University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2011 *Paper* 288

Targeted Maximum Likelihood Estimation of Natural Direct Effect

Wenjing Zheng* Mark J. van der Laan†

*Division of Biostatistics, University of California, Berkeley, wenjing.zheng@ucsf.edu

†Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper288

Targeted Maximum Likelihood Estimation of Natural Direct Effect

Wenjing Zheng and Mark J. van der Laan

Abstract

In many causal inference problems, one is interested in the direct causal effect of an exposure on an outcome of interest that is not mediated by certain intermediate variables. Robins and Greenland (1992) and Pearl (2000) formalized the definition of two types of direct effects (natural and controlled) under the counterfactual framework. Since then, identifiability conditions for these effects have been studied extensively. By contrast, considerably fewer efforts have been invested in the estimation problem of the natural direct effect. In this article, we propose a semiparametric efficient, multiply robust estimator for the natural direct effect of a binary treatment using the targeted maximum likelihood framework of van der Laan and Rubin (2006) and van der Laan and Rose (2011). The proposed estimator is asymptotically unbiased if either one of the following holds: i) the conditional outcome expectation given exposure, mediator, and confounders, and the mediated mean outcome difference are consistently estimated; (ii) the exposure mechanism given confounders, and the conditional outcome expectation are consistently estimated; or (iii) the exposure mechanism given confounders, and a ratio of conditional mediator densities are consistently estimated. Moreover, case (iii) implies in particular that estimation of the conditional mediator density may be replaced by consistent estimation of the exposure mechanism and the conditional distribution of exposure given confounders and mediator. If all three conditions hold, then the effect estimate is asymptotically efficient.

1 Introduction

The causal effect of an exposure (or *treatment*) on an outcome of interest is often times mediated by intermