

Proofs of theorems are provided in Section 2.A.

Call the causal parameter $E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0)$ the natural direct effect among the untreated, as it is the average of individual natural direct effects among those who have treatment $A = 0$. Theorem 1 is closely related to the identifiability results in M. van der Laan and Petersen (2004), and Assumptions 1 to 3 are analogous to the assumptions for identifiability of

$$DE(a) = E\Big\{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W)\Big\},$$

a generalized natural direct effect discussed in M. van der Laan and Petersen (2008) and for $E(Y_{aZ_0} - Y_{0Z_0})$, the natural direct effect.

The natural direct effect among the untreated (and the natural direct effect) depend on the counterfactual value Y_{aZ_0}, which can never be observed in a real life experiment, because it is the counterfactual outcome under two treatments that cannot occur simultaneously. Because of this, the randomization assumption is not enough to identify the natural direct effect among the untreated, though it is sufficient for identification of controlled direct effects (Petersen, Sinisi, and M. van der Laan, 2006).

Kosuke Imai, Luke Keele, and Yamamoto (2010) provide an alternate set of assumptions for identification of the natural direct effect which they call sequential ignorability. They assume $(Y_{a^*z}, Z_{a'}) \perp A \mid W$ and $Y_{a^*z} \perp Z_{a'} \mid A, W$ for $a^*, a' \in \mathcal{A}$. The natural direct effect among the untreated can also be identified by the sequential ignorability assumption (along with the positivity assumption.) The sequential ignorability assumption implies Assumptions 1 and 3 as shown in Section 2.B, but the converse is not true, so the sequential ignorability assumption is stronger. Even when Assumption 3 does not hold, the causal parameter $DEU(a)$ is interpretable as an average of controlled direct effects averaged with respect to the distribution of the counterfactual Z_0 conditional on the distribution of baseline covariates among the untreated group. For other discussions of identifiability of direct effects, see Avin, I. Shpitser, and J. Pearl (2005), Bullock, Green, and Ha (2010), Hafeman and Vanderweele (2011), Judea Pearl (2001, 2011), J. Robins and Richardson (2010), and J.M. Robins and Greenland (1992).

Under the randomization and positivity assumptions, we know that $DEU(a)$ is identifiable, and we can write $DEU(a)$ as a functional of the observed data generating distribution:

$$\begin{aligned}\Psi(P_0) =& E\Big(\sum_z \Big[\{E(Y \mid A=a, Z=z, W) \\ & - E(Y \mid A=0, Z=z, W)\} \\ & P(Z=z \mid A=0, W)\Big] \mid A=0\Big).\end{aligned} \tag{2.2}$$

Theorem 2. *(i) If A is completely randomized (i.e. $A \perp (W, Z_a, Y_{az})$), then $DEU(a) = DE(a)$. (ii) Additionally under Assumption 3, $E(Y_{aZ_0} - Y_{0Z_0}|A=0) = E(Y_{aZ_0} - Y_{0Z_0})$, so $DEU(a)$ is equal to the natural direct effect.*

In a randomized controlled trial (RCT) where subjects are randomly assigned to a treatment $a \in \mathcal{A}$ independent of baseline covariates W, the conditions for Theorem 2 (i) are satisfied. Furthermore, $A \perp (W, Z_a, Y_{az})$ implies that $A \perp Y_{az} \mid W$ and $A \perp Z_a \mid W$, so the only randomization assumption which is not automatically satisfied is $Z \perp Y_{az} \mid A, W$.

2.4 Estimation

In Section 2.3, we defined the statistical parameter that we are interested in estimating, $\Psi(P_0)$ in (2.2). Let $B = (W, Z)$ and without loss of generality let the exposure level of interest $a = 1$. The target statistical parameter can be written as

$$\psi_0 = \Psi(P_0) = E_0\{E_0(Y \mid A=1, B) - E_0(Y \mid A=0, B) \mid A=0\} \tag{2.3}$$

where E_0 is used to emphasize that an expectation is taken with respect to the true distribution P_0 as opposed to some other distribution belonging to $\mathcal{M}$. Under other causal models, $\Psi(P_0)$ can be interpreted as other interesting causal parameters. For example, if $B = Z$ and Assumptions 1 to 3 are strengthened to $(A, Z) \perp Y_{az}$, $A \perp Z_a$, $P_0(A = 1 \mid Z = z) \in (0, 1)$ almost everywhere, and $E(Y_{az} - Y_{0z} \mid Z_0 = z) = E(Y_{az} - Y_{0z})$, then $\Psi(P_0)$ is the natural direct effect as defined in Alan E. Hubbard, Jewell, and Mark J. van der Laan (2011). Under a different causal model, $\Psi(P_0)$ is equivalent to the so called average treatment effect among the untreated (Hahn, 1998; M. van der Laan, 2010). Hahn (1998) show that if $P_0(A = 1 \mid B)$, known as the propensity score in this setting, is known or belongs to a parametric family, the efficiency bound of $\Psi(P_0)$ is reduced relative to the model where $P_0(A = 1 \mid B)$ is unknown.

The functional Ψ is a mapping from the non-parametric statistical model $\mathcal{M}$ to $\mathbb{R}$. For a distribution $P \in \mathcal{M}$, let $\bar{Q}(a, b) = E_P(Y \mid A = a, B = b)$, $g(a \mid b) = P(A = a \mid B = b)$, and $Q_B(b) = P(B = b)$ for $a \in \{0, 1\}$ and $b \in \mathcal{B}$, the support of B. Let the subscript 0 denote the truth and the subscript n denote an estimate based on n independent observations $O_i = (B_i, A_i, Y_i)$ for $i = 1, \ldots, n$. For example, $\bar{Q}_0$ is the true conditional mean of Y and $\bar{Q}_n$ is an estimate. The mapping Ψ depends on P through $Q = (\bar{Q}, Q_B)$ and g, so

$$\Psi(Q, g) = \sum_b \left[\{\bar{Q}(1, b) - \bar{Q}(0, b)\} \frac{g(0|b)Q_B(b)}{\sum_b \{g(0|b)Q_B(b)\}}\right],$$

recognizing the abuse of notation.

Bickel et al. (1993) show that a regular estimator for a statistical parameter in a semiparametric model is asymptotically efficient, (i.e. the estimator has minimal asymptotic variance,) if it is asymptotically linear with influence curve (influence function) equal to the efficient influence curve. This minimal asymptotic variance is known as the semiparametric efficiency bound and is the variance of the efficient influence curve. The non-parametric model $\mathcal{M}$ is a special case of a semiparametric model where there are no restrictions on the possible distributions of O. The efficient influence curve for Ψ at $P \in \mathcal{M}$, derived in M. van der Laan (2010), is

$$D^*(P) = D^*(Q, g, \Psi(Q, g)) = \left\{ \frac{I(A=1)}{P(A=0)} \frac{g(0 \mid B)}{g(1 \mid B)} - \frac{I(A=0)}{P(A=0)} \right\} \{Y - \bar{Q}(A, B)\} + \frac{I(A=0)}{P(A=0)} \{\bar{Q}(1, B) - \bar{Q}(0, B) - \Psi(Q, g)\}$$

where $I(\cdot)$ is an indicator function. The semiparametric efficiency bound for an analogous statistical parameter, where the difference is conditioned on $A = 1$, is also derived in Hahn (1998).

The efficient influence curve for $\Psi(P_0)$ has the double robustness property. That is,

$$P_0 D^*(Q, g_0, \psi_0) = P_0 D^*(Q_0, g, \psi_0) = 0,$$

where $Pf := \int f(o) dP(o) = \sum_o f(o) P(O = o)$ is the expectation of f under distribution P. This means that if we have an estimator that solves the efficient influence curve equation, (i.e. $P_n D^*(Q_n, g_n, \Psi(Q_n, g_n)) = 0$,) then it is consistent if at least one of the estimators Q_n or g_n are consistent for Q_0 or g_0 under regularity conditions (M. van der Laan, 2010). Additionally, the efficiency bound is achieved if both Q_n and g_n are consistent estimators for Q_0 and g_0, so such an estimate is locally efficient at P_0.

In Alan E. Hubbard, Jewell, and Mark J. van der Laan (2011) and M. van der Laan (2010), a targeted minimum loss estimator (TMLE) is developed for $\Psi(P_0)$. The TMLE solves the efficient influence curve and is a locally efficient, double robust estimator. It is also a substitution or plug-in estimator in the sense that estimators for Q_0 and g_0 can be plugged into the mapping Ψ to calculate an estimate as

$$\Psi(Q_n, g_n) = \frac{1}{\sum_{i=1}^n I(A_i = 0)} \sum_{i=1}^n I(A_i = 0) \{\bar{Q}_n(1, B_i) - \bar{Q}_n(0, B_i)\} \tag{2.4}$$

for some estimates Q_n and g_n. That is, the estimate is the difference $\bar{Q}_n(1, B_i) - \bar{Q}_n(0, B_i)$ for each individual averaged with respect to the empirical distribution of B given $A = 0$. We review the TMLE for $\Psi(P_0)$ here.

Begin by constructing initial estimates for $\bar{Q}_0$ and g_0 called $\bar{Q}_n^0$ and g_n^0. If we have expert background knowledge about the functional forms of $\bar{Q}_0$ and g_0, they can be estimated by parametric models. In general there is not enough background knowledge to support

parametric models, and $\bar{Q}_n^0$ and g_n^0 should be constructed by some non-parametric data adaptive learning algorithm such as the super learner (Mark J van der Laan, Polley, and Alan E Hubbard, 2007), which combines machine learning algorithms and parametric models using cross validation. To calculate the TMLE, we update the initial estimates $\bar{Q}_n^0$ and g_n^0 to $\bar{Q}_n^*$ and g_n^*, and then plug them in to Ψ, so the final estimate is $\Psi(Q_n^*, g_n^*)$, where $Q_n^* = (\bar{Q}_n^*, Q_{Bn})$ and Q_{Bn} is the empirical distribution of B.

Suppose Y is binary. Other cases are discussed in Section 2.C. To update the initial estimates, for $j = 1, 2, \ldots$, calculate until convergence

$$\text{logit}(\bar{Q}_n^j(A, B)) = \text{logit}(\bar{Q}_n^{j-1}(A, B)) + \epsilon_{1n}^j C_1^{j-1}(A, B) \tag{2.5}$$

and

$$\text{logit}(g_n^j(0 \mid B)) = \text{logit}(g_n^{j-1}) + \epsilon_{2n}^j C_2^{j-1}(B) \tag{2.6}$$

where $\text{logit}(p) = \log(p/(1-p))$,

$$C_1^{j-1}(A, B) = \frac{I(A=1)}{P_n(A=0)} \frac{g_n^{j-1}(0 \mid B)}{g_n^{j-1}(1 \mid B)} - \frac{I(A=0)}{P_n(A=0)},$$

$$C_2^{j-1}(B) = P_n(A=0)^{-1}\{\bar{Q}_n^{j-1}(1, B) - \bar{Q}_n^{j-1}(0, B) - \Psi(Q_n^{j-1}, g_n^{j-1})\},$$

$P_n(A=0)$ is the empirical probability of $A=0$, and ϵ_{1n}^j and ϵ_{2n}^j maximum likelihood estimates in the logistic regression models in (2.5) and (2.6), respectively. The coefficients ϵ_{1n}^j and ϵ_{2n}^j can be calculated with standard logistic regression software where $\bar{Q}_n^{j-1}(A, B)$ and $g_n^{j-1}(0 \mid B)$ are offset terms. Convergence is reached when both $\epsilon_{1_n}^j$ and ϵ_{2n}^j are close to 0 and so estimates of $\bar{Q}_0$ and g_0 are changing very little. Set $\bar{Q}_n^* = \bar{Q}_n^j$ and $g_n^* = g_n^j$ at the last iteration.

Under regularity conditions on the initial estimates Q_n^0 and g_n^0, the TMLE is regular and asymptotically linear (Mark J. van der Laan and Sherri Rose, 2011), so $\sqrt{n}(\Psi(P_n^*) - \Psi(P_0)) \xrightarrow{d} N(0, \sigma^2)$. When Q_n^0 and g_n^0 are consistent estimators for Q_0 and g_0, the variance σ^2 is the variance of the efficient influence curve. In order to estimate the variance σ^2, we can use an estimate of the sample variance of the estimated influence curve $D^*(Q_n^*, g_n^*)$. Wald type hypothesis tests can be performed, and confidence intervals can be constructed with the estimated variance σ_n^2.

When either Q_n^0 or g_n^0 is not consistent, the influence curve based variance estimate is biased and not guaranteed to be conservative. If one assumes g_n^0 is a consistent MLE, then one can compute a correction term for the influence curve which only depends on the behavior of g_n^0 (M. J. van der Laan and J. M. Robins, 2003, Section 2.3.7). Alternatively, the non-parametric bootstrap can be used to estimate the variance of the TMLE in the standard way by resampling n observations many times from the original data and calculating the TMLE for each resampled dataset of n observations. The variance is estimated as the sample variance of the estimates of $\Psi(P_0)$ from each resampled data set. When initial estimates Q_n^0 and g_n^0 are differentiable functionals of the empirical distribution, as is the case for parametric maximum likelihood estimators, then the TMLE is also differentiable, so the bootstrap estimate of the variance is known to be consistent (Gill, Wellner, and Præ stgaard, 1989).

2.5 Sensitivity analysis

Sensitivity analyses have been proposed to investigate violations of $Y_{a^*z} \perp Z_{a'} \mid A, W$ of the sequential ignorability assumption when estimating the natural direct effect, which in general require specification of some model describing the violation of the assumption. Kosuke Imai, Luke Keele, and Yamamoto (2010) and Tchetgen Tchetgen and Ilya Shpitser (2011) require that the mediator take on two or a finite number of values. VanderWeele and Stijn Vansteelandt (2010) require specification of the relationship between a hypothetical unmeasured confounder and the observed variables. Here we propose an alternative method to investigate violations of $Z \perp Y_{az} \mid A, W$, implied by Assumption 1, that does not require the support of the mediator to be finite. Because the causal parameter $DEU(a)$ is identifiable under only Assumption 1 and the positivity assumption, we focus on deviations from Assumption 1 and not Assumption 3.

First assume $A \perp Z_a \mid W$ and $A \perp Y_{az} \mid W$, which is known when A is completely randomized as in an RCT. Let $E(Y_{az} \mid W = w) = m_{\bar{Q}_0}(a, z, w)$. If $m_{\bar{Q}_0}(a, Z, W) = \bar{Q}_0(a, (W, Z))$ does not hold almost everywhere, then $Z \perp Y_{az} \mid A, W$ must be violated. Suppose $m_{\bar{Q}_0} = m^{\alpha}_{\bar{Q}_0}$ is known up to $\bar{Q}_0$ and parameterized by real valued α. A sensitivity analysis is performed by estimating the causal parameter based on specified functions $m^{\alpha}_{\bar{Q}_0}$ using $\bar{Q}^*_n$ in place of $\bar{Q}_0$ to see how the estimated causal parameter can deviate from the statistical estimate under various violations of the randomization assumption. In Section 2.D, we discuss this approach in detail and include an example application to the data set presented in Section 2.7.

These methods can highlight how the deviation between the statistical and causal parameter behaves as a function of the violation of an assumption, but in general are not conclusive due to arbitrary choices regarding the parameterization of the violation of an assumption. Choosing such a parameterization and range of interpretable parameter values is very difficult, particularly in cases where the mediator is continuous or high dimensional. Nonetheless, these or similar methods can be useful in identifying departures from Assumption 1 or Assumption 3 in some cases.

2.6 Simulation study

To explore the performance of the TMLE in Section 2.4 we compare the TMLE to other types of estimators in a simulation study The first alternative estimator is known as the G-computation or maximum likelihood based estimator (MLE), and depends only on an initial estimate $\bar{Q}^0_n$. The estimate is computed by plugging $\bar{Q}^0_n$ into (2.4) and averaging with respect to the empirical distribution of B where $A = 0$. An inverse probability of treatment weighted (IPTW) type estimator is also presented, which is a function of an initial estimate of g_0. The estimate is computed as

$$\psi_n = n^{-1} \sum_i \left\{ \frac{I(A_i = 1)}{P_n(A_i = 0)} \frac{g^0_n(0 \mid B_i)}{g^0_n(1 \mid B_i)} - \frac{I(A_i = 0)}{P_n(A_i = 0)} \right\} Y_i.$$

We bound $g_n^0(1 \mid B)$ above 0.001 to mitigate the effect of extremely large weights on the estimate, a method discussed by Cole and Hernán (2008). See J M Robins, Hernán, and Brumback (2000) for a detailed treatment of IPTW estimators. Because these two estimators depend only on either Q or g, they are not double robust and we expect them to be biased if estimates of Q_0 or g_0 are not consistent.

Suppose we observe two independent baseline covariates. The first, W_1, has a Bernoulli distribution with mean 0.3, and the second, W_2 has a standard normal distribution. We also observe a binary treatment variable A, and a mediator Z. Suppose Z has a normal distribution with mean $|3W_1|$ and variance one, and A equals one with probability $\text{logit}^{-1}(-2.5 + 3W_1 + 0.2Z)$. Also suppose we observe a binary outcome, Y, which is one with probability $\text{logit}^{-1}(1.4A - 2.5Z + W1)$. Call the true distribution of $O = \{W_1, W_2, A, Z, Y\}$ P_0.

The statistical parameter $\psi(P_0) \approx 0.0872$ and the variance bound for a sample of size n is approximately $1.004/n$. The true parameter and variance bound were computed by Monte Carlo simulation. By the construction of P_0 we can see that the true $\bar{Q}_0$ is contained in a main terms logistic regression model including W_1, W_2, A, and Z as explanatory variables, and the true g_0 is contained in a main terms logistic regression model including W_1, W_2, and Z as explanatory variables. For sake of illustration, we construct initial estimates $\bar{Q}_n^0$ and g_n^0 using logistic regression, which we know will be consistent as long as all necessary independent variables are included in the model. In practice we would turn to data adaptive methods for the initial estimates when we do not have enough knowledge to guarantee that estimators based on parametric models are consentent for $\bar{Q}_0$ and g_0. In the simulations, the misspecified model for $\bar{Q}$ is a main terms logistic regression model with only A as an explanatory variable, and the misspecified model for g has only Z as an explanatory variable.

Results from $1,000$ datasets drawn from P_0 of size $n = 50$, $n = 200$ and $n = 1000$ are shown in Table 2.1. When the models are correctly specified, all three estimators have low bias, and the variance of TMLE estimates approaches the efficiency bound as sample size increases, demonstrating that the TMLE is locally efficient. We also see that bootstrap estimates of the variance are close to the observed variance. When the model for $\bar{Q}_0$ is misspecified, we see the MLE has a large bias which does not decrease with sample size. Similarly when the model for g_0 is misspecified, the IPTW estimator has a large bias. However, when one of the models for $\bar{Q}_0$ or g_0 is misspecified, TMLE still has low bias, demonstrating the double robustness property.

2.7 Application

To illustrate the TMLE, we use a subset of data from the COMBINE study, a multi-center RCT to evaluate the efficacy of medication, behavioral therapies, and their combinations to treat alcohol dependence (Anton et al., 2006). Naltrexone, one of the medical therapies in the study, is thought to act via reducing cravings for alcohol (Volpicelli et al., 1995). The combined behavioral intervention (CBI) integrated a variety of well-supported treatment

Table 2.1: Simulation results from an observational study. Variance bounds were 0.0201, 0.005, and 0.001 for sample sizes 50, 200, and 1000 respectively. Sample sizes are in parentheses.

Model	Bias			Observed Variance			Bootstrap Var. Est.		
	(50)	(200)	(1000)	(50)	(200)	(1000)	(50)	(200)	(1000)
Q, g correct									
TMLE	−0.007	−0.005	0.002	0.029	0.007	0.001	0.048	0.009	0.001
MLE	0.010	0.000	0.002	0.024	0.003	0.001	0.031	0.003	0.001
IPTW	−0.005	−0.005	−0.001	0.116	0.014	0.002	0.208	0.018	0.003
Q misspecified									
TMLE	−0.021	−0.013	−0.001	0.049	0.011	0.002	0.055	0.010	0.002
MLE	−0.024	−0.025	−0.023	0.014	0.004	0.001	0.015	0.004	0.001
g misspecified									
TMLE	0.004	0.000	0.002	0.022	0.003	0.001	0.029	0.004	0.001
IPTW	0.044	0.042	0.045	0.019	0.004	0.001	0.029	0.004	0.001

methods such as motivational interviewing and cognitive-behavioral skills training. An investigator may be interested in how CBI acts to reduce drinking. For this example, we define the outcome of interest Y as percent days of heavy drinking in the third month after treatment. We define our parameter of interest as the NDE on percent days of heavy drinking of CBI compared to a placebo not through reduction of cravings.

For the mediator, Z, we use a measurement of cravings collected at 4, 8, and 12 weeks after baseline. Because the trial is randomized, we know $A \perp Y_{az}$ and $A \perp Z_a$, and in order to identify the NDE we must be able to assume $Z \perp Y_{az} \mid W$ for some set of baseline covariates W in addition to Assumptions 2 and 3. For W, we include a measurement of baseline cravings and percent days abstinent from drinking prior to baseline.

Our dataset has 420 observations, with 227 patients assigned to CBI and 193 patients assigned to placebo. We present estimates of the NDE and also include estimates of the total effect for comparison. For initial estimates Q_n and g_n, we use a data adaptive super learner, combining GLM regression with all main terms, GLM with main terms and pairwise interaction terms chosen by stepwise selection, and the elastic net algorithm (Friedman, T Hastie, and Tibshirani, 2010; Zou and T. Hastie, 2005) with main and pairwise interaction terms. Results are presented in Table 2.2. Estimators include TMLE, MLE and IPTW. For the natural direct effect, all three estimates are negative, suggesting that CBI reduces percent days of heavy drinking relative to placebo by some mechanism other than by effecting cravings. Additionally, the estimates of the total effect are larger in magnitude, suggesting that some of the effect of CBI on percent days of heavy drinking is due to the effect on cravings.

To investigate the effect on the estimates due to leaving out a potentially important confounder from the set of baseline covariates, we present results where baseline percent days abstinent is excluded from W in Table 2.3. We know that treatment is independent of baseline percent days abstinent, so excluding it from W will not bias estimates of the total effect for

Table 2.2: Total and direct effect of CBI on percent days of heavy drinking mediated by cravings from COMBINE study.

	Estimate	Bootstrap CI	Bootstrap SE	Influence curve SE
Total effect				
TMLE	-5.20	(-10.07, -0.33)	2.48	2.50
MLE	-5.04	(-9.87, -0.21)	2.46	
IPTW	-5.23	(-10.14, -0.31)	2.51	
Natural direct effect				
TMLE	-4.33	(-8.51, -0.14)	2.13	2.02
MLE	-3.03	(-7.12, 1.06)	2.09	
IPTW	-4.93	(-9.6, -0.25)	2.39	

Table 2.3: Total and direct effect of CBI on percent days of heavy drinking mediated by cravings from COMBINE study, not adjusting for baseline percent days abstinent.

	Estimate	Bootstrap CI	Bootstrap SE	Influence curve SE
Total effect				
TMLE	-5.39	(-10.38, -0.39)	2.55	2.53
MLE	-5.29	(-10.34, -0.23)	2.58	
IPTW	-5.23	(-10.24, -0.21)	2.56	
Natural direct effect				
TMLE	-4.02	(-8.48, 0.45)	2.28	2.16
MLE	-3.74	(-8, 0.52)	2.17	
IPTW	-4.71	(-9.66, 0.25)	2.53	

any estimators, but we do not know if baseline percent days abstinent is a confounder of the mediator and outcome, so we do not know if differences in estimates observed here in the estimates for the NDE or $DE(1)$ are due to bias. The difference between estimates in Tables 2.2 and 2.3 is small relative to estimated standard errors, so in this case sample size is too small to determine if failing to include a potential confounder introduces bias. For all estimates, estimated standard errors are smaller when adjusting for baseline percent days abstinent, indicating that efficiency is gained even if the variable is not a confounder. In Section 2.D, we demonstrate an application of the sensitivity analysis proposed in Section 2.5 to this dataset.

2.8 Discussion

In this paper we proposed a new causal parameter called the natural direct effect among the untreated, and we provide and discuss identifiability results in Section 2.3. In Section 2.4, we describe a targeted minimum loss estimator that is a locally efficient and double robust substitution estimator for the statistical parameter $\Psi(P_0)$. In Theorem 2 we show when A is

completely randomized, such as in an RCT, this natural direct effect among the untreated is equal to the natural direct effect, and therefore the natural direct effect can be estimated with the method in Section 2.4. Even when A is not completely randomized, an estimate of $\Psi(P_0)$ can always be interpreted as the $DEU(a)$ under Assumption 1, that is, an average of direct effects weighted by the empirical distribution of baseline covariates W among the unexposed subjects with $A = 0$.

We point out that efficient estimators for $\Psi(P_0)$ in the non-parametric model are not fully efficient in the semiparametric model where A is completely randomized. When the knowledge that $A \perp W$ is ignored and $g_0(A \mid B) = P_0(A \mid W, Z)$ is estimated without restriction, some information about $\Psi(P_0)$ is lost and the efficient influence curve in this semiparametric model is not equal to D^* (Tchetgen Tchetgen and Ilya Shpitser, 2011; Zheng and Mark J van der Laan, 2012). Although the TMLE in Section 2.4 is not fully efficient when A is completely randomized, we argue it is still useful as an alternative and relatively simple estimator for the NDE that does not require estimation of the conditional density of the mediator in addition to being an estimator for the NDE among the untreated. Below we discuss other causal parameters to which the TMLE can be applied.

In addition to the NDE and the NDE among the untreated, researchers may also be interested in the NDE among the treated, defined as $E(Y_{aZ_0} - Y_{0Z_0} \mid A = a)$. Under appropriate identifiability conditions, this causal parameter corresponds to the statistical parameter

$$\begin{aligned}\Psi'(P_0) =& E\Big(\sum_z \big[\{E(Y \mid A = a, Z = z, W) \\ & - E(Y \mid A = 0, Z = z, W)\} \\ & P(Z = z \mid A = 0, W)\big] \mid A = a\Big).\end{aligned}$$

Because the conditional probability of Z is conditional on $A = 0$ inside the square brackets, but the expectation of the expression in square brackets is conditioned on $A = a$, $\Psi'(P_0)$ cannot be written in the form of (2.3) and cannot be estimated using a method similar to that in Section 2.4. However, when there are only two levels of treatment so A is binary, then $\Psi^*(P_0) = \Psi'(P_0)P_0(A = 1) + \Psi(P_0)P_0(A = 0)$ where

$$\begin{aligned}\Psi^*(P_0) =& E\Big(\sum_z \{E(Y \mid A = 1, Z = z, W) \\ & - E(Y \mid A = 0, Z = z, W)\} \\ & P(Z = z \mid A = 0, W)\Big)\end{aligned}$$

is the statistical parameter associated with the natural direct effect. We can write $\Psi'(P_0) = \{\Psi^*(P_0) - \Psi(P_0)P_0(A = 0)\}/P_0(A = 1)$. Based on this we can see that $\Psi'(P_0)$ can be estimated using an estimate for $\Psi^*(P_0)$ such as those proposed by Tchetgen Tchetgen and Ilya Shpitser (2011) and Zheng and Mark J van der Laan (2012) as well as an estimate for $\Psi(P_0)$ based on the methodology in Section 2.4.

Another causal parameter that may be of interest to researchers is called the indirect effect (IE) among the untreated, defined as $E(Y_{aZ_a} - Y_{aZ_0} \mid A = 0)$. This definition is analogous to the total indirect effect of J.M. Robins and Greenland (1992) and the indirect effect of M. van der Laan and Petersen (2004). Similarly to the total effect (TE), the TE among the untreated or average effect of treatment among the untreated (ATU), defined as $E(Y_{aZ_a} - Y_{0Z_0} \mid A = 0)$ in current notation, can be decomposed as the sum of the NDE among the untreated and the IE among the untreated. That is,

$$E(Y_{aZ_a} - Y_{0Z_0} \mid A = 0) = E(Y_{aZ_a} - Y_{aZ_0} \mid A = 0) + E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0).$$

Because of this decomposition, if the ATU and the NDE among the untreated are identifiable, the IE among the untreated can also be identified and can be estimated based on estimates of the ATU and the NDE among the untreated. Identifiability of the average treatment effect among the (un)treated is discussed in M. van der Laan (2010). Analogously, this relationship also holds for the TE among the treated, the NDE among the treated, and the IE among the treated so the indirect effect among the untreated can be estimated similarly.

A final alternative causal parameter of interest may be defined as $E(Y_{aZ_a} - Y_{0Z_a} \mid A = a)$. This parameter is similar to the NDE among the treated, but the intermediate variable is set to the value it would have been under treatment a instead of treatment 0. Under appropriate identifiability assumptions, this is equal to the statistical parameter

$$E_0\{E_0(Y \mid A = a, B) - E_0(Y \mid A = 0, B) \mid A = a\}. \tag{2.7}$$

This statistical parameter is similar to (2.3), but now the difference is conditional on $A = a$. An analogous estimator to that developed in Section 2.4 could be used to estimate (2.7).

2.A Proof of theorems

Proof of Theorem 1

For (i), by the randomization assumption we can write

$$\begin{aligned} P(Y = y \mid A = a, Z = z, W) &= P(Y_{az} \mid A = a, Z = z, W) \\ &= P(Y_{az} \mid W) \\ P(Z = z \mid A = a, W) &= P(Z_a = z \mid A = a, W) \\ &= P(Z_a = z \mid W), \end{aligned}$$

so

$$\begin{aligned} DEU(a) &= E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A = 0\} \\ &= E[E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A = 0, W\} \mid A = 0] \\ &= E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid A = 0, W) P(Z_0 = z \mid W) \mid A = 0\} \\ &= E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z_0 = z \mid W) \mid A = 0\} \\ &= E([\textstyle\sum_z \{E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W)\} \\ &\qquad P(Z = z \mid A = 0, W)] \mid A = 0), \end{aligned}$$

therefore $DEU(a)$ is identifiable. For (ii),

$$E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0) = E[E\{E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0, A = 0, W) \mid A = 0, W\} \mid A = 0]$$

and

$$\begin{aligned} & E\{E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0, A = 0, W) \mid A = 0, W\} \\ = \; & \textstyle\sum_z E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0 = z, A = 0, W) P(Z_0 = z \mid A = 0, W) \\ = \; & \textstyle\sum_z E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0 = z, W) P(Z_0 = z \mid W) \text{ by Assumption 1} \\ = \; & \textstyle\sum_z E(Y_{az} - Y_{0z} \mid Z_0 = z, W) P(Z_0 = z \mid W) \\ = \; & \textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z_0 = z \mid W) \text{ by Assumption 3} \end{aligned}$$

so

$$\begin{aligned} E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0) = \; & E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z_0 = z \mid W) \mid A = 0\} \\ = \; & DEU(a) \end{aligned}$$ □

Proof of Theorem 2

For (i),

$$\begin{aligned} DEU(a) = \; & E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W) \mid A = 0\} \\ = \; & \textstyle\sum_w \{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W)\} P(W = w|A = 0) \\ = \; & \textstyle\sum_w \{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W)\} P(W = w) \text{ by } A \text{ completely randomized} \\ = \; & E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W)\} \\ = DE(a) & \end{aligned}$$

The proof for (ii) follows from (i) of this theorem and the proof of Theorem 1 (ii). □

2.B Sequential ignorability implies Assumptions 1 and 3

Note $Z = Z_A$, and

$$\begin{aligned} (A, Z_A) \perp Y_{az} \mid W \iff & \{A \perp Y_{az} \mid W \\ & Z_A \perp Y_{az} \mid A, W\} \end{aligned}$$

For $a^*, a' \in \mathcal{A}$, we see $(Y_{a^*z}, Z_{a'}) \perp A \mid W$ implies $A \perp Z_a \mid W$ and $A \perp Y_{az} \mid W$. Additionally, $Y_{a^*z} \perp Z_{a'} \mid A, W$ implies $Y_{az} \perp Z_A \mid A, W$, so sequential ignorability implies Assumption 1.

Now,

$$\begin{aligned} E(Y_{a^*z} \mid Z_0 = z, W) = & E\{E(Y_{a^*z} \mid Z_0 = z, A, W) \mid Z_0 = z, W\} \\ = & E\{E(Y_{a^*z} \mid A, W) \mid Z_0 = z, W\} \text{ by } Y_{a^*z} \perp Z_{a'} \mid A, W \\ = & E\{E(Y_{a^*z} \mid A, W) \mid W\} \text{ by } (Y_{a^*z}, Z_{a'}) \perp A \mid W \\ = & E(Y_{a^*z} \mid W). \end{aligned}$$

This implies Assumption 3.

2.C Modifications to the TMLE algorithm for non-binary Y

If Y is not binary, but is bounded by 0 and 1, for example a proportion, the algorithm does not need to be modified because the negative quasibinomial likelihood is a valid loss function for estimating a conditional mean. Standard software for logistic regression can still be used for estimating ϵ_{1n}^j, though warning messages may be produced because the outcome variable is not binary. More generally, if Y is bounded by l and u with $l < u$, the algorithm can be modified by transforming Y to Y' bounded between 0 and 1. After estimating $\bar{Q}_n^0(A, B)$, calculate $Y' = (Y - l)/(u - l)$ and $\bar{Q}_n'^0(A, B) = (\bar{Q}_n^0(A, B) - l)/(u - l)$, and perform the updating steps with Y' and $\bar{Q}_n'^0$ in place of Y and $\bar{Q}_n^0$. After convergence, calculate the final estimate by multiplying $\Psi(Q_n^*, g_n^*)$ by $u - l$.

Alternatively, instead of updating the estimate Q_n^i on the logit scale, we can updated it on the linear scale by replacing (2.5) with

$$\bar{Q}_n^j(A, B) = \bar{Q}_n^{j-1}(A, B) + \epsilon_{1n}^j C_1^{j-1}(A, B)$$

where ϵ_{1n}^j is estimated with maximum likelihood or least squares in the linear model

$$\bar{Q}(A, B) = \epsilon_1^j C_1^{j-1}(A, B) + \bar{Q}_n^{j-1}(A, B).$$

In small samples, when Y is in fact bounded by between l and u, a linear update can yield final estimates that do not respect these bounds. For example, suppose Y is a proportion and therefore between 0 and 1. This implies ψ_0 must be between -1 and 1, but a linear update could potentially yield estimates outside of this interval.

2.D Sensitivity analysis

Assume $A \perp Z_a \mid W$ and $A \perp Y_{az} \mid W$ of the randomization assumption, but $Z \perp Y_{az} \mid A, W$ is not necessarily true. Following proof of Theorem 1 (i) and given $m_{\bar{Q}_0}(a, z, w) = E(Y_{az} \mid W = w)$, the causal parameter $DEU(a)$ is calculated as

$$\begin{aligned}
DEU(a) = \;& E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A = 0\} \\
= \;& E[E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A = 0, W\} \mid A = 0] \\
= \;& E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid A = 0, W) P(Z_0 = z \mid W) \mid A = 0\} \\
= \;& E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z = z \mid A = 0, W) \mid A = 0\} \\
& \text{by } A \perp Z_a \mid W \text{ and } A \perp Y_{az} \mid W, \\
= \;& E([\textstyle\sum_z \{m_{\bar{Q}_0}(a, z, W) - m_{\bar{Q}_0}(0, z, W)\} P(Z = z \mid A = 0, W)] \mid A = 0)
\end{aligned}$$

Possible choices for $m_{\bar{Q}}$ are

$$m_{\bar{Q}}^{\alpha}(a, z, w) = \bar{Q}(a, (w, z)) + \alpha r(a, z, w)',$$

or, if Y_{az} is bounded by l and u with $l < u$,

$$m^{\alpha}_{\bar{Q}}(a,z,w) = (u-l)\text{logit}^{-1}\left[\text{logit}\left(\frac{\bar{Q}(a,(w,z))-l}{u-l}\right) + \alpha r(a,z,w)'\right] + l$$

where $\alpha \in \mathbb{R}^d$ and r is a known d dimensional function. We call these two choices of $m_{\bar{Q}}$ a linear deviation and a logistic deviation, respectively. For both of these choices of $m_{\bar{Q}}$, $m^{\alpha}_{\bar{Q}_0} = \bar{Q}_0$ when $\alpha = 0$.

For a given α, call the estimate of $DEU(a)$ ψ^{α}_n, and calculate it as

$$\psi^{\alpha}_n = \frac{\sum_{i=1}^n I(A_i = 0)[m^{\alpha}_{\bar{Q}^*_n}(a, Z_i, W_i) - m^{\alpha}_{\bar{Q}^*_n}(0, Z_i, W_i)]}{\sum_{j=1}^n I(A_j = 0)}.$$

In general, ψ^{α}_n is consistent when the initial estimate $\bar{Q}^0_n$ is consistent. When the linear deviation is used for $m^{\alpha}_{\bar{Q}}$, we can write

$$\begin{aligned} DEU(a) =& E_0(\bar{Q}_0(a,(Z,W)) - \bar{Q}_0(0,(Z,W)) \mid A = 0) \\ &+ E_0(\alpha r(a,Z,W)' - \alpha r(0,Z,W)' \mid A = 0) \\ =& \psi_0 + E_0(\alpha r(a,Z,W)' - \alpha r(0,Z,W)' \mid A = 0) \end{aligned}$$

and analogously

$$\psi^{\alpha}_n = \psi_n + \frac{\sum_{i=1}^n I(A_i = 0)[\alpha r(a,Z_i,W_i)' - \alpha r(0,Z_i,W_i)']}{\sum_{j=1}^n I(A_j = 0)}$$

where ψ_n is the TMLE of ψ_0. In this case, the estimate ψ^{α}_n is doubly robust, because ψ_n is consistent for ψ_0 when either $\bar{Q}^0_n$ or g^0_n is consistent, and

$$E_0(\alpha r(a,Z,W)' - \alpha r(0,Z,W)' \mid A = 0)$$

is estimated with an empirical mean of i.i.d. random variables, which is consistent by the law of large numbers.

A sensitivity analysis can be performed by calculating ψ^{α}_n for $\alpha \in \{\alpha_1, \ldots \alpha_k\}$. The deviation between ψ^{α}_n and ψ^0_n can be interpreted as the sensitivity of the analysis to a violation of $Z \perp Y_{az} \mid A, W$ described by $m_{\bar{Q}_0}$ as a function of α.

In an application to the data set from the COMBINE study in Section 2.7, we parametrize the deviation from the $Z \perp Y_{az} \mid A, W$ assumption in two ways using the logistic deviation. In the first we set $r(a,z,w) = \bar{z}$ where $\bar{z}$ is the average level of the mediator, cravings, measured at three time points, and standardized to have mean zero and variance one. For the second, we set $r(a,z,w) = w_c$ where w_c is the baseline value of cravings again standardized to have mean zero and variance one.

Results are plotted in Figures 2.1 and 2.2. The open circle and the confidence interval at $\alpha = 0$ are equal to the estimate of the NDE in Table 2 in the main paper, where baseline percent days abstinent is included in W. The pointwise 95% confidence limits were calculated based on the bootstrapped standard error at each value of α. For both deviations, we see that the 95% confidence interval only excludes 0 for α near 0. This indicate that our results may not be robust to violations of the randomization assumption for the deviations investigated.

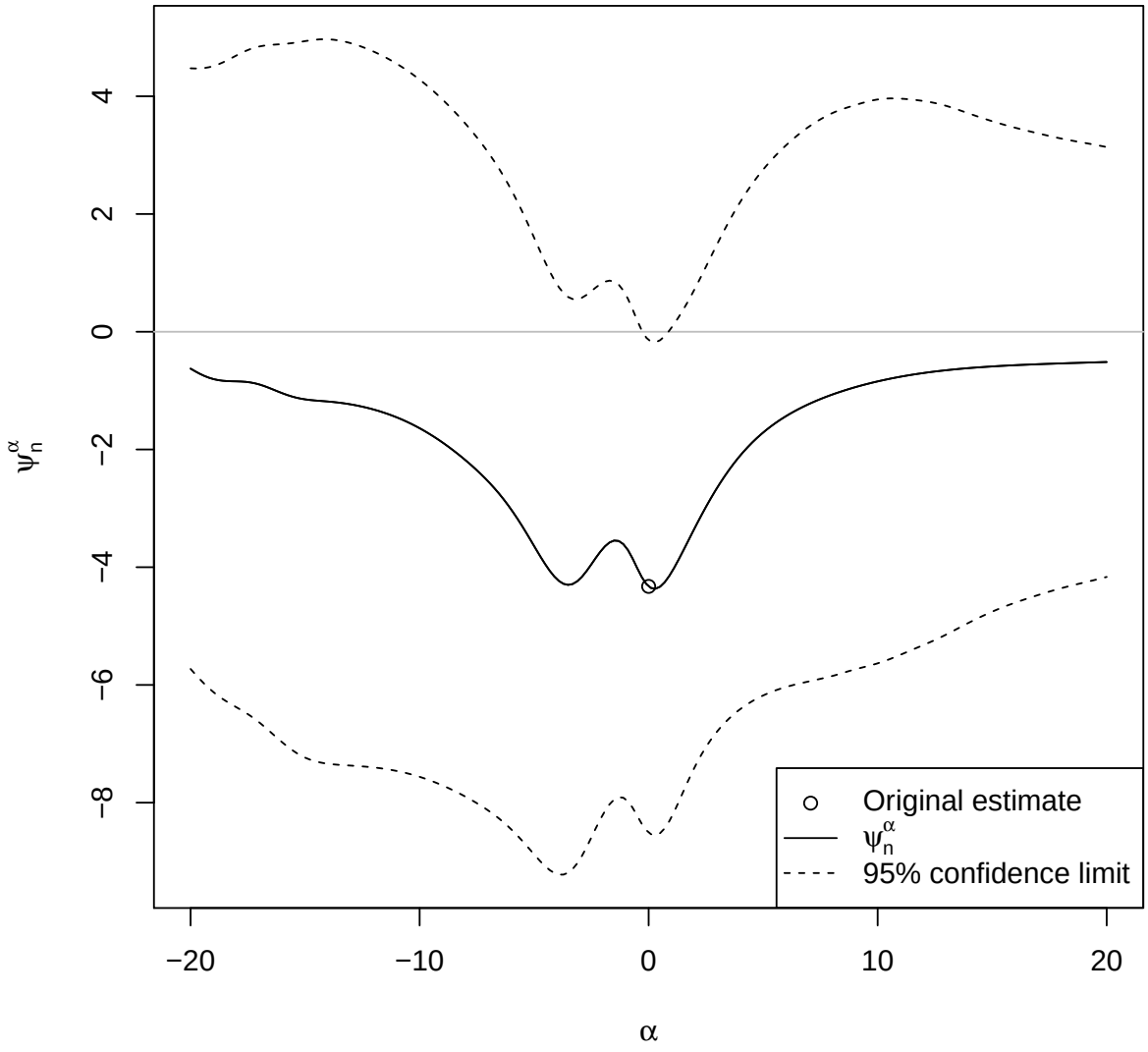

Figure 2.1: Sensitivity analysis of deviation $\bar{z}$

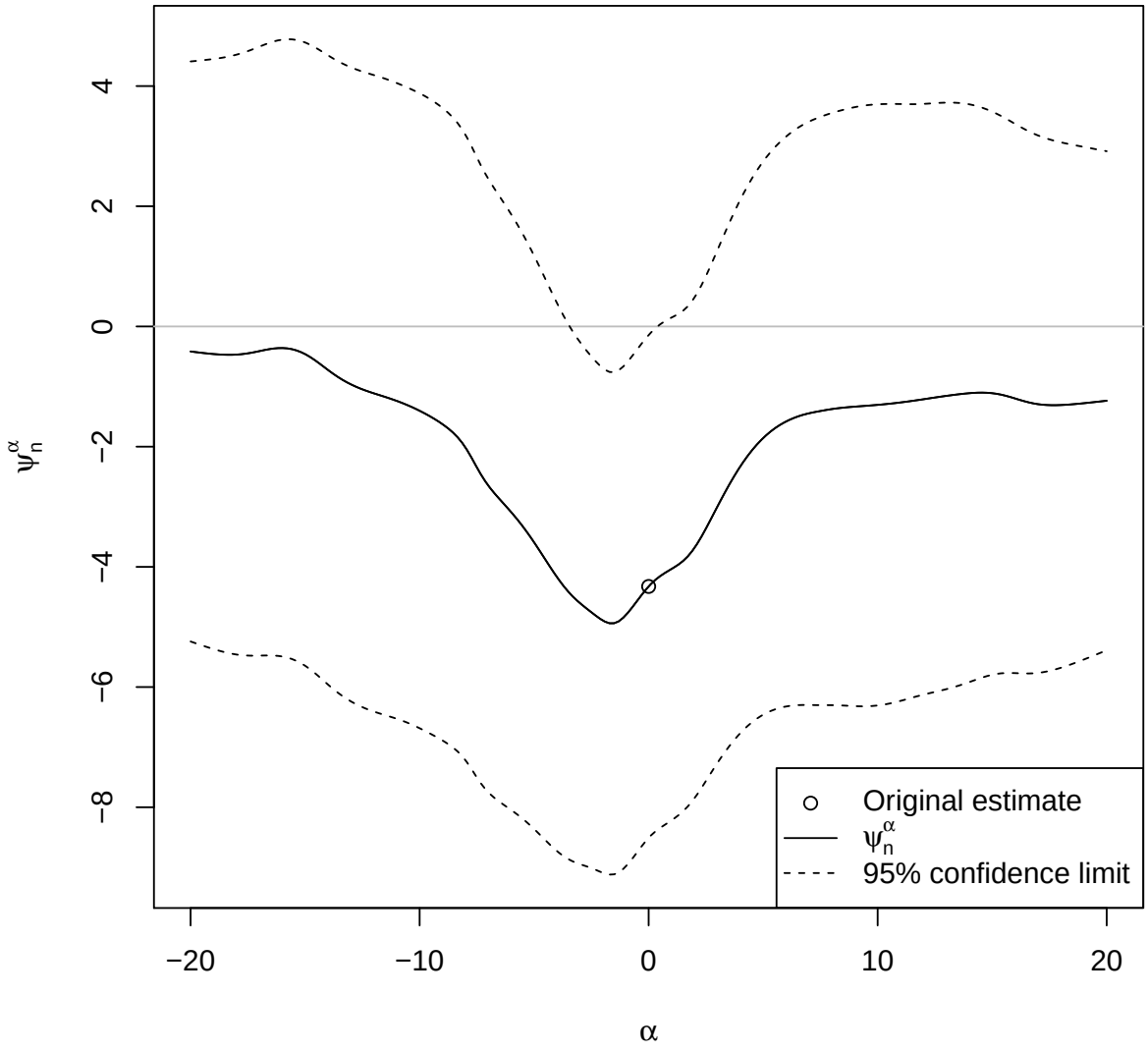

Figure 2.2: Sensitivity analysis of deviation w_c

Chapter 3

Balancing score adjusted targeted minimum loss-based estimation

3.1 Introduction

Estimators based on the propensity score (PS), the probability of receiving a treatment given baseline covariates, are popular for estimation of causal effects such as the average treatment effect (ATE), average treatment effect among the treated (ATT), or the average outcome under treatment. Such methods can be thought of as adjusting for the propensity score in place of baseline covariates, and generally require consistent estimation of the propensity score if it is not known. Common propensity score methods include stratification or subclassification (Austin, 2010; Lunceford and Davidian, 2004; Rosenbaum and D.B. Rubin, 1984), inverse probability of treatment weighting (IPTW) (J M Robins, Hernán, and Brumback, 2000; Rosenbaum, 1987), and propensity score matching (Caliendo and Kopeinig, 2008; Dehejia and Wahba, 2002; Rosenbaum and D.B. Rubin, 1983).

A "balancing score" as defined by Rosenbaum and D.B. Rubin (1983) is a function of baseline covariates such that treatment and baseline covariates are independent conditional on that function. The propensity score is perhaps the most well known example of a balancing score, but balancing scores are more general. Typically, propensity score based methods are said to be consistent when the true propensity score is consistently estimated. Methods that adjust for the propensity score nonparametrically, such as matching or stratification by the propensity score, actually only need that the estimated propensity score converge to some balancing score in order for the parameter of interest to be estimated consistently. However, we are not aware of specific claims in the literature that particular propensity score based methods are consistent under this weaker condition. We say that an estimator using the propensity score or other balancing score has the balancing score property if it is consistent when the estimated propensity score converges to a balancing score.

Though not guaranteed in general, it is possible for an estimated propensity score based on a misspecified model to converge to a balancing score that is not equal to the true propensity

score. Propensity score based estimators that have the balancing score property are robust to this sort of estimator misspecification of the PS, while other propensity score based estimators are not. The balancing score property is desirable because, even though most such estimators were initially developed based on the PS specifically, they inherit this robustness for free. Estimators with the balancing score property are in general not efficient.

An efficient estimator is one that achieves the minimum asymptotic variance of all regular estimators. In many cases, for example when estimating the ATE, ATT, and average outcome under treatment, doubly robust estimators can be constructed. A doubly robust estimator is one that relies on an estimate of both the propensity score and of the outcome regression, the conditional mean of the outcome given baseline covariates and treatment. Doubly robust estimators are consistent if either the estimated propensity score or outcome regression is consistent. Examples include targeted minimum loss-based estimation (TMLE) (Mark J. van der Laan and Sherri Rose, 2011; Mark J. van der Laan and Daniel Rubin, 2006) and augmented inverse probability of treatment weighted estimation (A-IPTW) (James M Robins, Rotnitzky, and Zhao, 1994; M. J. van der Laan and J. M. Robins, 2003). In addition to being doubly robust, both TMLE and A-IPTW are efficient when both the propensity score and outcome regression are consistently estimated.

In this article, we discuss a general class of estimators that have the balancing score property. We also construct a targeted minimum loss-based estimator (TMLE) (Mark J. van der Laan and Sherri Rose, 2011; Mark J. van der Laan and Daniel Rubin, 2006) with the balancing score property. This new TMLE not only has the benefit of the robustness provided by the balancing score property, it also is a locally efficient, doubly robust plug-in estimator. This means that our new estimator retains all of the attractive properties of a traditional TMLE while gaining robustness that other estimators with the balancing score property enjoy when the propensity score only converges to a balancing score.

In Section 3.2, we introduce notation and define the statistical parameter we wish to estimate. In Section 3.3 we describe a TMLE for the statistical parameter. In Section 3.4 we discuss the balancing score property and describe the proposed new TMLE. In Section 3.5 we compare the performance of the new estimator to a traditional TMLE as well as other common estimator and conclude with a discussion in Section 3.6. A list of notation used throughout the article is provided in Section 3.A. Some results and proofs not included in the main text are in Section 3.B and two modifications to the TMLE algorithm are presented in Section 3.C. An example implementation of the proposed new TMLE in R (R Core Team, 2013) is provided in Section 3.D.

3.2 Preliminaries

Consider the random variable $O = (W, A, Y)$ where W is a real valued vector, A is binary with values in $\{0, 1\}$ and Y is univariate real number. Call the probability distribution of O $P_0 \in \mathcal{M}$ where $\mathcal{M}$ is the statistical model. Assume $P_0(A = 1 \mid W) > 0$ for almost every W. This is sometimes called a positivity assumption. Define the parameter mapping Ψ from

$\mathcal{M}$ to $\mathbb{R}$ that maps P to $E_P(E_P(Y \mid A = 1, W))$ where E_P denotes expected value under probability distribution $P \in \mathcal{M}$.

Suppose $A = 1$ indicates some treatment of interest and $A = 0$ represents some control or reference treatment, W represents a vector of baseline covariates measured before treatment, and Y represents some outcome measured after treatment. Then under additional causal assumptions, $\Psi(P_0)$ can be interpreted as a causal quantity. In particular, we may assume that observed treatment A is independent of the counterfactual outcome had each observation received treatment 1 given covariates W. This is known as the randomization assumption or the "no unmeasured confounders" assumption, and the validity depends on the particular application. Under the randomization positivity assumptions, $\Psi(P_0)$ can be interpreted as the average outcome had everyone in the population received treatment 1. In this paper we focus on estimation of the statistical parameter $\Psi(P_0)$, but other similar statistical parameters can, under assumptions, be interpreted as causal parameters such as the ATE or the ATT (Hahn, 1998).

For a probability distribution $P \in \mathcal{M}$, $\bar{Q}(a, w) = E_P(Y \mid A = a, W = w)$ is the regression of the outcome on covariates and treatment. Let $Q_W(w) = P(W = w)$ be the distribution of baseline covariates. The conditional distribution of treatment on baseline covariates is called $g(a \mid w) = P(A = a \mid W = w)$, and define the propensity score as $\bar{g}(w) = g(1 \mid w)$, the probability of treatment given covariates w. The parameter mapping Ψ depends on P only through $Q = (\bar{Q}, Q_W)$, so recognizing the abuse of notation, we sometimes write $\Psi(P) = \Psi(Q) = \Psi(\bar{Q}, Q_W)$.

For a distribution $P \in \mathcal{M}$, we make no assumptions on the outcome regression $\bar{Q}$ or on the distribution Q_W of W. We may put some restriction on possible functions g, for example we may know that $P(A \mid W)$ depends only on a subset of W. The model $\mathcal{M}$ is therefore nonparametric or semiparametric.

Let $O_1, \ldots, O_n$ be a data set of n independent and identically distributed random variables drawn from P_0 where $O_i = (W_i, A_i, Y_i)$. We use the subscript 0 to denote the true probability distribution, and n to denote an estimate based on a dataset of size n, so, for example, E_0 denotes expectation with respect to P_0, $\bar{Q}_0(a, w) = E_0(Y \mid A = a, W = w)$, and $\bar{Q}_n$ is an estimate of $\bar{Q}_0$. Let $\psi_0 = \Psi(P_0)$.

3.3 Targeted minimum loss-based estimation

A plug-in estimator takes an estimate of the distribution P_0, or relevant parts of P_0, and plugs it into the parameter mapping Ψ. In this case, Ψ depends on P through $\bar{Q}$ and Q_W. Using an estimate $\bar{Q}_n$ of $\bar{Q}_0$, and letting Q_{Wn} be the empirical distribution of W, we can

calculate the plug-in estimate as

$$\begin{aligned}\Psi(Q_n) &= \int_w \bar{Q}_n(1, W) dQ_{Wn}(w) \\ &= \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n(1, W_i).\end{aligned}$$

That is, we take the mean of $\bar{Q}_n(1, W)$ with respect to the empirical distribution of W. Plug-in estimators are desirable because they fully utilize known global constraints of Q_0 (by using an estimate Q_n that satisfies these constraints) and guarantee that estimates are in the parameter space, even in small samples. Non-plug-in estimators such as IPTW, can produce estimates outside of the parameter space. For instance if our estimand is a probability, a method like IPTW could yield an estimate outside of $[0, 1]$ when the sample size is small.

Targeted minimum loss-based estimation is a general framework for constructing a plug-in estimator for ψ_0 with additional properties such as efficiency. TMLE takes an initial estimate of the outcome regression $\bar{Q}_0$, say $\bar{Q}_n^0$, and, using an estimate $\bar{g}_n(W)$ of the propensity score, updates it to $\bar{Q}_n^*$. Using the empirical distribution of W along with the updated $\bar{Q}_n^*$, the final estimate is calculated as $\Psi(\bar{Q}_n^*, Q_{Wn})$. The updated $\bar{Q}_n^*$ is constructed in such a way that the final estimate is efficient or attains other properties. We now review some background and a specific implementation of the TMLE procedure for $\Psi(P_0)$.

An estimator that is asymptotically linear can be written as

$$\sqrt{n}(\psi_n - \psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IC(P_0)(O_i) + o_P(1)$$

for some mean zero function $IC(P_0)$ where $o_P(1)$ is a term that converges in probability to 0. The function $IC(P_0)$ is called the influence curve of the estimator at P_0. For an estimator to be efficient, that is, to have the minimum asymptotic variance among all regular estimators, it must be asymptotically linear with influence curve equal to the so called efficient influence curve (Bickel et al., 1993; Mark J. van der Laan and Sherri Rose, 2011). The efficient influence curve for a particular parameter mapping Ψ depends on the model. For our model, regardless of the model for g_0, the efficient influence curve at a $P \in \mathcal{M}$ written in terms of Q and g is

$$D^*(\bar{Q}, Q_W, g)(O) = \frac{A}{g(1 \mid W)}(Y - \bar{Q}(A, W)) + \bar{Q}(1, W) - \Psi(\bar{Q}, Q_W).$$

A derivation of the efficient influence curve is presented in Mark J. van der Laan and Sherri Rose (2011, Chapter 4).

Suppose for now Y is binary or bounded by 0 and 1. A modification to the algorithm and a different TMLE are described in Section 3.C if this is not the case. The initial estimate $\bar{Q}_n^0$ can be obtained via a parametric model for $E_0(Y \mid A, W)$, such as a generalized linear model (McCullagh and Nelder, 1989), or with a data adaptive machine learning algorithm such as the SuperLearner algorithm (Mark J van der Laan, Polley, and Alan E Hubbard,

2007; Mark J. van der Laan and Sherri Rose, 2011), which combines parametric and data adaptive estimators using cross-validation.

The updating step is defined by a choice of loss function L for Q such that $E_0 L(Q)(O)$ is minimized at Q_0, and a working parametric submodel with finite dimensional real valued parameter ϵ, $\{Q(\epsilon) : \epsilon\}$ such that $Q(0) = Q$. The submodel is typically chosen so that the efficient influence curve is in the linear span of the components of the "score" $\frac{d}{d\epsilon} L(Q(\epsilon)(O))$ at $\epsilon = 0$. When L is the negative log likelihood, $\frac{d}{d\epsilon} L(Q(\epsilon)(O))$ is the score in the usual sense. Starting with $k = 0$, the empirical risk minimizer $\epsilon_n^k = \arg\min_\epsilon \sum_{i=1}^n L(Q_n^k(\epsilon))(O_i)$ is calculated and Q_n^k is updated to $Q_n^{k+1} = Q_n^k(\epsilon_n^k)$. The process is iterated until $\epsilon^k \approx 0$, sometimes converging in one step. Details can be found in (Mark J. van der Laan, 2010; Mark J van der Laan, 2010; Mark J. van der Laan and Sherri Rose, 2011; Mark J. van der Laan and Daniel Rubin, 2006).

Define the loss function $L(Q)(O) = L_Y(\bar{Q})(O) + L_W(Q_W)(O)$ where

$$L_Y(\bar{Q})(O) = -Y \log(\bar{Q}(A, W)) - (1 - Y) \log(1 - \bar{Q}(A, W)).$$

and $L_W(Q_W)(O) = -\log(Q_W(W))$. When Y is binary, $L_Y(\bar{Q})(O)$ is the negative conditional log likelihood of the Bernoulli distribution. Because Y is at least bounded by 0 and 1 if not binary, $L_Y(\bar{Q})(O)$ is a valid loss function for the conditional mean. That is, $\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L_Y(\bar{Q})(O)$ (Gruber and Mark J van der Laan, 2010). The function $L_W(Q_W)(O)$ is the negative log likelihood of the distribution of W, and its true mean is minimized by Q_{W0}. Thus, the sum loss function is a valid loss function for $Q_0 = (\bar{Q}_0, Q_{W0})$.

For a working submodel for $\bar{Q}$, we use

$$\bar{Q}(\epsilon)(A, W) = \text{logit}^{-1}\left[\text{logit}(\bar{Q}(A, W)) + \epsilon \frac{A}{g(1 \mid W)}\right]$$

indexed by ϵ. We call this a logistic working model because it is a logistic regression model with offset $\text{logit}\bar{Q}(A, W)$ and single covariate $\frac{A}{g(1|W)}$. The score of this model at $\epsilon = 0$ is

$$\frac{A}{g(1 \mid W)}(Y - \bar{Q}(A, W)).$$

For Q_W, we can use as working submodel

$$Q_W(\epsilon')(W) = \{1 + \epsilon'[\bar{Q}(1, W) - \Psi(Q)]\} Q_W(W)$$

which has score $\bar{Q}(1, W) - \Psi(\bar{Q}, Q_W)$ at $\epsilon' = 0$. We can see that the efficient influence curve $D^*(P_0)$ can be written as a linear combination of the scores of these submodels when $Q = Q_0$ and $g = g_0$.

The estimate ϵ_n^0 can be calculated using standard logistic regression software with $\text{logit}(\bar{Q}_n^0(A, W))$ as a fixed offset term, and $\frac{A}{g_n(1|W)}$ as a covariate. By using the empirical distribution of W as an initial estimate for Q_{Wn}^0, and negative log likelihood loss function

for L_W, the empirical risk is already minimized at Q^0_{Wn}, so $\epsilon'^0_n = 0$ and no update is needed. In this case, the algorithm converges in one step, because $\frac{A}{g_n(1|W)}$ is not updated between iterations, so an additional update to $\bar{Q}^1_n$ will yield $\epsilon^1_n = 0$. The estimate $\bar{Q}^*_n = \bar{Q}^0_n(\epsilon^0_n)$ and the TMLE estimate of $\Psi(P_0)$ is calculated as

$$\Psi(\bar{Q}^*_n, Q_{Wn}) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}^*_n(1, W_i).$$

Under regularity conditions, the TMLE is asymptotically linear and doubly robust, meaning that if the initial estimate $\bar{Q}^0_n$ is consistent for $\bar{Q}_0$, or $\bar{g}_n$ is consistent for $\bar{g}_0$, then $\Psi(\bar{Q}^*_n, Q_{Wn})$ is consistent for $\Psi(P_0)$. Additionally, when both $\bar{Q}^0_n$ and g_n are consistent, the influence curve of the TMLE is equal to the efficient influence curve, so the estimator achieves the semiparametric efficiency bound. Precise regularity conditions for asymptotic linearity and efficiency are presented in Section 3.B in Theorem 5.

3.4 Balancing score property and proposed estimator

A function b of W is called a balancing score if $A \perp W \mid b(W)$ (Rosenbaum and D.B. Rubin, 1983). Trivially, $b(W) = W$ is a balancing score, and by definition of the propensity score, $\bar{g}_0(W)$, is a balancing score. In general, any function $b(W)$ is a balancing score if and only if there exists some function f such that $\bar{g}_0(W) = f(b(W))$ (Rosenbaum and D.B. Rubin, 1983, Theorem 2). For example any monotone transformation of the propensity is a balancing score. Such a function is called a "balancing score" because, conditional on $b(W)$, the distribution of W between the treated and untreated observations is equal or balanced. That is, $P_0(W \mid A = 1, b(W)) = P_0(W \mid A = 0, b(W))$. Rosenbaum and D.B. Rubin (1983) show that adjusting for a balancing score yields the same estimand as adjusting for the full set of covariates W which we state in Lemma 1 and offer a different proof in Section 3.B.

Lemma 1. *If $b(W)$ is a balancing score under distribution P, then $E_P(E_P(Y|A = 1, b(W))) = \Psi(P)$.*

This result gives rise to methods for estimating $\Psi(P_0)$ based on a balancing score and not on an estimate of $\bar{Q}_0$. The propensity score is the balancing score most commonly used for estimating $\Psi(P_0)$, and frequently used estimators include propensity score matching, stratification, and IPTW. When the propensity score is not known, these estimators rely on an estimated propensity score $\bar{g}_n$, and, under regularity conditions, are consistent when $\bar{g}_n$ is consistent for $\bar{g}_0$. The IPTW estimator, in particular, requires that $\bar{g}_n$ converges to $\bar{g}_0$ for consistency. However, many of these methods, such as propensity score matching and stratification by the propensity score, can be seen as nonparametrically adjusting for the propensity score and only rely on the propensity score being a balancing score. For these estimators, it is sufficient for $\bar{g}_n$ to converge to some balancing score under P_0. We call this property the balancing score property.

In practice, an estimator $\bar{g}_n$ can approximate a balancing score well but not converge to the true propensity score. A parametric logistic regression estimator will estimate some function of the covariates that is a projection of $\bar{g}_0$ onto the model determined by the parametrization of the estimator. If the parametric estimator is correctly specified, this projection will be $\bar{g}_0$. Depending on the true $\bar{g}_0$ and distribution of covariates, it is possible for this projection to be a balancing score or at least approximate some balancing score when the estimator is not correctly specified. For example, suppose the true $\bar{g}_0$ depends on higher order interactions of covariates. Though not the case in general, in some settings a main terms logistic regression may approximate a balancing score well. We explore such a setting via simulation in Section 3.5. In another example, suppose $\bar{g}_0$ depends on covariates in an additive on the logit scale but not necessarily linear or even smooth way. A logistic regression estimator with linear or possibly higher order polynomial main terms may again approximate some balancing score.

Estimators based only on the propensity score are not doubly robust. We now construct a locally efficient doubly robust estimator with the balancing score property. We start with initial estimators $\bar{Q}_n$ for $\bar{Q}_0$ and $\bar{g}_n$ for $\bar{g}_0$. We then update $\bar{Q}_n$ by nonparametrically regressing Y on A and $\bar{g}_n(W)$ using $\bar{Q}_n(A, W)$ as an offset. Similarly to the TMLE procedure in Section 3.3, we use this updated estimate of $\bar{Q}_0$ to estimate ψ_0 by plugging it in to the parameter mapping Ψ along with the empirical distribution of W.

To update $\bar{Q}_n$ by further adjusting for A and $\bar{g}_n$, we specify a working model and loss function pair. The working model and loss function pair is somewhat analogous to that in the updating step in the TMLE procedure described in Section 3.3. The loss function can be the same as that in the TMLE procedure's updating step, but it need not be. Define $\bar{Q}$ and b to be the limits of $\bar{Q}_n$ and $\bar{g}_n$, respectively, as $n \to \infty$. Let Θ be the class of all functions of A and $b(W)$, and let θ be some function in that class. Here $\bar{Q}$ is not necessarily $\bar{Q}_0$ and b is not necessarily $\bar{g}_0$ or even a balancing score. For concreteness, consider two working model and loss function pairs: a logistic working model

$$\bar{Q}^{b,\theta}(A, W) = \text{logit}^{-1}[\text{logit}(\bar{Q}(A, W)) + \theta(A, b(W))] \tag{3.1}$$

with loss function

$$L'(\bar{Q}^{b,\theta})(O) = -Y \log(\bar{Q}^{b,\theta}(A, W)) - (1 - Y) \log(1 - \bar{Q}^{b,\theta}(A, W)),$$

which is the negative log likelihood loss when Y is binary, and a linear working model

$$\bar{Q}^{b,\theta}(A, W) = \bar{Q}(A, W) + \theta(A, b(W)) \tag{3.2}$$

with loss function

$$L'(\bar{Q}^{b,\theta})(O) = (Y - \bar{Q}^{b,\theta}(A, W))^2,$$

the squared error loss. In both working models, we leave the function θ unspecified. We can view a working model used for the updating step in the TMLE procedure as a special case of

the working model here by restricting θ to have the form

$$\theta(A, b(W)) = \epsilon \frac{A}{b(W)}$$

where ϵ is real, using notation $b(W)$ in place of $g(1 \mid W)$ as used in Section 3.3.

Define

$$\theta_0 = \arg\min_{\theta \in \Theta} E_0 L'(\bar{Q}^{b,\theta})(O).$$

Given $\bar{Q}$, the limit of some estimate for $\bar{Q}_0$, one can think of θ_0, a function of A and $b(W)$, as the residual bias between $E_0(\bar{Q}(A, W) \mid A, b(W))$ and $E_0(Y \mid A, b(W))$ on either the logistic or linear scale. When the initial estimator $\bar{Q}_n$ is consistent, so $\bar{Q} = \bar{Q}_0$, $\theta_0(A, b(W))$ will be 0, because $\bar{Q}$ will already be fully adjusting for A and $b(W)$.

Suppose for now that we have an estimate of θ_0 which we call θ_n. We return to the problem of estimating θ_0 later in this section. Calculate the update of $\bar{Q}_n$ as $\bar{Q}_n^{\bar{g}_n,\theta_n}$ and using this updated regression, a final estimate of ψ_0 is calculated as $\Psi(\bar{Q}_n^{\bar{g}_n,\theta_n}, Q_{Wn})$, which we call a doubly robust balancing score adjusted (DR-BSA) plug-in estimator. In Theorem 3 in Section 3.B, we show that the DR-BSA estimator doubly robust in the sense that it is consistent when either $\bar{Q} = \bar{Q}_0$ or θ_n consistently estimates θ_0 and b is a balancing score.

When initial estimator $\bar{Q}_n$ does not consistently estimate $\bar{Q}_0$, consistency of the DR-BSA estimate requires that b is a balancing score and θ_0 is consistently estimated. To weaken this requirement, we now construct a TMLE with the balancing score property by using $\bar{Q}_n^0 = \bar{Q}_n^{g_n,\theta_n}$ as the initial estimate in the TMLE procedure in Section 3.3 and updating it to $\bar{Q}_n^*$. The TMLE of $\Psi(P_0)$ is calculated as $\Psi(\bar{Q}_n^*, Q_{Wn})$. We call this a balancing score adjusted TMLE (BSA-TMLE). In Theorem 4 in Section 3.B, we show that the BSA-TMLE is consistent if any of the three conditions hold: (1) $\bar{Q} = \bar{Q}_0$, (2) $b = \bar{g}_0$, or (3) b is a balancing score and θ_n consistently estimates θ_0. The BSA-TMLE is therefore doubly robust in the usual sense and also has the balancing score property. The BSA-TMLE is a TMLE as described in Section 3.3 where in addition to attempting to adjust for W, the initial estimator $\bar{Q}_n^0$ is making an extra attempt to adjust for a balancing score. If θ_0 is consistently estimated, then like the standard TMLE, when both the initial estimates of $\bar{Q}_0$ and g_0 are consistent, the influence curve of the BSA-TMLE is the efficient influence curve. Therefore, under regularity conditions, the BSA-TMLE is locally efficient and keeps all of the attractive properties of TMLE while also having the balancing score property.

We now return to the problem of estimating θ_0. The working model in the definition of θ_0 depends is $\bar{Q}^{b,\theta}$ which depends on limits $\bar{Q}$ and b. To estimate θ_n, we use $\bar{Q}_n^{\bar{g}_n,\theta}$ as the working model. If $\bar{g}_n(W)$ is discrete and θ_0 is estimated in a saturated parametric model, $\Psi(\bar{Q}_n^{\bar{g}_n,\theta_n}, Q_{Wn})$ is exactly a TMLE as proved in Lemma 2 in Section 3.B. When $\bar{g}_n(W)$ is not discrete, it can be discretized into k categories based on quantiles. The parameter θ_0 can be estimated with a saturated parametric model with standard logistic regression software with dummy variables for each stratum and treatment combination, and logit$\bar{Q}_n(A, W)$ as an offset. When $\bar{Q}_n(A, W)$ is unadjusted for W, for example $\bar{Q}_n$ is estimated in a GLM with only an intercept and treatment as a main term, this reduces to usual propensity score

stratification. In general, when the number of categories k is fixed and does not grow with sample size, stratification is not consistent, though one hopes that the residual bias is small (Lunceford and Davidian, 2004). If k is too large, there is a possibility of all observations in a particular stratum having the same value for A, in which case $\theta_n(A, W)$ is not well defined. In many applications, the number of strata is often set based on the rule of thumb $k = 5$ recommended by Rosenbaum and D.B. Rubin (1984). Though the stratification estimator of ψ_0 is not root-n consistent when k is fixed, the BSA-TMLE removes this remaining bias if g_n consistently estimates the true propensity score while preserving the balancing score property. In practice, the number of strata k can be chosen based on cross-validation in such a way that it can grow with sample size.

Alternatively, when $\bar{g}_n(W)$ is not discrete or has many levels, θ_0 can be estimated in an generalized additive model(Simon N. Wood, 2011) with $\bar{Q}_n$ as an offset. We can parameterize this model as

$$\begin{aligned}\bar{Q}_n^{\bar{g}_n,\theta}(A, W) = \text{logit}^{-1}[&\text{logit}(\bar{Q}_n(A, W))+ \\ &A\theta_1(\bar{g}_n(W)) + (1 - A)\theta_2(\bar{g}_n(W))]\end{aligned} \tag{3.3}$$

with $\theta = (\theta_1, \theta_2)$ where θ_1 and θ_2 are unspecified. Other parametric or nonparametric methods can be used and cross-validation based SuperLearning can be used to select the best weighted combination of estimators for θ_0 (Mark J van der Laan, Polley, and Alan E Hubbard, 2007; Mark J. van der Laan and Sherri Rose, 2011). When the linear model (3.2) is used, $\theta_0(A, W) = E_0(Y - \bar{Q}(A, W) \mid A, g_n(1 \mid W))$. In this case, a nearest neighbor or kernel regression can be used where residuals from the initial estimate, $R_i = Y_i - \bar{Q}_n(A_i, W_i)$, are treated as an outcome. This is similar to the bias corrected matching estimator presented by Abadie and Imbens (2011).

3.5 Simulations

We demonstrate properties of the proposed BSA-TMLE in various scenarios, and compare it to other estimators. The estimators compared in simulations include a plug-in estimator based on just the initial estimator of $\bar{Q}_0$ without balancing score adjustment, DR-BSA plug-in estimators without a TMLE update, non-doubly robust BSA plug-in estimators, an inverse probability of treatment weighted estimator, and a TMLE using an initial estimator for $\bar{Q}_0$ not directly adjusted for a balancing score.

The plug-in estimator not adjusted for a balancing score is calculated as $\Psi(\bar{Q}_n, Q_{W_n})$ with $\bar{Q}_n$ as defined in Section 3.4. We call this the simple plug-in estimator. The DR-BSA plug-in estimator uses the balancing score adjusted $\bar{Q}_n^0$ as in Section 3.4 and is calculated as $\Psi(\bar{Q}_n^0, Q_{W_n})$. The non-doubly robust BSA plug-in estimator adjusts for the balancing score, but uses as initial $\bar{Q}_n$ an unadjusted estimate that is not a function of W. The non-DR-BSA plug-in estimator can be thought of as only adjusting for $g_n(1 \mid W)$ and not the whole

Table 3.1: Summary of properties of compared estimators

Estimator	Plug-in	Consistent if			Efficient if
		$Q_n \to Q_0$	$\bar{g}_n \to \bar{g}_0$	$\bar{g}_n \to$ BS	$\bar{Q}_n \to \bar{Q}_0$ & $\bar{g}_n \to \bar{g}_0$
Simple plug-in	✓	✓			
BSA	✓		✓	✓	
DR-BSA	✓	✓	✓	✓	✓†
IPTW			✓		
TMLE	✓	✓	✓		✓
BSA-TMLE	✓	✓	✓	✓	✓

†We do now show formally that the DR-BSA estimator is asymptotically linear.

covariate vector W. The IPTW estimator is calculated as

$$n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{g_n(1 \mid W_i)}.$$

The estimators we compare are summarized in Table 3.1.

In the simulation studies, we use two methods for adjusting the initial estimator with the propensity score. All simulations were conducted in R (). The initial estimator $\bar{Q}_n$ was adjusted with either a generalized additive model (GAM) in (3.3), or a nearest neighbor approach analogous to propensity score matching. The non-DR-BSA plug-in estimator based on nearest neighbors reduces exactly to a propensity score matching estimator. The GAM was fitted with the `mgcv` package (Simon N. Wood, 2011) and the nearest neighbor/propensity score matching type estimator was implement with the `Matching` package (Sekhon, 2011).

The initial estimates for $\bar{Q}_0$ and $\bar{g}_0$ are estimated using generalized linear models. Specifically, $\bar{g}_0$ is estimated using logistic regression, and $\bar{Q}_0$ is estimated with least squares when Y is continuous, and logistic regression when Y is binary. To investigate robustness to various kinds of model misspecification, models are either correctly specified, or some relevant covariates are excluded.

The data generating distribution in the simulations was as follows. Baseline covariates W_1, W_2 and W_3 have independent uniform distributions on $[0, 1]$. Treatment A is Bernoulli with mean

$$\text{logit}^{-1}(\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_1 W_2).$$

Outcome Y is either Bernoulli or normal with variance 1 and mean

$$m(\alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 A),$$

where m is logit^{-1} if Y is Bernoulli, or the identity if Y is normal. All estimators were evaluated on $1,000$ datasets of size $n = 100$ and $n = 1,000$. Bias, variance, and mean squared error (MSE) are calculated for each estimator.

Table 3.2: Simulation results for distribution one with $\bar{Q}_n$ unadjusted and $\bar{g}_n$ correctly specified but transformed with Beta CDF

Estimator	**n=100**			**n=1000**		
	Bias	Variance	MSE	Bias	Variance	MSE
BSA, NN	0.0276	0.0180	0.0188	0.0026	0.0018	0.0018
BSA, GAM	0.0075	0.0163	0.0163	0.0041	0.0015	0.0015
IPTW	−0.0249	0.0087	0.0093	−0.0246	0.0010	0.0016
TMLE	0.1063	0.0111	0.0224	0.1082	0.0010	0.0127
BSA-TMLE, NN	0.0276	0.0180	0.0188	0.0026	0.0018	0.0018
BSA-TMLE, GAM	0.0070	0.0164	0.0165	0.0037	0.0015	0.0015

In the first scenario, which we call distribution one, $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (-3, 2, 2, 0.5)$ and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (-3, 1, 1, 0, 5)$ so W_1 and W_2 are confounders, and the propensity score depends on the product W_1W_2. The true parameter $\psi_0 \approx 0.0985$ and the variance bound is approximately $1.5691/n$. The variance bound of a parameter in a semiparametric model is the minimum asymptotic variance that a regular estimator can achieve, and depends on the parameter mapping Ψ and the true distribution P_0 (Bickel et al., 1993). This is analogous with the Cramér-Rao bound in a parametric model. An estimator that asymptotically achieves the variance bound is called efficient.

The first set of results in Table 3.2 demonstrate the balancing score property. The initial estimate $\bar{Q}_n$ is unadjusted. A correct logistic regression model is specified for $\bar{g}_0$, but predictions are transformed by the Beta cumulative distribution function with both shape parameters equal to 2. Although artificial, this means that $\bar{g}_n$ converges to a monotone transformation of $\bar{g}_0$, which is a balancing score, but does not converge to the true $\bar{g}_0$. We can see that the TMLE not adjusted for the propensity score and the IPTW estimators are not consistent as the bias is not decrease substantially when sample size increase. Conversely, methods where the initially estimate $\bar{Q}_n$ is adjusted with the propensity score, are consistent, as bias is decreasing quickly with sample size.

Table 3.3 shows similar performance in a more realistic scenario. In this setting, the initial estimator for $\bar{Q}_n$ is unadjusted, but the logistic regression model for the propensity score is misspecified by excluding the interaction term W_1W_2. Here predictions are not transformed. Here $\bar{g}_n$ is close to but not exactly a balancing score, but it is close enough that the bias in estimators that nonparametrically adjust for $\bar{g}_n$ is small. The IPTW estimator, however, is still biased at large n because $\bar{g}_n$ is not converging to $\bar{g}_0$. In this case TMLE performs well even with an unadjusted initial estimator but this is not guaranteed when $\bar{g}_n$ is misspecified.

Table 3.4 examines the performance of estimators when the model for $\bar{g}_0$ is misspecified, (only including W_1 in the logistic regression model,) but the initial estimate $\bar{Q}_n$ is a correctly specified model. Here we see that estimates that rely only on estimated propensity score, (the non-doubly robust BSA estimators and IPTW,) fail to be consistent, but estimates that

Table 3.3: Simulation results for distribution one with $\bar{Q}_n$ unadjusted, and $\bar{g}_n$ misspecified but close to a balancing score

Estimator	**n=100**			**n=1000**		
	Bias	Variance	MSE	Bias	Variance	MSE
BSA, NN	0.0311	0.0166	0.0176	0.0027	0.0016	0.0016
BSA, GAM	0.0147	0.0159	0.0161	0.0033	0.0014	0.0014
IPTW	0.0390	0.0410	0.0425	0.0357	0.0025	0.0037
TMLE	0.0096	0.0172	0.0173	0.0098	0.0016	0.0017
BSA-TMLE, NN	0.0311	0.0166	0.0176	0.0027	0.0016	0.0016
BSA-TMLE, GAM	0.0101	0.0189	0.0190	−0.0042	0.0015	0.0016

Table 3.4: Simulation results for distribution one with $\bar{Q}_n$ correctly specified and $\bar{g}_n$ misspecified

Estimator	**n=100**			**n=1000**		
	Bias	Variance	MSE	Bias	Variance	MSE
Simple plug-in	0.0071	0.0120	0.0120	0.0011	0.0013	0.0013
BSA, NN	0.1190	0.0126	0.0268	0.1064	0.0014	0.0128
DR-BSA, NN	0.0064	0.0139	0.0140	0.0003	0.0015	0.0015
BSA, GAM	0.1139	0.0116	0.0246	0.1096	0.0012	0.0133
DR-BSA, GAM	0.0152	0.0129	0.0132	0.0015	0.0013	0.0013
IPTW	0.1061	0.0115	0.0228	0.1035	0.0012	0.0119
TMLE	0.0076	0.0129	0.0130	0.0009	0.0013	0.0013
BSA-TMLE, NN	0.0064	0.0139	0.0140	0.0003	0.0015	0.0015
BSA-TMLE, GAM	0.0154	0.0133	0.0136	0.0014	0.0013	0.0013

use the correctly specified initial estimate of $\bar{Q}_0$, are consistent. Importantly, even when the initial estimate is adjusted with the completely misspecified $\bar{g}_n$, final estimates are still consistent when the initial $\bar{Q}_n$ is correctly specified.

In a second scenario, called distribution two, Y is conditionally normal with $\alpha = (0, 10, 8, 0, 2)$ and $\beta = (-1, 0, 0, 3, 0)$. Here Y depends on W_1 and W_2 but A does not, so they are not confounders. Additionally, A depends on W_3, but Y does not, so W_3 is an instrumental variable. In this setting, because none of the baseline covariates are confounders, an unadjusted estimator of ψ_0 will be consistent but not efficient, because it will fail to take into account the relationship with the non-confounding baseline covariates W_1 and W_2. Here, the true ψ_0 is 2 and the variance bound is approximately $5.1979/n$.

Table 3.5 shows results from distribution two where the initial estimate for $\bar{Q}_0$ is the least squares estimate from a linear regression model with A, W_1, W_2, and W_3 are main terms,

Table 3.5: Simulation results from distribution two with $\bar{Q}_n$ correctly specified and $\bar{g}_n$ correctly specified and includes an instrumental variable

Estimator	**n=100**			**n=1000**		
	Bias	Variance	MSE	Bias	Variance	MSE
Simple plug-in	−0.0112	0.0505	0.0506	0.0007	0.0048	0.0048
BSA, NN	0.0080	0.1815	0.1815	0.0020	0.0185	0.0185
DR-BSA, NN	−0.0108	0.0578	0.0579	0.0024	0.0059	0.0060
BSA, GAM	−0.0061	0.3207	0.3208	−0.0008	0.0097	0.0097
DR-BSA, GAM	−0.0112	0.0565	0.0566	0.0010	0.0051	0.0051
IPTW	−0.0072	0.7559	0.7560	−0.0021	0.0231	0.0231
TMLE	−0.0182	0.0575	0.0578	0.0009	0.0052	0.0052
BSA-TMLE, NN	−0.0108	0.0578	0.0579	0.0024	0.0059	0.0060
BSA-TMLE, GAM	−0.0181	0.0587	0.0590	0.0009	0.0053	0.0053

and the initial estimate for the propensity score is the MLE from a logistic regression model with main terms W_1, W_2, and W_3. Here we see that, although all estimators have low bias, those that only adjust for $\bar{g}_n$, (the non-doubly robust BSA estimators and IPTW,) have much higher variance than those with a correctly specified initial estimate. This demonstrates the importance in terms of efficiency of attempting to estimate $\bar{Q}_0$ well with the initial estimate even when confounding is not a concern.

3.6 Discussion

In this paper we discuss the balancing score property of estimators that nonparametrically adjust for the propensity score. We see in simulations that, even when the propensity score estimator is not consistent, $\Psi(P_0)$ can be estimated with low bias if the estimate of the propensity score approximates a balancing score well enough. Additionally, we introduce a balancing score adjusted TMLE which has the balancing score property and is also doubly robust and locally efficient, and provide regularity conditions for asymptotic linearity in Section 3.B.

In order for an estimator to have the balancing score property, we need to estimate some balancing score. We acknowledge that in practice, one does not expect an estimate of the propensity score to converge exactly to a balancing score that is not g_0 in general. However, because the propensity score is a single element of the large class of balancing scores, the condition that an estimated propensity score g_n converges to some balancing score is strictly weaker than requiring g_n to converge to g_0. When g_n fails to converge to g_0, we may still have a chance at approximating a balancing score, and the proposed BSA-TMLE can still reduce bias relative to an estimator that requires that g_n converges to g_0 without sacrificing double robustness or efficiency.

We now discuss some some possible generalizations to the work in this paper and areas for further research. The estimators present in this paper are for the statistical parameter $E_0[E_0(Y \mid A = 1, W)]$, which, under assumptions, can be interpreted as the population mean of a variable Y when Y is subject to missingness (Kang and Schafer, 2007). The results and similar estimators are immediately applicable to other interesting statistical parameters such as

$$E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$$

and

$$E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 1]$$

which, under non-testable causal assumptions, can be interpreted as causal parameters called the ATE or ATT, respectively (Hahn, 1998; Mark J. van der Laan and Sherri Rose, 2011). Additionally, the results are immediately generalizable to the estimation of parameters in marginal structural models (James M. Robins, 1997; Rosenblum and Mark J van der Laan, 2010).

Propensity score based methods are most often applied in settings where the treatment variable is binary. In settings where the treatment variable is not binary, Kosuke Imai and Van Dyk (2004) generalize the notion of the propensity score to the propensity function, the conditional probability of observed treatment given covariates. Kosuke Imai and Van Dyk (2004) show that the propensity function is a balancing score. When the propensity function can be characterized by a finite dimensional parameter, one can estimate parameters of the distribution of counterfactuals by adjusting for the dimensional characterization of the propensity function in place of all covariates. Using the approach of Kosuke Imai and Van Dyk (2004), the methods in this paper may be extended to develop estimators that are doubly robust and efficient with the balancing score property for more general situations where treatment is categorical or potentially even continuous.

Traditionally, propensity score based estimators estimate the propensity score based on how well $\bar{g}_n$ approximates the true $\bar{g}_0$. Collaborative targeted minimum loss-based estimation (CTMLE) is a method that chooses an estimator for the propensity score based on how well it helps reduce bias in the estimation of $\Psi(P_0)$ in collaboration with an initial estimate of $\bar{Q}_0$ using cross-validation (Mark J van der Laan and Gruber, 2010; Mark J. van der Laan and Sherri Rose, 2011). In doing so, CTMLE attempts to adjust the propensity score for the most important confounders first, and avoid adjustment for instrumental variables. This can lead to improvements in efficiency and robustness to violations of the assumption $P_0(A = a|W) > 0$. Applying an analogous techniques of estimator selection for balancing score adjusted estimators is an area of further research.

3.A Notation

- $O = (W, A, Y)$: observed data structure
  - W: vector of covariates

  - A: treatment indicator, 0 or 1
  - Y: univariate outcome
- P: a distribution of O
- $\mathcal{M}$: statistical model, set of possible probability distributions P
- $E_p(\cdot)$: expectation under distribution P
- $Q = (\bar{Q}, Q_W)$
  - $\bar{Q}(a, w) = E_P(Y \mid A = a, W = w)$
  - $Q_W(w) = P(W = w)$
- $g(a \mid w) = P(A = a \mid W = w)$
- $\bar{g}(w) = g(1 \mid W)$, also called the propensity score when .
- Ψ: statistical parameter mapping from $\mathcal{M}$ to $\mathbb{R}$.
  - In particular, $\Psi(P) = E_P[E_P(Y \mid A = 1, W)]$
  - Also written as $\Psi(Q)$
- $\psi = \Psi(P)$
- Subscript 0: indicates the truth, e.g. $\psi_0 = \Psi(P_0)$ is the true parameter value
- Subscript n: indicates an estimate based on n observations, e.g. $\bar{Q}_n$ is an estimate of $\bar{Q}_0$
- $\bar{Q}_n^0$ an initial estimate of $\bar{Q}_0$
- L: loss function
- L_Y: loss function for $\bar{Q}$
- L_W: loss function for Q_W
- $Q(\epsilon)$ a working submodel through Q
- IC: an influence curve
- D^*: the efficient influence curve
- $\bar{Q}_n^*$ a TMLE updated estimate of some initial $\bar{Q}_n^0$
- $b(w)$: some function of w that is a potential balancing score

- θ: some function of a and $b(w)$
- $\bar{Q}^{b,\theta}$: a working submodel through $\bar{Q}$ for a particular b and θ
- L' a loss function for $\bar{Q}^{b,\theta}$, used in Section 3.4

3.B Some results and proofs

Proof of Lemma 1. In this proof, E means expectation with respect to P. First note that $E(Y \mid A = 1, W, b(W)) = E(Y \mid A = 1, W)$ because b is a function of only W. Next,

$$E[E(Y \mid A = 1, W) \mid A = 1, b(W)] = E[E(Y \mid A = 1, W) \mid b(W)]$$

because the inner conditional expectation is a function of only W and $W \perp A \mid b(W)$ when b is a balancing score. Thus,

$$\begin{aligned} E[E(Y \mid A = 1, b(W))] =& E\{E[E(Y \mid A = 1, W, b(W)) \mid A = 1, b(W)]\} \\ =& E\{E[E(Y \mid A = 1, W) \mid A = 1, b(W)]\} \\ =& E\{E[E(Y \mid A = 1, W) \mid b(W)]\} \\ =& E[E(Y \mid A = 1, W)] \\ =& \Psi(P) \end{aligned}$$

□

Theorem 3. *Assume*

$$\Psi((\bar{Q}_n^{g_n,\theta_n}, Q_{Wn})) - \Psi((\bar{Q}^{b,\theta_0}, Q_{W0})) \to 0, \text{ as } n \to \infty.$$

In addition, assume that either $\bar{g}$ is a balancing score or $\bar{Q} = \bar{Q}_0$. Then $\Psi((\bar{Q}_n^{g_n,\theta_n}, Q_{Wn}))$ is consistent for ψ_0.

Proof. By definition of θ_0, we have

$$E_0[h(A, b(W))(Y - \bar{Q}^{b,\theta_0}(A, W))] = 0$$

for all functions h of A and $b(W)$. Rosenbaum and D.B. Rubin (1983, Theorem 2) show that b is a balancing score if and only if there exists a function f so that $\bar{g}_0(w) = f(b(w))$ a.e., so we can select the function

$$h(A, b(W)) = \frac{A}{f(b(W))} = \frac{A}{\bar{g}_0(W)}.$$

In addition, we also have that $E_0\bar{Q}^{b,\theta_0}(1, W) - \Psi((\bar{Q}^{b,\theta_0}, Q_{W,0})) = 0$. This proves that

$$P_0 D^*(\bar{Q}^{b,\theta_0}, Q_{W,0}, g_0) = 0,$$

where D^* is the efficient influence curve of Ψ at P, and notation

$$P\phi = \int \phi(o)dP(o)$$

for some function ϕ of O and distribution P. Since $P_0D^*(\bar{Q}, Q_W, g_0) = \psi_0 - \Psi(Q)$, this shows

$$\Psi((\bar{Q}^{b,\theta_0}, Q_{W0})) = \Psi((\bar{Q}_0, Q_{W0}))$$

This proves that under the stated consistency condition, we indeed have that $\Psi((\bar{Q}_n^{g_n,\theta_n}, Q_{Wn}))$ is consistent for ψ_0. This proves the consistency under the condition that b is a balancing score.

Consider now the case that $\bar{Q} = \bar{Q}_0$. Then $\theta_0 = 0$ and thus $\bar{Q}^{b,\theta_0} = \bar{Q}_0$. Thus, the limit $\Psi((\bar{Q}^{b,\theta_0}, Q_{W0})) = \Psi((\bar{Q}_0, Q_{W0}))$, which proves the second claim of the theorem. □

Theorem 4. *Assume*

$$\Psi((\bar{Q}_n^{g_n,\theta_n}(\epsilon_n), Q_{Wn})) - \Psi((\bar{Q}^{b,\theta_0}(\epsilon_0), Q_{W0})) \to 0, \text{ as } n \to \infty,$$

where $\epsilon_0 = \arg\min_\epsilon P_0L(\bar{Q}^{b,\theta_0}(\epsilon))$.

In addition, assume that b is a balancing score, or $\bar{Q} = \bar{Q}_0$. *Then* $\epsilon_0 = 0$ *and* $\Psi((\bar{Q}_n^{\bar{g}_n,\theta_n}(\epsilon_n), Q_{Wn}))$ *is consistent for* ψ_0.

Proof. Firstly, assume b is a balancing score so by Rosenbaum and D.B. Rubin (1983, Theorem 2) there exists a mapping f so that $g_0(w) = f(b(w))$ a.e.. In the proof of the previous theorem we showed that

$$E_0\frac{A}{b(W)}(Y - \bar{Q}^{b,\theta_0}(A,W)) = E_0\frac{A}{g_0(W)}(Y - \bar{Q}^{b,\theta_0}(A,W)) = 0.$$

The left-hand side equals $\frac{d}{d\epsilon}P_0L(\bar{Q}^{b,\theta_0}(\epsilon))\Big|_{\epsilon=0}$ and this score equation in ϵ is solved by ϵ_0. This proves that $\epsilon_0 = 0$ under the assumption that this score equation $P_0L(\bar{Q}^{b,\theta_0}(\epsilon)) = 0$ has a unique solution. The latter follows from the fact that the submodel with single parameter ϵ has an expected loss that is strictly convex.

This now proves that the limit $\Psi((\bar{Q}^{b,\theta_0}(\epsilon_0), Q_{W0})) = \Psi((\bar{Q}^{b,\theta_0}, Q_{W,0}))$ so that we can apply the previous theorem which shows that the latter limit equals ψ_0. This proves the consistency of the TMLE when b is a balancing score.

Consider now the case that $\bar{Q} = \bar{Q}_0$. Then $\theta_0 = 0$ and thus $\bar{Q}^{b,\theta_0} = \bar{Q}_0$. Thus, the limit $\Psi((\bar{Q}^{b,\theta_0}, Q_{W0})) = \Psi((\bar{Q}_0, Q_{W0}))$, which proves the consistency under the condition that $\bar{Q} = \bar{Q}_0$. In the latter case, it also follows that $\epsilon_0 = 0$. □

Lemma 2. *If* $\bar{g}_n$ *takes only discrete values with support G, then* $\Psi((\bar{Q}_n^{\bar{g}_n,\theta_n}, Q_{Wn}))$ *is a TMLE if* θ_0 *is estimated as* θ_n *using MLE in a saturated parametric model*

$$logit\bar{Q}_n^{g_n,\theta}(a,w) = logit(\bar{Q}_n(A,W)) + \sum_{\substack{a\in\{0,1\}\\ c\in G}} \theta_{a,c}I(A = a, \bar{g}_n(W) = c) \tag{3.4}$$

where $\bar{Q}_n$ *is some initial estimator for* $\bar{Q}_0$ *and I is the indicator function.*

Proof of Lemma 2. The MLE θ_n (or empirical risk minimizer for the negative quasi-binomial log likelihood, if Y is not binary), solves the score equations for each parameter $\theta_{a,c}$:

$$0 = \sum_{i=1}^{n} I(A_i = a, \bar{g}_n(W_i) = c)(Y - \bar{Q}_n^{g_n,\theta_n}(A_i, W_i)).$$

Additionally, any function h of A and $\bar{g}_n(W)$ is in the linear span of basis functions $I(A = a, \bar{g}_n(W) = c)$ for all $a \in \{0, 1\}$, $c \in G$, so

$$0 = \sum_{i=1}^{n} h(A_i, \bar{g}_n(W_i))(Y - \bar{Q}_n^{g_n,\theta_n}(A_i, W_i)).$$

In particular, the above equation is solved when $h(a, w) = \frac{a}{\bar{g}_n(w)}$, which is the score from the parametric submodel in (3.4). Thus if the TMLE update is applied to the initial estimate $\bar{Q}_n^0 = \bar{Q}_n^{\bar{g}_n,\theta_n}$, $\epsilon_n = 0$, and $\bar{Q}_n^* = \bar{Q}_n^0$ so $\Psi((\bar{Q}_n^{\bar{g}_n,\theta_n}, Q_{Wn}))$ is a TMLE. □

Theorem 5. *Define* $\Phi_1(Q) = P_0\bar{Q}\frac{\bar{g}-\bar{g}_0}{\bar{g}}$ *and* $\Phi_2(g) = P_0(\bar{Q} - \bar{Q}_0)\frac{\bar{g}}{\bar{g}_0}$. *Assume* $D^*(Q_n^*, g_n)$ *falls in a* P_0*-Donsker class with probability tending to 1;* $P_0\{D^*(Q_n^*, g_n) - D^*(Q, g)\}^2 \to 0$ *in probability as* $n \to \infty$*;*

$$\begin{aligned} P_0(\bar{Q}_0 - \bar{Q}_n^*)(\bar{g}_0 - \bar{g}_n)\frac{(\bar{g} - \bar{g}_n)}{\bar{g}\bar{g}_n} &= o_P(1/\sqrt{n}); \\ P_0(\bar{Q}_n^* - \bar{Q})(\bar{g}_n - \bar{g})/\bar{g} &= o_P(1/\sqrt{n}); \\ P_0(\bar{Q} - \bar{Q}_0)(\bar{g} - \bar{g}_0)/\bar{g} &= 0; \end{aligned}$$

$\Phi_1(\bar{Q}_n^*)$ *and* $\Phi_2(\bar{g}_n)$ *are asymptotically linear estimators of* $\Phi_1(\bar{Q})$ *and* $\Phi_2(\bar{g})$ *with influence curves* IC_1 *and* IC_2*, respectively.*

Then $\Psi(Q_n^*)$ *is asymptotically linear with influence curve* $D^*(Q, g) + IC_1 + IC_2$.

Proof. Since $P_0D^*(Q, g) = \psi_0 - \Psi(Q) + P_0(\bar{Q}_0 - \bar{Q})(\bar{g}_0 - \bar{g})/\bar{g}$ (e.g, Zheng and Mark J van der Laan (2010, 2012)), where we use the notation $\bar{Q}(W) = \bar{Q}(1, W)$, this results in the identity:

$$\Psi(Q_n^*) - \psi_0 = (P_n - P_0)D^*(Q_n^*, g_n) + P_0(\bar{Q}_0 - \bar{Q}_n^*)(\bar{g}_0 - \bar{g}_n)/\bar{g}_n.$$

The first term equals $(P_n - P_0)D^*(Q, g) + o_P(1/\sqrt{n})$ if $D^*(Q_n^*, g_n)$ falls in a P_0-Donsker class with probability tending to 1, and $P_0\{D^*(Q_n^*, g_n) - D^*(Q, g)\}^2 \to 0$ in probability as $n \to \infty$ (van der Vaart, 1998; van der Vaart and A., 1996). We write

$$P_0(\bar{Q}_0 - \bar{Q}_n^*)(\bar{g}_0 - \bar{g}_n)/\bar{g}_n = P_0(\bar{Q}_0 - \bar{Q}_n^*)(\bar{g}_0 - \bar{g}_n)/\bar{g} + P_0(\bar{Q}_0 - \bar{Q}_n^*)(\bar{g}_0 - \bar{g}_n)\frac{(\bar{g} - \bar{g}_n)}{\bar{g}\bar{g}_n}.$$

Assume that the last term is $o_P(1/\sqrt{n})$. We now write

$$\begin{aligned} &P_0(\bar{Q}_0 - \bar{Q}_n^*)(\bar{g}_0 - \bar{g}_n)/\bar{g} = P_0(\bar{Q}_n^* - \bar{Q} + \bar{Q} - \bar{Q}_0)(\bar{g}_n - \bar{g} + \bar{g} - \bar{g}_0)/\bar{g} \\ &= P_0(\bar{Q}_n^* - \bar{Q})(\bar{g}_n - \bar{g})/\bar{g} + P_0(\bar{Q}_n^* - \bar{Q})(\bar{g} - \bar{g}_0)/\bar{g} \\ &+P_0(\bar{Q} - \bar{Q}_0)(\bar{g}_n - \bar{g})/\bar{g} + P_0(\bar{Q} - \bar{Q}_0)(\bar{g} - \bar{g}_0)/\bar{g} \\ &\equiv P_0(\bar{Q}_n^* - \bar{Q})(\bar{g}_n - \bar{g})/\bar{g} + \Phi_1(\bar{Q}_n^*) - \Phi_1(\bar{Q}) \\ &+\Phi_2(\bar{g}_n) - \Phi_2(\bar{g}) + P_0(\bar{Q} - \bar{Q}_0)(\bar{g} - \bar{g}_0)/\bar{g}, \end{aligned}$$

where $\Phi_1(Q) = P_0\bar{Q}\frac{\bar{g}-\bar{g}_0}{\bar{g}}$ and $\Phi_2(g) = P_0(\bar{Q} - \bar{Q}_0)\frac{\bar{g}}{\bar{g}_0}$. We assume that the first term is $o_P(1/\sqrt{n})$, the last term equals zero (i.e., either $g = g_0$ or $\bar{Q} = \bar{Q}_0$), and $\Phi_1(\bar{Q}^*_n)$ and $\Phi_2(\bar{g}_n)$ are asymptotically linear estimators with influence curves IC_1 and IC_2, respectively. This proves $\Psi(Q^*_n)$ is asymptotically linear with influence curve $D^*(Q, g) + IC_1 + IC_2$. □

3.C TMLE when Y is not bounded by 0 and 1

If Y is not bounded by 0 and 1, but we can assume Y is bounded by l and u with $-\infty < l < u < \infty$, Y can be transformed to $Y^\dagger = \frac{Y-l}{u-l}$. Similarly $\bar{Q}^0_n$ can be transformed to $\bar{Q}^{0\dagger}_n = \frac{\bar{Q}^0_n - l}{u-l}$. The procedure described in Section 3.3 can be applied to the data structure $(W, A, Y^\dagger)$ using $\bar{Q}^{0\dagger}_n$ as initial estimator, and the final estimate can be transformed back to the original scale as $\Psi((\bar{Q}^*_n, Q_{Wn})) * (u - l) + l$. When l and u are not known, they can be set to the minimum and maximum of the observed Y as described in ().

For completeness we can define an alternative TMLE using a linear working model where

$$\bar{Q}^0_n(\epsilon)(A, W) = \bar{Q}^0_n(A, W) + \epsilon\frac{A}{g_n(1 \mid W)}$$

with loss function

$$L_Y(\bar{Q})(O) = (Y - \bar{Q}(A, W))^2$$

the squared error loss. Here, $\epsilon_0 = \arg\min_\epsilon E_0 L_Y(\bar{Q})(O)$ can be estimated by standard least squares regression software, with $\bar{Q}^0_n(A, W)$ as an offset.

Asymptotically, a TMLE using a linear working model (or linear fluctuation) is the equivalent to a TMLE with a logistic working model, but in practice can perform poorly. This is because if $g_n(1 \mid W_i)$ is very small for some observations, which is more likely in small samples, ϵ^0_n can be large in absolute value, having a large effect on $\bar{Q}^*_n$ with a linear fluctuation, which is unbounded. Because of this, if it is reasonable to bound Y by some l and u, it the logistic working model is recommended because $\bar{Q}^*_n$ always respects these bounds, even if ϵ^0_n is large.

3.D Example implementation of a BSA-TMLE estimator in R

```
bsatmle <- function(QnA1, QnA0, gn1, A, Y, family="binomial") {
    # computes estimates of E(E(Y|A=1, W)) (called ey1 in the
    # output), E(E(Y|A=0, W)) (called ey0), and
    # E(E(Y|A=1, W)) - E(E(Y|A=1, W)) (called ate)
    #
    # Inputs:
    # QnA1, QnA0: vectors, initial estimates of \bar{Q}_n(1, W)
```

```
#             and \bar{Q}_n(O, W)
# gn1: vector, estimates of g_n(1|W)
# A: vector, indicator of treatment
# Y: vector, outcome
# family: "binomial" for logistic fluctuation, "gaussian"
#         for linear fluctuation.
#         if "binomial", Y should be binary or bounded
#         by 0 and 1

if (!require(mgcv)) stop("mgcv package is required")
if (family=="binomial") {
    #use quasibinomial to suppress error messages about
    #non-integer Y
    family <- "quasibinomial"
    link <- qlogis
} else {
    link <- identity
}

QnAA <- ifelse(A==1, QnA1, QnA0)

# Use a generalized additive model to estimate theta_0
# using the initial estimate of \bar{Q}
gamfit <- gam(Y~factor(A)+s(gn1, by=factor(A))+offset(off),
    family, data=data.frame(A=A, gn1=gn1, off=link(QnAA)))

#Get predictions from gam fit
QnA1.gam <- predict(gamfit, type="response",
    newdata=data.frame(A=1, gn1=gn1, off=link(QnA1)))
QnA0.gam <- predict(gamfit, type="response",
    newdata=data.frame(A=0, gn1=gn1, off=link(QnA0)))
QnAA.gam <- ifelse(A==1, QnA1.gam, QnA0.gam)

# compute a/g_n(1|W)
hA1 <- 1 / gn1
hA0 <- -1 / (1 - gn1)
hAA <- ifelse(A==1, hA1, hA0)

#using glm, fluctuate the gam-updated initial fit of \bar{Q}
glmfit <- glm(Y~-1+h + offset(off), family,
    data=data.frame(h=hAA, off=link(QnAA.gam)))
```

```
    QnA1.star <- predict(glmfit, type="response",
        newdata=data.frame(h=hA1, off=link(QnA1.gam)))
    QnA0.star <- predict(glmfit, type="response",
        newdata=data.frame(h=hA0, off=link(QnA0.gam)))

    #compute the final estimates
    ey1 <- mean(QnA1.star)
    ey0 <- mean(QnA0.star)
    ate <- ey1-ey0

    list(ey1=ey1, ey0=ey0, ate=ate)
}
```

Chapter 4

Scalable Causal Infernece

4.1 Introduction

As the size of data sets grow, computational time becomes the limiting factor in statistical and machine learning problems. Methods that scale super-linearly in the number of observations are not practical as the number of observations grows extremely large or possibly infinite. Developing scalable methods for machine learning is an active area of research, and in recent years, procedures have been developed that scale well as sample sizes grow for tasks like classification and regression.

For causal inference and effect estimation, we often want to estimate a relatively low dimensional statistical parameter in a semiparametric or nonparametric model. Semiparametric efficient estimators of pathwise differentiable target parameters have been developed using some general approaches such as one-step estimation (Bickel et al., 1993), estimating equation methodology (James M Robins, Rotnitzky, and Zhao, 1994; M. J. van der Laan and J. M. Robins, 2003), and targeted minimum loss-based estimation (TMLE) (M. J. van der Laan and S. Rose, 2011; Mark J. van der Laan and Daniel Rubin, 2006). Though the target parameter is much lower in dimension than the number of covariates or features per observation in the data set, estimation of the target parameter often requires estimation of high dimensional functions such as a conditional mean or conditional probability. Usually computational complexity of such estimators is not taken into account. In this article, we introduce scalable methods for estimating parameters in a semi-parametric model with applications to causal inference.

What does it mean for a method to be scalable? First we consider so called batch methods which are traditionally used to fit statistical and machine learning estimators. Such methods typically update some intermediate estimate iteratively, where each iteration is computed using the whole data set—the whole "batch". Iteration is stopped when the estimate meets some predefined criterion. The computational complexity of batch methods as a function of the number of samples n measures the number of operations needed to compute the estimate. When a data set fits in main memory, a batch method that takes $O(n)$ operations per iteration

and dozens or hundreds of iterations may be practical.

In the age of big data, the computational complexity of batch methods does not correspond to wall clock time. It is simple to see why. To compute the wall clock time from the computational complexity of an algorithm, we only need to know the "constant" hidden in the big O notation representing the time per operation. This number turns out to not be constant with respect to n. As the size of data sets grow too large to fit in main memory, the time to access each data point increases quickly as data must be read from disk or across a network. This data access time quickly dominates the total computation time. Batch methods that need many iterations using the whole data set can become orders of magnitude slower as n increases linearly and are no longer practical.

In contrast, incremental methods update an estimate using a relatively small, fixed number of observations, a mini-batch, at a time. They make more updates in total but relatively few passes through a data set compared to batch methods. Because each update only depends on a fixed number of observations, the time per update does not grow with n. An important incremental method is stochastic (or mini-batch) gradient descent (SGD) which is commonly used to fit predictive models like large scale generalized linear models (GLMs), neural networks, and support vector machines (Bottou, 2010). There are many variants of SGD (Bottou, 2012; Duchi, Hazan, and Y. Singer, 2011; Zeiler, 2012), but they all revolve around a common theme: make some approximation of the gradient of an objective function of the full data set with only a few observations, and perform gradient descent with that approximation. With a sufficiently large n, SGD can perform as well as a batch method like batch gradient descent with only a few passes over the data set, and the computation time can be orders of magnitude faster (Bottou, 2010).

Online methods can be thought of as incremental methods which only see each piece of data once and therefore only make a single pass through a data set. Such methods are particularly useful in settings where data becomes available constantly, and updating an estimate with new data by refitting on all data available is prohibitively expensive. Under some conditions, SGD-like methods have been shown to perform asymptotically as well as batch methods in one pass through the data set (Murata, 1998; Xu, 2011).

In this article, we consider a method to be scalable if it is online, the required memory is bounded and does not depend on total sample size, and the computational complexity of an update on each new fixed-size mini-batch of data is constant. In Section 4.2 we formulate the estimation problem and introduce some notation. In Section 4.3 we introduce an online one-step estimator for a pathwise differentiable parameter and an online targeted one-step estimator in Section 4.4. In Section 4.5, we review some results on stochastic gradient descent and discuss how SGD can be used to compute initial estimators of components of our final estimator. In Section 4.6, we describe the implementation of online one-step and online targeted one-step estimators for estimation of the average treatment effect and investigate their performance via a simulation study. We conclude with a discussion in Section 4.7.

4.2 Formulation of the online estimation problem

Let $O_1, \ldots, O_n$ be a set of n independent and identically distributed observations with probability distribution $P_0 \in \mathcal{M}$ where $\mathcal{M}$ is a statistical model. Let $0 = n_0 < n_1 < n_2 < \ldots < n_K = n$. Here $n = n_K$ represents the total sample size, while n_j represents the sample size at the jth mini-batch. Each mini-batch j adds a next group of $n_j - n_{j-1}$ observations O_i with $i = n_{j-1} + 1, \ldots, n_j$. We do not assume that the number of new samples at each mini-batch, $n_j - n_{j-1}$ is converging to infinity, but instead, we assume that K and thus n_K converge to infinity. For simplicity, we assume $n_j - n_{j-1} = m$ is constant.

Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be a Euclidean target parameter mapping of interest so that $\Psi(P_0)$ denotes the desired estimand we want to learn from the data. Suppose that $\Psi(P) = \Psi_1(Q(P))$ for some parameter mapping Ψ_1 and parameter $P \to Q(P)$ on $\mathcal{M}$, so that $\Psi(P)$ only depends on P through a smaller part $Q(P)$. Recognizing the abuse of notation, we denote Ψ_1 with Ψ for convenience.

Assume that Ψ is pathwise differentiable at P for each $P \in \mathcal{M}$ and let $D^*(P)$ be the efficient influence curve (EIC) of $\Psi : \mathcal{M} \to \mathbb{R}^d$ at P. The efficient influence curve is defined as the canonical gradient of the pathwise derivative along parametric paths through P. That is, for any path $\{P(\epsilon) : \epsilon\} \subset \mathcal{M}$ through P with score $S = \frac{d}{d\epsilon} \log P(\epsilon)\Big|_{\epsilon=0}$ at $\epsilon = 0$, we have

$$\frac{d}{d\epsilon}\Psi(P(\epsilon))\Big|_{\epsilon=0} = PD^*(P)S.$$

where we use the notation $Pf = \int f(o)dP(o)$. This canonical gradient is uniquely defined as the only gradient $D(P)$ (i.e, each component is an element of $L_0^2(P)$ and $PD^*(P)S = PD(P)S$ for all scores S) whose components are also an element of the so called tangent space $T(P)$ defined as the closure of the linear span of all the scores generated by the class of parametric paths.

Suppose that $D^*(P)$ only depends on P through $Q(P)$ and an nuisance parameter $G(P)$ defined on the model $\mathcal{M}$ and that we can write the efficient influence curve as a function of $Q(P)$, $G(P)$. To emphasize this we will use the notation $D^*(P) = D^*(Q(P), G(P))$ for some $(Q, G) \mapsto D^*(Q, G)$. The efficient influence curve determines the efficiency of an estimator of $\psi_0 = \Psi(P_0)$. An estimator ψ_n is an asymptotically efficient estimator of ψ_0 if and only if

$$\psi_n - \psi_0 = (P_n - P_0)D^*(P_0) + o_P(1/\sqrt{n}).$$

In other words, an estimator attains the smallest asymptotic variance among the class of all regular estimators if and only if the estimator is asymptotically linear with influence curve equal to the efficient influence curve.

Let $R(P, P_0)$ be defined by

$$P_0 D^*(Q, G) = \Psi(P_0) - \Psi(P) + R(P, P_0),$$

where, by the fact that $D^*(P)$ is the canonical gradient of the pathwise deriative so $(P_0 - P)D^*(P)$ can be interpreted as a first order expansion of Ψ, $R(P, P_0)$ is a second order

remainder that can be explicitly determined give Ψ and D^*. Equivalently, in terms of $D^*(P) = D^*(Q, G)$ and $\Psi(P) = \Psi(Q)$, we have

$$P_0 D^*(Q, G) = \Psi(Q_0) - \Psi(Q) + R(Q, G, Q_0, G_0) \tag{4.1}$$

for a specified second order term $R()$.

Let $\mathbf{O}_k = (O_{n_{k-1}+1}, \ldots, O_{n_k})$ represent the $m = n_k - n_{k-1}$ observations making up mini-batch k. For notational convenience, we define

$$D_k^*(P)(\mathbf{O}_k) = \frac{1}{m} \sum_{i=n_{k-1}+1}^{n_k} D^*(P)(O_i).$$

Before we proceed with presenting our proposed online estimators of ψ_0 in the next sections, we first formally define what we mean with an online estimator.

Definition 1. *An online estimator of a parameter ψ_0 based on a sequence of mini-batches $\mathbf{O}_1, \mathbf{O}_2, \ldots$ is a sequence of estimators $(\psi_k : k = 1, \ldots)$ with ψ_k being an estimator based on $\mathbf{O}_1, \ldots, \mathbf{O}_k$ satisfying the following property: there exist certain functions f_1 and f_2, and a sequence of estimators $(\eta_k : k = 1, \ldots)$ with $\eta_k = f_2(\mathbf{O}_k, \eta_{k-1})$, so that*

$$\psi_k = f_1(\mathbf{O}_k, \eta_{k-1}),\ k = 1, \ldots.$$

This definition can be applied to our target parameter ψ_0, but also to define an online estimator of (Q_0, G_0). Whether or not a particular online estimator is scalable in the sense described in Section 4.1 depends on choice of f_1 and f_2. In particular, we need that the memory required to store η_k and the computational complexity of evaluating f_1 and f_2 given $\mathbf{O}_k$ and η_{k-1} are bounded and do not depend on k. In general, it may be not be possible to choose η_k, f_1 and f_2 that meet this condition, but in Section 4.6 we present an example and describe a class of problems in which it is possible.

Let $((Q_k, G_k) : k = 1, \ldots,)$ be an online estimator of (Q_0, G_0). For example, this might be estimators using a stochastic gradient descent algorithm based on a high dimensional parametric model. In the next sections we will propose two online estimators of ψ_0 that map this online estimator $((Q_k, G_k) : k = 1, \ldots)$ into an online estimator ψ_k of ψ_0, so that ψ_k is only a function of $(Q_{k-1}, G_{k-1}, \psi_{k-1})$, and possibly a few more online low-dimensional statistics, and the new mini-batch $\mathbf{O}_k$.

A crucial ingredient in the analysis of our proposed online estimators is the following identity that is an immediate consequence of (4.1):

$$P_{0,k} D_k^*(Q_{k-1}, G_{k-1}) = \Psi(Q_0) - \Psi(Q_{k-1}) + R(Q_{k-1}, G_{k-1}, Q_0, G_0) \tag{4.2}$$

for $k = 1, \ldots, K$ where we used the notation $P_{0,k} f(\mathbf{O}_k) = \int f(\mathbf{O}_k) dP_{0,k}(\mathbf{O}_k)$ and

$$dP_{0,k}(\mathbf{O}_k) = \prod_{i=n_{k-1}+1}^{n_k} dP_0(O_i)$$

is the probability distribution of $\mathbf{O}_k$ implied by the common probability distribution P_0 and the fact that all O_i are independent. Note that we also have $P_{0,k}D_k^*(Q_{k-1}, G_{k-1}) = E_0(D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k) \mid \mathcal{F}_{k-1})$ is the conditional expectation of the random variable $D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k)$ (a function of $\mathbf{O}_1, \ldots, \mathbf{O}_k$), given $\mathcal{F}(k-1) = (\mathbf{O}_1, \ldots, \mathbf{O}_{k-1})$.

We will assume that an initial estimator $Q_{k=0}, G_{k=0}$ is given, so that the online procedure can be initiated with this choice. In practice this might be an estimator based on an initial mini-batch that is further ignored in our definition on the online estimator. However, one could also simply define $(Q_{k=0}, G_{k=0}) = (Q_1, G_1)$, i.e., as the online estimator based on the first mini-batch, since this choice does not affect the asymptotics (i.e., it only affects the impact of the first $n_1 = m$ observations in the online estimator which is asymptotically negligible as $K \to \infty$).

4.3 Online one-step estimator

In the batch setting, one-step estimation is one way of constructing an efficient estimator in a semi-parametric model (Bickel et al., 1993). Given Q_n and G_n, estimates of Q_0 and G_0, respectively, a one-step estimator is computed by taking a plug-in estimator $\Psi(Q_n)$ and updating it with a step in the direction of the empirical EIC:

$$\Psi(Q_n) + \frac{1}{n}\sum_{i=1}^{n} D^*(Q_n, G_n).$$

We now present an online version of the one-step estimator which is asymptotically efficient. Suppose we have some procedure for computing initial online estimates of Q_0 and G_0. In Section 4.6 we give a specific example of a useful initial estimator. Denote estimates of Q_0 and G_0 computed using the first $j-1$ mini-batches $Q_{n_{j-1}}$ and $G_{n_{j-1}}$ and initialize them appropriately. We define the online one-step estimator at the jth batch as

$$\psi_k = \frac{1}{k}\sum_{j=1}^{k} \Psi(Q_{j-1}) + \frac{1}{k}\sum_{j=1}^{k} \frac{1}{m} \sum_{i=n_{j-1}+1}^{n_j} D^*(Q_{j-1}, G_{j-1})(O_i).$$

This can be written equivalently as

$$\psi_k = \frac{k-1}{k}\psi_{k-1} + \frac{1}{k}\left[\Psi(Q_{k-1}) + \frac{1}{m}\sum_{i=n_{k-1}+1}^{n_k} D^*(Q_{k-1}, G_{k-1})(O_i)\right]. \tag{4.3}$$

We have the following theorem with a proof presented in Appendix 4.7.

Theorem 6. *Define* $\bar{M}(K) = \sum_{k=1}^{K} M_k$, *where*

$$M_k = D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k) - P_{0,k}D_k^*(Q_{k-1}, G_{k-1}).$$

Let $\Sigma_k^2 = E_0 M_k^2 = E_0 M_k M_k^\top$, *and* $\Sigma^2(K) = \frac{1}{K}\sum_{k=1}^K \Sigma_k^2$. *Define also*

$$\bar{R}(K) = \frac{1}{K}\sum_{k=1}^K R_0(Q_{k-1}, G_{k-1}, Q_0, G_0).$$

If we assume

1. *for some* $M < \infty$ $\max_k \mid D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k) \mid < M < \infty$ *with probability* 1,
2. $\bar{R}(K) = o_P(1/\sqrt{K})$,
3. $$\frac{1}{K}\sum_{k=1}^K P_{0,k} D_k^*(Q_{k-1}, G_{k-1})^2 - E_0 \frac{1}{K}\sum_{k=1}^K P_{0,k} D_k^*(Q_{k-1}, G_{k-1})^2 \to 0$$
   in probability as $K \to \infty$ *where* $D_k^*(Q_{k-1}, G_{k-1})^2 = D_k^*(Q_{k-1}, G_{k-1}) D_k^*(Q_{k-1}, G_{k-1})^\top$, *and*
4. $\liminf_{K\to\infty} \lambda \Sigma^2(K) \lambda > 0$ *for all* λ,

then

$$\psi_K - \psi_0 = \bar{M}(K)/K + o_P(1/\sqrt{K})$$

and

$$\Sigma(K)^{-1} \frac{\bar{M}(K)}{\sqrt{K}} \Rightarrow_D N(0, I), \text{ as } K \to \infty.$$

This implies

$$\sqrt{K}\Sigma(K)^{-1}(\psi_K - \psi_0) \Rightarrow_D N(0, I), \text{ as } K \to \infty.$$

If also $\Sigma^2 = \lim_{k\to\infty} \Sigma(k)^2$ *exists and is positive definite, then*

$$\sqrt{K}(\psi_K - \psi_0) \Rightarrow_D N(0, \Sigma^2), \text{ as } K \to \infty,$$

so

$$\sqrt{Km}(\psi_K - \psi_0) \Rightarrow_D N(0, \Sigma^2/m)$$

as $K \to \infty$, *and* $\Sigma^2/m = P_0 D^*(Q_0, G_0)^2$ *is the efficiency bound, so* ψ_K *is an asymptotically efficient estimator of* ψ_0.

Finally, consider the following estimator of $\Sigma^2(K)$*:*

$$\hat{\Sigma}^2(K) = \frac{1}{K}\sum_{k=1}^K \{D_k^*(Q_{k-1}, G_{k-1}(\mathbf{O}_k) - \bar{D}_K\}^2,$$

where $\bar{D}_K = \frac{1}{K}\sum_{k=1}^K D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k)$. *We have* $\hat{\Sigma}^2(K) - \Sigma^2(K) \to 0$ *in probability as* $K \to \infty$, *and if* Σ^2 *exists, then we also have* $\hat{\Sigma}^2(K) \to \Sigma^2$ *in probability as* $K \to \infty$.

By Theorem 6, under regularity conditions, the online one-step estimator ψ_K is an asymptotically efficient estimator for ψ_0 if the remainder term $\bar{R}(K)$ is sufficiently small $o_P(1/\sqrt{(K)})$. The form of $\bar{R}(K)$ depends on particular model and parameter mapping, but in general, this tells us something about the necessary rate of convergence of online initial estimators of Q_0 and G_0. In cases where (parts of) Q or G can be expressed as an optimum of a known loss function, (for example a generalized linear model), those parts can be consistently estimated with stochastic gradient descent or similar algorithms.

4.4 Online targeted one-step estimation

In the batch setting, targeted minimum loss-based estimation (TMLE) is a framework for constructing asymptotically efficient plug-in estimators. To define a TMLE, we choose a loss function $(Q, O) \to L(Q)(O)$ such that

$$Q_0 = \arg\min_Q P_0 L(Q).$$

We also choose a parametric working model through Q which depends on G $\{Q(\epsilon \mid G) : \epsilon\}$ such that the linear span of $\frac{d}{d\epsilon} L(Q(\epsilon \mid G))(O)\Big|_{\epsilon=0}$ contains $D^*(Q, G)$. Starting with initial estimate Q_n^0, set $Q_n^j = Q_n^{j-1}(\epsilon_n^{j-1} \mid G_n)$ where $\epsilon_n^{j-1} = \arg\min_\epsilon \frac{1}{n}\sum_{i=1}^n L(Q_n^{j-1}(\epsilon \mid G_n)(O_i))$ for $j = 1, 2, \ldots$ until convergence. Convergence is reached when $\epsilon_n^j \approx 0$ or $\frac{1}{n}\sum_{i=1}^n D^*(Q_n^j, G_n) \approx 0$. In some cases, the algorithm converges in one iteration. When convergence is reached, the final estimate is calculated as $\Psi(Q_n^*)$ where $Q_n^* = Q_n^J$ at the last iteration J. For more details and examples see Mark J. van der Laan and Sherri Rose (2011).

Under regularity conditions, the TMLE $\Psi(Q_n^*)$ is asymptotically linear and efficient, like the one-step estimator. An advantage of TMLE is that it is a plug-in estimator, computed by plugging in a good estimate of Q_0 to the parameter mapping Ψ. Being a plug-in estimator guarantees that the estimate respects the global constraints of the model and in particular that the estimate of target parameter is in the parameter space, which is not true in general for a one-step estimator in finite samples. Though the main motivation for online methods is huge data sets, estimates can be quite unstable when a relatively small number of mini-batches have been processed. Taking inspiration from batch TMLE, we present the online targeted one-step estimator.

The main idea is that we want to update an initial estimate at mini-batch k, Q_k, to Q_k^* so that

$$\frac{1}{K}\sum_{k=1}^K D_k^*(Q_{k-1}^*, G_{k-1})(\mathbf{O}_k), \tag{4.4}$$

which we call the online efficient influence curve estimating equation, is small as K increases. We then use (Q_k^*, G_k) as online initial estimates of (Q_0, G_0) to compute the online one-step estimator in Section 4.3 and call this an online targeted one-step estimator.

If eq. (4.4) is sufficiently small, in particular $o_p(1/\sqrt{K})$, then the online one-step estimator at mini-batch k in eq. (4.3) is asymptotically equivalent to

$$\frac{1}{K}\sum_{k=1}^{K}\Psi(Q_k^*),$$

which can be used as our estimate of ψ_0. If we cannot guarantee that the online efficient influence curve estimating equation is solved up to an $o_P(1/\sqrt{K})$ term, we hope that updating Q_k to Q_k^* still improves the performance of the online one-step estimator.

It may not be clear how to compute the update to Q_k^* in general. In batch TMLE, the updating step is sometimes an iterative procedure, but in problems where it converges in one iteration, the update can be extended to the online case. When this is the case, we define $Q_k^* = Q_k(\epsilon_k \mid G_k)$ where ϵ_k is an online estimator of ϵ. We describe a concrete example in Section 4.6.

4.5 Initial estimators with stochastic gradient descent

Often, we can express initial estimators of parts of Q_0 or G_0 as as an optimum of an empirical risk. For example, we may need to estimate some conditional mean or conditional probability. In those cases, stochastic gradient descent based estimators may be a natural choice. We review some relevant literature on SGD and variants in this section.

Suppose we parametrize our estimator through some loss function $O \to L'(\theta)(O)$ and define the target parameter as $\theta_0 = \arg\min_\theta P_0 L'(\theta)$ for $\theta \in \Theta \subseteq \mathbb{R}^p$. For example, if $O = (X, Y)$ where $Y \in \{0, 1\}$ and $X \in \mathbb{R}^p$, we can estimate $P_0(Y = 1 \mid X)$ using a working logistic regression model by choosing

$$L'(\theta)(O) = -Y \log(\text{logistic}(\theta^\top X)) - (1 - Y)\log(1 - \text{logistic}(\theta^\top X))$$

where $\text{logistic}(z) = 1/(1 + \exp(-z))$. We could also choose to include a regularization term in the loss function. Other examples include least squares regression models and linear support vector machines.

For a data set with empirical distribution P_n, call the true optimum of the empirical mean of the loss function, also known as the empirical risk, $\hat{\theta}_n$. That is,

$$\hat{\theta}_n = \arg\min_\theta P_n L'(\theta) = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} L'(\theta)(O_i).$$

When $L'(\theta) = -\log p_\theta$ for some parametric model $\{p_\theta : \theta \in \Theta\}$, $\hat{\theta}_n$ is the maximum likelihood estimator. Let $V_\theta = P_0 \frac{d^2}{d\theta^2} L'(\theta)$. Under mild regularity conditions (see e.g., (van der Vaart, 1998)), we have $\hat{\theta}_n$ is asymptotically normally distributed with asymptotic variance

$$V_{\theta_0}^{-1} P_0 \left[\frac{d}{d\theta_0} L'(\theta_0)^\top \frac{d}{d\theta_0} L'(\theta_0)\right] V_{\theta_0}^{-1}.$$

Stochastic gradient descent is an iterative optimization routine which takes a small step in the direction of a single randomly selected observation from the data set. In practice, the data set is usually shuffled or assumed to be in random order and processed sequentially. Let

$$\theta_{t+1} = \theta_t - \gamma_t \Gamma_t \frac{d}{d\theta_t} L'(\theta_t)(O_t) \tag{4.5}$$

where γ_t is a scalar step size or learning rate, Γ_t is a $d \times d$ matrix, and O_t is the observation used at the t-th step (Bottou, 2010). After some number of steps, we hope that θ_t is sufficiently close to the true optimum $\hat{\theta}_n$ of the empirical risk. In particular, we hope that n steps is enough so that the SGD estimate θ_n after a single pass through the data set is a reasonable estimate of θ_0.

In the simplest version of SGD Γ_t is some constant times the identity matrix. Other variants replace Γ_t with an appropriate diagonal matrix (e.g., an approximation of the diagonal elements of $V_{\theta_t}^{-1}$) as in Adagrad (Duchi, Hazan, and Y. Singer, 2011) and Adadelta (Zeiler, 2012), which are methods that tend to work well in practice. Murata (1998) shows that the mean and variance of θ_t depend on the learning rate γ_t and the eigenvalues of the matrix $\Gamma_t V_{\hat{\theta}_n}$. Second order SGD takes the curvature of the loss function into account, using some Γ_t that approximates $V_{\hat{\theta}_n}^{-1}$. Murata (1998) shows that when $\Gamma_t = V_{\hat{\theta}_n}^{-1}$ and γ_t is asymptotically $1/t$, θ_n, the second order SGD estimate after a single pass through the data set, is asymptotically equivalent with the true empirical optimum $\hat{\theta}_n$. That is, asymptotically, the variance of second order SGD divided by the variance of $\hat{\theta}_n$ converges to 1 as $n \to \infty$. Murata (1998) shows that, if Γ_t is constant and some weak conditions hold, then θ_n has bias of $O(1/n^{\lambda_d})$, where λ_d is the smallest eigenvalue of $\Gamma_n V_{\hat{\theta}_n}$, and the variance is $O(1/n)$ if $\lambda_d > 1/2$.

Though optimal, due to the high dimension of p, second order SGD is rarely used in practice because it is often too expensive to compute and store (an estimate of) $V_{\theta_n}^{-1}$. Averaged stochastic gradient descent (ASGD) is another different but related method to SGD which is very simple to implement. The ASGD estimate at step t is simply

$$\bar{\theta}_t = \frac{1}{t} \sum_{i=1}^{t} \theta_i$$

where θ_i is the SGD estimate at step i as in (4.5), Γ_t is the identity matrix times a constant, and γ_t now goes to 0 slower than $1/t$. Polyak and Juditsky (1992) and Xu (2011) show that in a single pass through the data set, $\bar{\theta}_n$ is also asymptotically optimal and thus equivalent with $\hat{\theta}_n$. Xu (2011) note that ASGD is not frequently used in practice possibly due to required tuning and the possibly huge number of observations required to reach the asymptotic performance, but it is shown in simulations that with some careful tuning, ASGD can perform very well.

There are many other variants of stochastic gradient descent type optimization routines. For more details and some insightful notes on implementation details, see (Bottou, 2012) and references therein.

4.6 Online efficient estimation of the average treatment effect

Suppose our observed data structure is $O = (W, A, Y) \sim P_0$ where $W \in \mathbb{R}^p$, A is Bernoulli, and Y is univariate and $\mathcal{M}$ is the non-parametric model. Define $\Psi(P) = E_P[E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W)]$. Under a causal model, this statistical parameter is equal to the average treatment effect (ATE)(Holland, 1986; Neyman, 1990; J. Pearl, 2009; J.M. Robins, 1987a,b; D.B. Rubin, 2006; D. B. Rubin, 1974). Note that $\Psi(P)$ only depends on P through $Q(P) = (Q_W, \bar{Q})$ where Q_W is the marginal distribution of W and $\bar{Q}(A, W) = E_P(Y \mid A, W)$.

The efficient influence curve is given by

$$D^*(Q, G)(O) = \frac{2A - 1}{G(A \mid W)}(Y - \bar{Q}(A, W)) + \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(Q),$$

where $G(A \mid W) = P(A \mid W)$ (M. J. van der Laan and J. M. Robins, 2003; Mark J. van der Laan and Sherri Rose, 2011). The nuisance parameter G_0 is sometimes called the treatment mechanism or propensity score. Note that $D^*(Q, G) = D^*(\bar{Q}, Q_W, G)$.

We have

$$P_0 D^*(Q, G) = \Psi(Q_0) - \Psi(Q) + R_0(\bar{Q}, G, \bar{Q}_0, G_0),$$

where

$$\begin{aligned} R_0(\bar{Q}, G, \bar{Q}_0, G_0) =& E_{P_0}(\bar{Q} - \bar{Q}_0)(1, W)\frac{G - G_0}{G}(1 \mid W) - \\ & E_{P_0}(\bar{Q} - \bar{Q}_0)(0, W)\frac{G - G_0}{G}(0 \mid W). \end{aligned}$$

This also defines the efficient influence curve $D_k^*(Q, G)(\mathbf{O}_k)$ for the k-th batch $\mathbf{O}_k$, and the corresponding identity

$$P_{0,k} D_k^*(Q, G) = \Psi(Q_0) - \Psi(Q) + R_0(\bar{Q}, G, \bar{Q}_0, G_0).$$

We also note that $R(\bar{Q}_0, G, \bar{Q}_0, G_0) = R(\bar{Q}, G_0, \bar{Q}_0, G_0) = 0$, so the efficient influence curve has the so called double robustness property. Though Theorem 6 presents conditions for asymptotic efficiency that rely on consistency of Q_k, in this case the online one-step estimator can remain asymptotically linear if either Q_k or G_k is consistent, but not necessarily both. Such a more general theorem for the online one-step estimator would be a completely analogue to these types of theorems presented for a batch one-step estimator, and is not repeated here.

Online one-step estimator for the ATE

To construct an online one-step estimator for ψ_0, we first need to choose initial estimators for Q_0 and G_0. The treatment mechanism G_0 is a conditional probability, so we can choose as an

estimator a logistic regression model for $G_0(1 \mid W)$ fit by SGD. The outcome regression $\bar{Q}_0$ is a conditional mean which we can also estimate with a generalized linear model fit with SGD. If Y is binary, we can choose logistic regression, or if Y is continuous, we may choose linear regression. We note that if Y is continuous and (scaled to be) bounded between 0 and 1, it is often useful in practice to choose the negative quasi-binomial log likelihood as a loss function for $\bar{Q}_0$, essentially performing logistic regression on a continuous outcome. This guarantees that estimates are also bounded between 0 and 1(Gruber and Mark J van der Laan, 2010).

Finally, we specify an estimator for Q_{W0}. A natural choice is the empirical distribution of W, but storing the empirical distribution requires order n storage and quickly becomes impractical and does not fit our definition of a scalable estimator in Section 4.1. We ignore this issue temporarily. Let

$$D_1^*(Q,G)(O) = \frac{2A-1}{G(A \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(1,W) - \bar{Q}(0,W)$$

and compute the estimate of ψ_0 and mini-batch k by eq. (4.3):

$$\begin{aligned}
\psi_k =& \frac{k-1}{k}\psi_{k-1} + \frac{1}{k}\left[\Psi(Q_{k-1}) + \frac{1}{m}\sum_{i=n_{k-1}+1}^{n_k} D^*(Q_{k-1},G_{k-1})(O_i)\right] \\
=& \frac{k-1}{k}\psi_{k-1} + \frac{1}{k}\left[\Psi(Q_{k-1}) + \frac{1}{m}\sum_{i=n_{k-1}+1}^{n_k} D_1^*(Q_{k-1},G_{k-1})(O_i) - \Psi(Q_{k-1})\right] \\
=& \frac{k-1}{k}\psi_{k-1} + \frac{1}{k}\frac{1}{m}\sum_{i=n_{k-1}+1}^{n_k} D_1^*(Q_{k-1},G_{k-1})(O_i) \\
=& \frac{k-1}{k}\psi_{k-1} + \frac{1}{k}\frac{1}{m}\sum_{i=n_{k-1}+1}^{n_k}\Bigg[\frac{2A_i-1}{G_{k-1}(A_i \mid W_i)}(Y - \bar{Q}_{k-1}(A_i,W_i)) \\
& \qquad + \bar{Q}_{k-1}(1,W_i) - \bar{Q}_{k-1}(0,W_i)\Bigg].
\end{aligned}$$

We see that because $D^*(Q,G)$ has the form $D_1^*(Q,G) - \Psi(Q)$, and D_1^* does not depend on Q_W, we can compute the estimate ψ_k without evaluating $\Psi(Q)$ directly, and therefore we do not need to store the empirical distribution of W for all mini-batches at once.

Online targeted one-step estimator for the ATE

A batch TMLE procedure for ψ_0 can be developed that updates an initial Q_n to Q_n^* in one iteration (Gruber and Mark J van der Laan, 2010). For example, assuming Y or bounded between 0 and 1 for simplicity, we can choose as loss function $L(Q) = L_Y(\bar{Q}) + L_W(Q_W)$ where

$$L_Y(\bar{Q})(O) = -Y\log(\bar{Q}(A,W)) - (1-Y)\log(1-\bar{Q}(A,W))$$

and L_W is the negative log likelihood loss. We can then choose

$$\text{logit}\bar{Q}(\epsilon \mid G)(O) = \text{logit}\bar{Q}(A, Y) + \epsilon \frac{2A - 1}{G(A, W)}$$

as a working parametric submodel through $\bar{Q}$. The parameter can then be estimated as

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} L_Y(\bar{Q}_n(\epsilon \mid G_n)(O_i))$$

given initial estimators $\bar{Q}_n$ and G_n. The initial estimate of $\bar{Q}_n$ is then update to $\bar{Q}_n^* = \bar{Q}_n(\epsilon_n \mid G_n)$. If we choose the empirical distribution of W as the initial estimate of Q_{W0}, the negative log likelihood loss function L_W is already minimized, so no further update of initial estimate Q_{Wn} is needed. The final TMLE estimate of ψ_0 is then computed as $\Psi(Q_n*) = \Psi(\bar{Q}_n^*, Q_{Wn})$.

We use this batch estimator as a starting point for developing an online targeted one-step estimator for ψ_0. Using $\bar{Q}(\epsilon \mid G)$ as defined above for a working model, we need to choose how to estimate ϵ at each mini-batch. One choice is to choose ϵ_k to be the minimizer of the loss L_Y on mini-batch k using $\bar{Q}_k$ as an initial offset:

$$\epsilon_k = \arg\min_{\epsilon} \sum_{i=n_{k-1}+1}^{n_k} L_Y(\bar{Q}_k(\epsilon \mid G_k)(O_i)). \tag{4.6}$$

For the chosen working model, this requires fitting a logistic regression on m observations at each mini-batch, but the regression is univariate and m is relatively small, so this will generally be fairly fast. One potential issue is that the variance of ϵ_k will be of order $1/m$, so $\bar{Q}_k^*$ will not converge even when $\bar{Q}_k$ does. This may not be an issue in practice if m is not too small. Instead, we may choose ϵ_k to be the mean of the mini-batch specific risk minimizers in eq. (4.6) for mini-batches 1 to k so that the variance of ϵ_k is of order $1/k$.

Alternatively, we may also choose to estimate ϵ_k with stochastic gradient descent or some variant, using the initial $\bar{Q}_k$ as an offset at each mini-batch. The computation of ϵ_k will be faster with SGD than minimizing an empirical risk at each mini-batch, but may be more sensitive to tuning parameters.

Simulations

We evaluate the statistical performance of the online one-step estimator and online targeted one-step estimator for the average treatment effect in a simulation study. For each observation, we make $p = 2000$ independent draws from a uniform distribution on $[-1, 1]$ for W. We then draw A from a Bernoulli distribution with success probability

$$\frac{1}{1 + \exp(-0.75[W(1) + W(2) + W(3) + W(4)])}$$

where $W(j)$ is the jth component of W. Finally, we draw Y from a Bernoulli distribution with success probability

$$\frac{1}{1+\exp(1+0.5[W(1)+W(2)+W(3)+W(4)]-0.3A)}.$$

The value of ψ_0, the true parameter of interest, is approximately 0.060. The first 4 components of the covariate vector W are confounders because they are related to both Y and A. In this data generating distribution, confounding is strong enough that failing to adjust for confounders will result in substantial bias. A naive estimate of ψ_0 that does not adjust for W is approximately -0.026. The asymptotic variance bound, $P_0D(Q_0, G_0)^2$, for this data generating distribution is approximately 0.95.

For initial estimators of G_0 and $\bar{Q}_0$, we use mini-batch gradient descent with a learning rate of the form $a/(1+bk)$ for mini-batch k. When estimating G_0, we include an intercept and each component of W as main terms. For $\bar{Q}_0$, we include an intercept, each component of W, and A as main terms.

We also investigate the performance of our estimators when one of the initial estimators of G_0 or $\bar{Q}_0$ is badly misspecified. When the estimate of G_0 is misspecified, we use an intercept only model, and when the estimate of $\bar{Q}_0$ is misspecified, we include an intercept and A as main terms.

First, we compare the online one-step and online targeted one-step estimators using SGD to compute ϵ_k for the online targeted one-step estimators on data sets up to size $n_K = 5,000,000$ with mini-batch size $m = 100$. For estimators of both G_0 and $\bar{Q}_0$, we choose $0.1/(1+0.001k)$ as the learning rate, though it is not necessary for the learning rate to be the same for both. For computing ϵ_k, we chose a learning rate of $0.1/(1+0.01k)$. We call this Simulation 1.

We compute the bias and variance of each estimator at each mini-batch k from $1,000$ simulations and plot the results in Figures 4.1 and 4.2. In Figure 4.1, bias scaled by n_k is plotted, and we want to see the absolute value of scaled bias converge to 0 as sample size increase. Not surprisingly, bias can be quite large when relatively few mini-batches have been processed. We see that when both estimators of $\bar{Q}_0$ and G_0 are correctly specified, bias reaches nearly 0 as sample size increases. When one of $\bar{Q}_0$ or G_0 is misspecified, bias is larger because we are relying on double robustness of the estimators. In that case, a larger sample size is needed for the correctly specified estimator of $\bar{Q}_0$ or G_0 compensate for the misspecified estimator. In particular, when the estimator for $\bar{Q}_0$ is misspecified, we see that the online one-step estimator has a relatively higher bias which is very slowly decreasing to 0. This may be because the learning rate for the correctly specified estimator of G_0 is tuned poorly. The corresponding online targeted one-step estimator has lower bias, so it appears that the targeting step is particularly helpful in this case.

In Figure 4.2, we plot a smoothed scaled variance for each estimator by cumulative sample size. We see that the variances stabilize slightly below the theoretical variance bound indicated by the red line. This indicates that both the online one-step and online targeted one-step estimators are efficient when sample size is sufficiently large and both estimators of $\bar{Q}_0$ and

G_0 are correctly specified. We note that when one of the initial estimators of $\bar{Q}_0$ or G_0 is misspecified, the estimator for ψ_0 is not guaranteed to be efficient, but in this data generating distribution we do not observe any loss of efficiency. When only a few mini-batches have been processed, we also see that the online targeted one-step estimator has lower variance than the online one-step estimator.

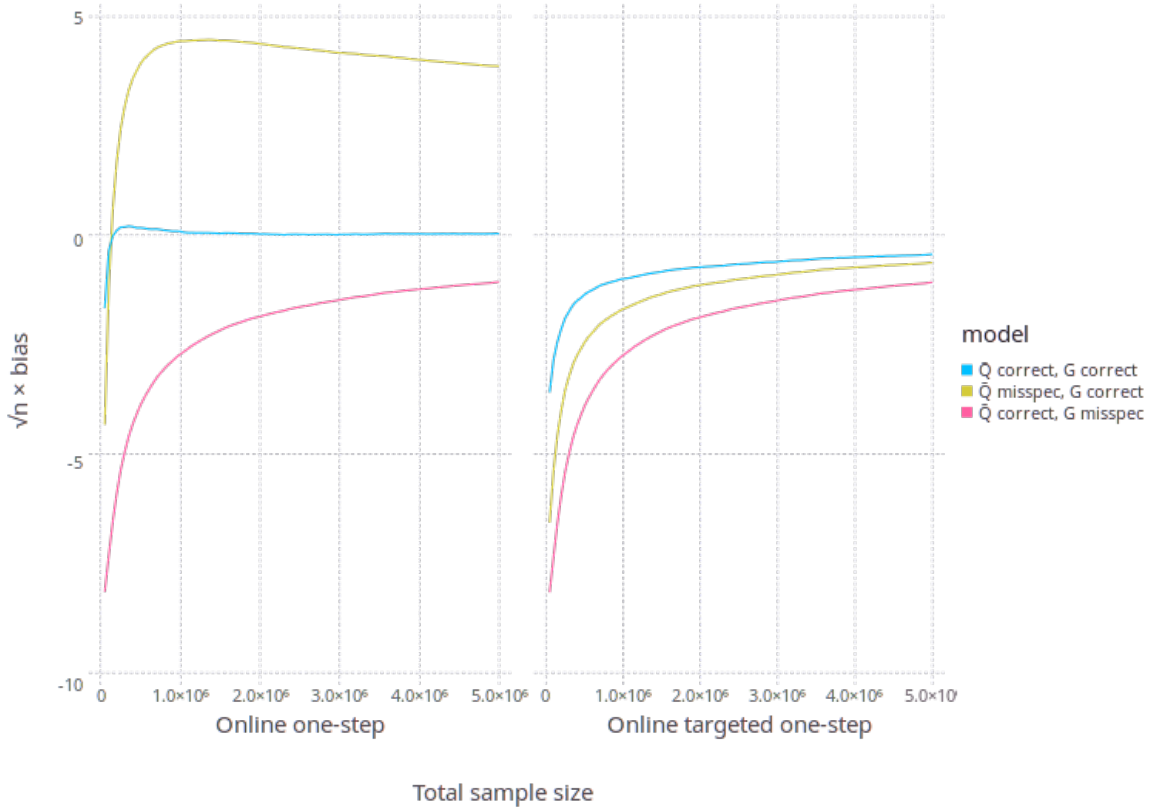

Figure 4.1: Simulation 1, bias scaled by $\sqrt{n_k}$ for the online one-step and online targeted one-step estimators.

We also investigated the performance of the online targeted one-step estimator when ϵ_k is computed by minimizing the empirical risk of L_Y at each mini-batch as in eq. (4.6). Both bias and variance of the online targeted one-step estimator for ψ_0 are almost the same as when ϵ_k is computed with SGD as in Simulation 1, so results are not shown.

In Simulation 2, we try adjusting the learning rate for the estimator of G_0 to see how sensitive the estimators of ψ_0 are to tuning parameters. We now use a learning rate of $0.1/(1+0.005k)$ at mini-batch k for the estimator of G_0, and use the same choices for $\bar{Q}_0$ and computing ϵ_k as in Simulation 1. We plot the bias in fig. 4.3 from $1,000$ simulated data sets. Now we see that the online one-step estimator has a much lower bias when the estimator for

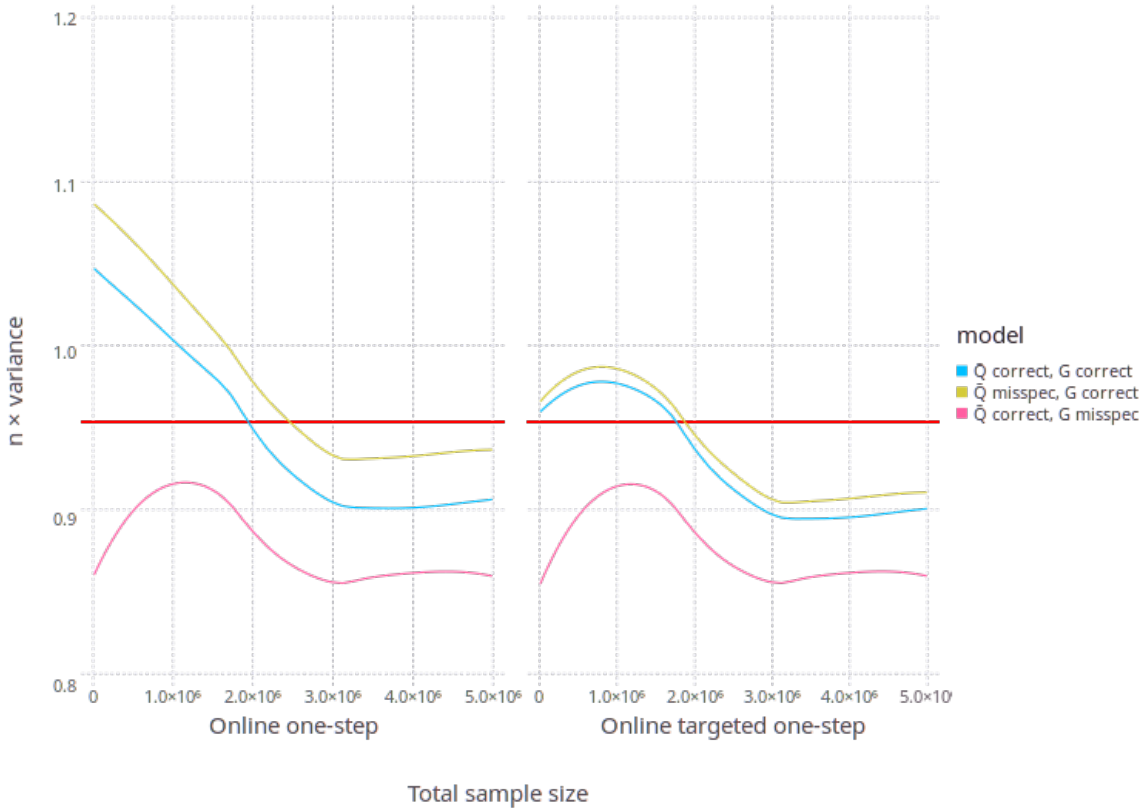

Figure 4.2: Simulation 1, smoothed variance scaled by n_k for the online one-step and online targeted one-step estimators.

$\bar{Q}_0$ is misspecified than in Simulation 1. We also see that when one of the estimators for $\bar{Q}_0$ is misspecified, the online targeted one-step estimator has somewhat higher bias than the online one-step estimator. The variance of both estimators are similar to those in Simulation 1, and results are not shown.

We tried other combinations of tuning parameters for learning rates and found, unsurprisingly, that performance of estimators can vary greatly with tuning parameters. Usually the performance of the online targeted one-step estimator was less sensitive to tuning parameters for the initial estimators, in particular for the estimator of G_0, but this was not always the case.

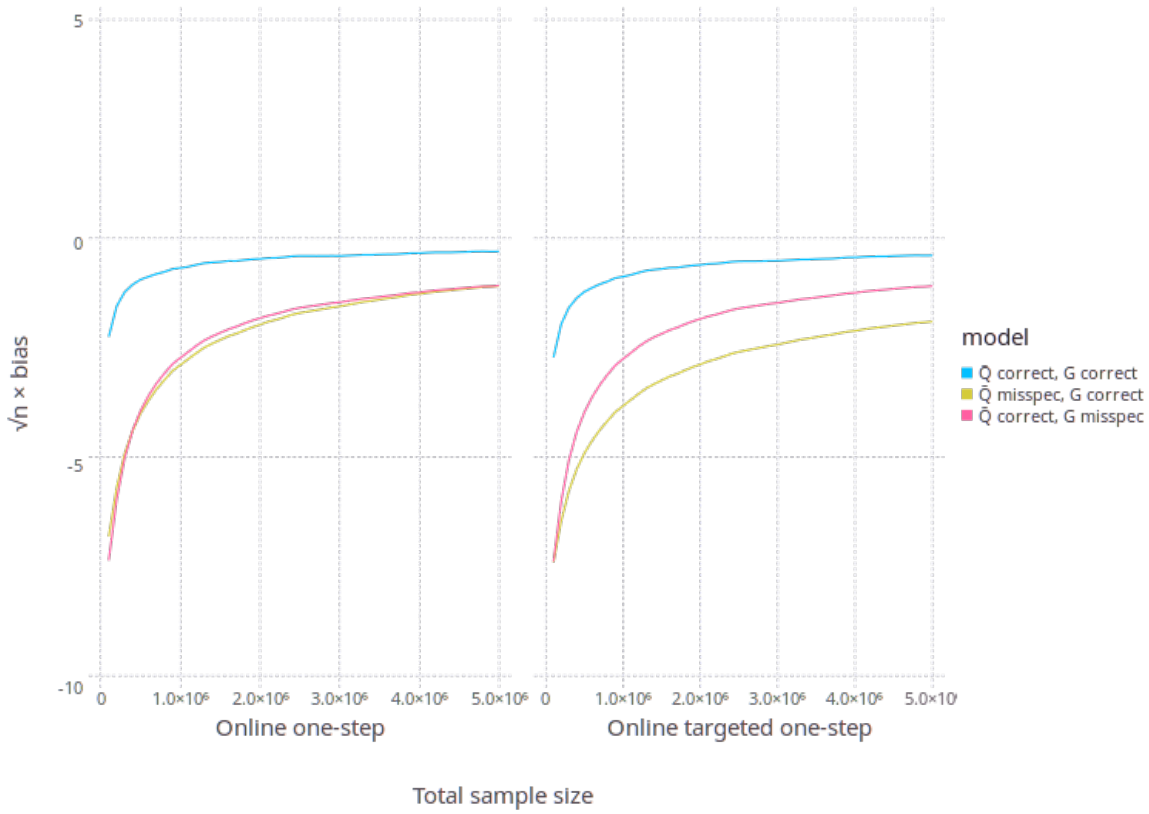

Figure 4.3: Simulation 2, bias scaled by $\sqrt{n_k}$ for the online one-step and online targeted one-step estimators.

4.7 Discussion

In this article, we introduce some asymptotically efficient and online scalable methods for estimating a pathwise differentiable parameter in a single pass through a data set. We describe particular implementations in an example where we estimate the average treatment effect.

In the simulations in Section 4.6, we use stochastic gradient descent to fit main terms a main terms generalized linear model (GLMs) estimators for $\bar{Q}_0$ and G_0. Because we know how the data are generated, simple estimators for $\bar{Q}_0$ and G_0 are sufficient, but this will usually not be the case. In order to consistently estimate the targeted parameter ψ_0, we need to consider general online initial estimators.

A main terms GLM fit with SGD can be extended by using more flexible basis functions, and we can easily add $\ell 1$ or $\ell 2$ regularization terms to the objective function (Bottou, 2010). SGD optimization methods are also applicable to other machine-learing estimators, a popular

example being multi-layer neural networks which can be very flexible (LeCun et al., 1998). Neural networks are frequently trained using SGD though typically one still makes many passes over the data set. Despite this, relatively simple neural networks may be useful when trained in a single pass.

Recently, other online estimators have been developed that are not necessarily based on SGD. Examples include generalized additive models (Simon N Wood, Goude, and Shaw, 2014), online boosting (Oza, 2005), random forests (Abdulsalam, Skillicorn, and Martin, 2007; Saffari et al., 2009), and bayesian semiparametric regression (Luts, Broderick, and Wand, 2014).

There is not a clear way to choose between different initial estimators or even tuning parameters for a single estimator. In the batch setting, one approach to this problem is to use cross-validation to choose one or a combination of estimators with a model stacking algorithm such as the super learner algorithm (Mark J van der Laan, Polley, and Alan E Hubbard, 2007). An area of future research is to extend this approach to the online setting where each mini-batch is used to estimate the out-of-sample risk before updating an estimator on a mini-batch. Multiple estimators could then be fit concurrently, and a combination of candidate estimators could be used as initial estimators for online one-step or online-targeted onestep estimators.

In Section 4.4, we note that if the online efficient influence curve equation is sufficiently small, we can use a mean of plug-in estimators as an estimate of ψ_0 and avoid the additional step in the direction of the influence curve, but we do not discuss a way to guarantee that this is the case. Investigating this further by implementing some of the ideas of Mark J van der Laan and Lendle (2014) and evaluating the performance in practice is an area of future work.

Appendix A.

In this appendix we prove Theorem 6 from Section 4.3.

Proof. We have that $(\bar{M}(k) : k = 1, \ldots)$ is a discrete martingale w.r.t. $\mathcal{F}_k = (\mathbf{O}_1, \ldots, \mathbf{O}_k)$: that is, $E_0(\bar{M}(K) \mid \mathcal{F}(k)) = \bar{M}(k)$ for $k \leq K$. Define

$$W^2(K) = \frac{1}{K} \sum_{k=1}^{K} E_0(M_k^2 \mid \mathcal{F}_{k-1}) = \frac{1}{K} \sum_{k=1}^{K} P_{0,k} M_k^2$$

By assumption, $W^2(K) - \Sigma^2(K) \to_{K \to \infty} 0$ in probability.

We have

$$\begin{aligned} \psi_K &= \frac{1}{K} \sum_{k=1}^{K} \{\Psi(Q_{k-1}) + D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k) - P_{0,k} D_k^*(Q_{k-1}, G_{k-1}))\} \\ &\quad + \frac{1}{K} \sum_{k=1}^{K} P_{0,k} D_k^*(Q_{k-1}, G_{k-1}). \end{aligned}$$

By the identity (4.1), we can write

$$P_{0,k}D_k^*(Q_{k-1},G_{k-1}) = \Psi(Q_0) - \Psi(Q_{k-1}) + R(Q_{k-1},G_{k-1},Q_0,G_0),\ k = 1,\ldots,K,$$

Then we have

$$\frac{1}{K}\sum_{k=1}^{K} P_{0,k}D_k^*(Q_{k-1},G_{k-1}) = \psi_0 - \frac{1}{K}\sum_{k=1}^{K}\Psi(Q_{k-1}) + \frac{1}{K}\sum_{k=1}^{K} R_0(Q_{k-1},G_{k-1},Q_0,G_0).$$

Substitution of this in the last expression yields now

$$\psi_K - \psi_0 = \frac{\bar{M}(K)}{K} + \bar{R}(K),$$

where

$$\bar{M}(K) = \sum_{k=1}^{K}\left\{D_k^*(Q_{k-1},G_{k-1})(\mathbf{O}_k) - P_{0,k}D_k^*(Q_{k-1},G_{k-1}))\right\}.$$

We assumed that $\bar{R}(K) = o_P(1/\sqrt{K})$ (or equivalently, $\bar{R}(K) = o_P(1/\sqrt{n})$). We now note that $\bar{M}(K) = \sum_{k=1}^{K} M_k$, where $E_0(M_k \mid \mathbf{O}_1,\ldots,\mathbf{O}_{k-1}) = 0$. Thus, $E_0(\bar{M}(K) \mid \mathbf{O}_1,\ldots,\mathbf{O}_k) = \bar{M}(k)$, which proves that $(\bar{M}(k) : k)$ is a discrete martingale process. Application of Theorem 7, presented below, to $\bar{M}(K)$ establishes the conclusions of Theorem 6, and, in particular, $\frac{\bar{M}(K)}{\sqrt{K}}$ converges to $N(0,\Sigma^2)$. Finally, the fact that $\Sigma^2/m = P_0D^*(Q_0,G_0)^2$ is easily verified. The consistency of the estimator of $\Sigma^2(K)$ is a consequence of Sen and J. Singer (1993), formally presented by Theorem 8 below. □

The proof relies on establishing weak convergence of the process $(\bar{M}(K)/\sqrt{K} : K)$ as $K \to \infty$. For that purpose we apply a central limit theorem for discrete martingales. An example of such a theorem is given in Sen and J. Singer (1993), resulting in Theorem 17 in Mark J. van der Laan (2008). In our context this Theorem 17 translates into the following one.

Theorem 7. *Let* $\bar{M}(K) = \sum_{k=1}^{K} M_k$, $M_k = (M_{k1},\ldots,M_{kd})$, $E_0(M_k \mid \mathcal{F}_{k-1}) = 0$, *where* $\mathcal{F}_k = (\mathbf{O}_1,\ldots,\mathbf{O}_k)$. *In our case,* $M_k = D_k^*(Q_{k-1},G_{k-1})(\mathbf{O}_k) - P_{0,k}D_k^*(Q_{k-1},G_{k-1})$.
***D**efinitions: Let*

$$\Sigma_k^2 \equiv E_0M_k^2 \equiv E_0M_kM_k^\top,$$

and

$$V_k^2 \equiv E_0\left(M_k^2 \mid \mathcal{F}_{k-1}\right) = P_{0,k}M_k^2.$$

Let

$$\Sigma^2(K) \equiv \frac{1}{K}\sum_{k=1}^{K}\Sigma_k^2 = E_0\frac{1}{K}\sum_{k=1}^{K}P_{0,k}M_k^2$$

and

$$W^2(K) \equiv \frac{1}{K}\sum_{k=1}^{K}V_k^2 = \frac{1}{K}\sum_{k=1}^{K}P_{0,k}M_k^2.$$

***A**ssumptions: Assume that for some $M < \infty$ $\max_k \mid D_k^*(Q_{k-1}, G_{k-1})(\mathbf{O}_k) \mid M < \infty$ with probability 1; $\liminf \lambda\Sigma(k)^2\lambda > 0$ for all λ (or that $\Sigma^2 = \lim_{k\to\infty} \Sigma(k)^2$ exists and is a positive definite covariance matrix); and that component wise*

$$\frac{1}{K}\sum_{k=1}^{K} P_{0,k}D_k^*(Q_{k-1}, G_{k-1}) - E_0\frac{1}{K}\sum_{k=1}^{K} P_{0,k}D_k^*(Q_{k-1}, G_{k-1}) \to 0 \tag{4.7}$$

in probability as $K \to \infty$.
***C**onclusion: Then,*

$$\Sigma(K)^{-1}\frac{\bar{M}(K)}{\sqrt{K}} \Rightarrow_D N(0, I), \text{ as } K \to \infty,$$

and, if $\Sigma^2(K) \to \Sigma^2$, as $K \to \infty$, for some positive definite covariance matrix Σ^2, then

$$\frac{\bar{M}(K)}{\sqrt{K}} \Rightarrow_D N(0, \Sigma^2), \text{ as } K \to \infty.$$

Theorem 8. *Under the conditions stated in Theorem 7, we have that*

$$\hat{\Sigma}^2(K) - \Sigma(K)^2 \to 0 \text{ in probability, as } K \to \infty,$$

and, if $\Sigma^2(K) \to \Sigma^2$, as $K \to \infty$, for a positive definite matrix Σ^2, then this also implies $\hat{\Sigma}^2(K) \to \Sigma$ in probability, as $K \to \infty$.

Bibliography

Abadie, A. and G.W. Imbens (2011). "Bias-corrected matching estimators for average treatment effects". In: *Journal of Business & Economic Statistics* 29.1, pp. 1–11.

Abdulsalam, Hanady, David B Skillicorn, and Patrick Martin (2007). "Streaming random forests". In: *Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th International*. IEEE, pp. 225–232.

Albert, J.M. (2008). "Mediation analysis via potential outcomes models". In: *Statistics in Medicine* 27.8, pp. 1282–1304.

Albert, J.M. and S. Nelson (2011). "Generalized causal mediation analysis". In: *Biometrics* 67.3, pp. 1028–1038.

Anton, R.F. et al. (2006). "Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence". In: *JAMA* 295.17, pp. 2003–2017.

Austin, Peter C (2010). "The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies." In: *Statistics in medicine* 29.20, pp. 2137–48. ISSN: 1097-0258. DOI: 10.1002/sim.3854.

Avin, C., I. Shpitser, and J. Pearl (2005). "Identifiability of Path-Specific Effects". In: *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 357–363.

Bickel, Peter J. et al. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press. ISBN: 0801845416.

Bottou, Léon (2010). "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010*. Springer, pp. 177–186.

— (2012). "Stochastic gradient descent tricks". In: *Neural Networks: Tricks of the Trade*. Springer, pp. 421–436.

Bullock, John G, Donald P Green, and Shang E Ha (2010). "Yes, but what's the mechanism? (don't expect an easy answer)." In: *Journal of personality and social psychology* 98.4, pp. 550–8. ISSN: 1939-1315.

Caliendo, M. and S. Kopeinig (2008). "Some practical guidance for the implementation of propensity score matching". In: *Journal of economic surveys* 22.1, pp. 31–72.

Cole, S.R. and M.A. Hernán (2008). "Constructing inverse probability weights for marginal structural models". In: *American journal of epidemiology* 168.6, pp. 656–664.

Dehejia, R.H. and S. Wahba (2002). "Propensity score-matching methods for nonexperimental causal studies". In: *Review of Economics and statistics* 84.1, pp. 151–161.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization". In: *The Journal of Machine Learning Research* 999999, pp. 2121–2159.

Friedman, J, T Hastie, and R Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1.

Gill, R.D., J.A. Wellner, and J. Præ stgaard (1989). "Non-and semi-parametric maximum likelihood estimators and the von Mises method (part 1)". In: *Scandinavian Journal of Statistics*, pp. 97–128.

Goetgeluk, S., S. Vansteelandt, and E. Goetghebeur (2008). "Estimation of controlled direct effects". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5, pp. 1049–1066.

Gruber, Susan and Mark J van der Laan (2010). "A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome". In: *The International Journal of Biostatistics* 6.1.

Hafeman, Danella M and Tyler J Vanderweele (2011). "Alternative assumptions for the identification of direct and indirect effects." In: *Epidemiology (Cambridge, Mass.)* 22.6, pp. 753–764. ISSN: 1531-5487. DOI: 10.1097/EDE.0b013e3181c311b2.

Hahn, Jinyong (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica* 66.2, pp. 315–331. ISSN: 0012-9682. DOI: 10.2307/2998560.

Holland, P.W. (1986). "Statistics and Causal Inference". In: *J Am Stat Assoc* 81.396, pp. 945–960.

Hubbard, Alan E., Nicholas P. Jewell, and Mark J. van der Laan (2011). "Targeted Learning: Causal Inference for Observational and Experimental Data". In: New York: Springer. Chap. 8.

Imai, K., L. Keele, and D. Tingley (2010). "A General Approach to Causal Mediation Analysis". In: *Psychological Methods* 15.4, pp. 309–334.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010). "Identification, inference and sensitivity analysis for causal mediation effects". In: *Statistical Science* 25.1, pp. 51–71. ISSN: 0883-4237. DOI: 10.1214/10-STS321.

Imai, Kosuke and David A Van Dyk (2004). "Causal inference with general treatment regimes". In: *Journal of the American Statistical Association* 99.467.

Jo, Booil et al. (2011). "The use of propensity scores in mediation analysis". In: *Multivariate Behavioral Research* 46.3, pp. 425–452.

Kang, J.D.Y. and J.L. Schafer (2007). "Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statistical science* 22.4, pp. 523–539.

LeCun, Yann et al. (1998). "Efficient BackProp". In: *Neural Networks: Tricks of the Trade*, pp. 546–546.

Lendle, Samuel D, Bruce Fireman, and Mark J van der Laan (2015). "Balancing score adjusted targeted minimum loss-based estimation". In: *Journal of Causal Infernece*.

Lendle, Samuel D, Meenakshi S Subbaraman, and Mark J van der Laan (2013). "Identification and efficient estimation of the natural direct effect among the untreated". In: *Biometrics* 69.2, pp. 310–317.

Lunceford, J.K. and M. Davidian (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study". In: *Statistics in medicine* 23.19, pp. 2937–2960.

Luts, J., T. Broderick, and M. P. Wand (2014). "Real-Time Semiparametric Regression". In: *Journal of Computational and Graphical Statistics* 23.3, pp. 589–615.

McCullagh, P. and J.A. Nelder (1989). *Generalized linear models*. Vol. 37. Chapman & Hall/CRC.

Murata, Noboru (1998). "A statistical study of on-line learning". In: *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK*.

Neyman, J. (1990). "On the application of probability theory to agricultural experiments". In: *Statistical Science* 5, pp. 465–480.

Oza, Nikunj C (2005). "Online bagging and boosting". In: *Systems, man and cybernetics, 2005 IEEE international conference on*. Vol. 3. IEEE, pp. 2340–2345.

Pearl, J. (2009). *Causality: models, reasoning, and inference*. 2nd. New York: Cambridge.

Pearl, Judea (2001). "Direct and indirect effects". In: *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann, pp. 411–420.

— (2011). *The mediation formula: a guide to the assessment of causal pathways in nonlinear models*. Tech. rep. July. UCLA, pp. 1–38. URL: `http://escholarship.org/uc/item/0hz9x8pc.pdf`.

Petersen, M.L., S.E. Sinisi, and M.J. van der Laan (2006). "Estimation of direct causal effects." In: *Epidemiology* 17.3, pp. 276–284.

Polyak, Boris T and Anatoli B Juditsky (1992). "Acceleration of stochastic approximation by averaging". In: *SIAM Journal on Control and Optimization* 30.4, pp. 838–855.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Robins, J M, M a Hernán, and B Brumback (2000). "Marginal structural models and causal inference in epidemiology." In: *Epidemiology* 11.5, pp. 550–60. ISSN: 1044-3983.

Robins, J. and T. Richardson (2010). *Alternative Graphical Causal Models and the Identification of Direct Effect*. Working Paper 100. Center for Statistics and the Social Sciences, University of Washington.

Robins, James M. (1997). "Marginal Structural Models". In: *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, pp. 1–10.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Robins, J.M. (1987a). "A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods". In: *J Chron Dis (40, Supplement)* 2, 139s–161s.

Robins, J.M. (1987b). "Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]". In: *Comput. Math. Appl.* 14.9-12, pp. 923–945. ISSN: 0097-4943.

Robins, J.M. and Sander Greenland (1992). "Identifiability and exchangeability for direct and indirect effects". In: *Epidemiology* 3.2, pp. 143–155. ISSN: 1044-3983.

Rosenbaum, P.R. (1987). "Model-based direct adjustment". In: *Journal of the American Statistical Association* 82.398, pp. 387–394.

Rosenbaum, P.R. and D.B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, p. 41.

— (1984). "Reducing bias in observational studies using subclassification on the propensity score". In: *Journal of the American Statistical Association* 79.387, pp. 516–524.

Rosenblum, Michael and Mark J van der Laan (2010). "Targeted maximum likelihood estimation of the parameter of a marginal structural model." In: *The international journal of biostatistics* 6.2, Article 19. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1238.

Rubin, D.B. (2004). "Direct and Indirect Causal Effects via Potential Outcomes*". In: *Scandinavian Journal of Statistics* 31.2, pp. 161–170.

— (2006). *Matched Sampling for Causal Effects*. Cambridge, MA: Cambridge University Press.

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

Saffari, Amir et al. (2009). "On-line random forests". In: *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, pp. 1393–1400.

Sekhon, Jasjeet S. (2011). "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R". In: *Journal of Statistical Software* 42.7, pp. 1–52. URL: http://www.jstatsoft.org/v42/i07/.

Sen, P.K. and J.M. Singer (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York.

Tchetgen Tchetgen, Eric J and Ilya Shpitser (2011). *Semiparametric theory for causal mediation analysis : efficiency bounds, multiple robustness, and sensitivity analysis*. Working Paper 130. Harvard University Biostatistics Working Paper Series. URL: http://www.bepress.com/harvardbiostat/paper130.

van der Laan, M. J. and J. M. Robins (2003). *Unified methods for censored longitudinal data and causality*. New York: Springer. ISBN: 0387955569.

van der Laan, M. J. and S. Rose (2011). *Targeted learning causal inference for observational and experimental data*. New York: Springer. ISBN: 978-1-4419-9782-1.

van der Laan, Mark J. (2008). *The Construction and Analysis of Adaptive Group Sequential Designs*. Tech. rep. 232, www.bepress.com/ucbbiostat/paper232. University of California, Berkeley.

— (2010). "Targeted Maximum Likelihood Based Causal Inference: Part I". In: *The International Journal of Biostatistics* 6.2. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1211.

van der Laan, Mark J (2010). "Targeted Maximum Likelihood Based Causal Inference: Part II". In: *The international journal of biostatistics* 6.2. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1241.

van der Laan, Mark J and Susan Gruber (2010). "Collaborative double robust targeted maximum likelihood estimation." In: *The international journal of biostatistics* 6.1. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1181.

van der Laan, Mark J and Samuel D Lendle (2014). *Online Targeted Learning*. Working Paper 330. Berkeley, CA: U.C. Berkeley Division of Biostatistics. URL: http://biostats.bepress.com/ucbbiostat/paper330.

van der Laan, Mark J, Eric C Polley, and Alan E Hubbard (2007). "Super learner." In: *Statistical applications in genetics and molecular biology* 6.1. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1309.

van der Laan, Mark J. and Sherri Rose (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer. ISBN: 1441997814.

van der Laan, Mark J. and Daniel Rubin (2006). "Targeted Maximum Likelihood Learning". In: *The International Journal of Biostatistics* 2.1. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1043.

van der Laan, M.J. (2010). *Estimation of causal effects of community based interventions*. Working Paper 268. U.C. Berkeley Division of Biostatistics Working Paper Series. URL: http://www.bepress.com/ucbbiostat/paper268/.

van der Laan, M.J. and M.L. Petersen (2004). *Estimation of direct and indirect causal effects in longitudinal studies*. Working Paper 155. U.C. Berkeley Division of Biostatistics Working Paper Series. URL: http://www.bepress.com/ucbbiostat/paper155/.

— (2008). "Direct effect models". In: *International Journal of Biostatistics* 4.1. DOI: 10.2202/1557-4679.1064.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK New York, NY, USA: Cambridge University Press.

van der Vaart, A. W. and Wellner J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

VanderWeele, Tyler J (2009). "Marginal structural models for the estimation of direct and indirect effects." In: *Epidemiology* 20.1, pp. 18–26.

VanderWeele, Tyler J and Stijn Vansteelandt (2010). "Odds ratios for mediation analysis for a dichotomous outcome." In: *American journal of epidemiology* 172.12, pp. 1339–48. ISSN: 1476-6256.

Vansteelandt, S. (2009). "Estimating Direct Effects in Cohort and Case–Control Studies". In: *Epidemiology* 20.6, p. 851.

Volpicelli, J.R. et al. (1995). "Effect of naltrexone on alcohol "high" in alcoholics." In: *The American Journal of Psychiatry; The American Journal of Psychiatry* 152.4.

Wood, Simon N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1, pp. 3–36. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2010.00749.x.

Wood, Simon N, Yannig Goude, and Simon Shaw (2014). "Generalized additive models for large data sets". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Xu, Wei (2011). "Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent". In: *CoRR* abs/1107.2490. URL: http://arxiv.org/abs/1107.2490.

Zeiler, Matthew D (2012). "ADADELTA: An adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701*.

Zheng, Wenjing and Mark J van der Laan (2010). *Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation*. Working Paper 273. U.C. Berkeley Division of Biostatistics Working Paper Series. URL: http://www.bepress.com/ucbbiostat/paper273/.

— (2012). "Targeted maximum likelihood estimation of natural direct effects." In: *The international Journal of Biostatistics* 8.1.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.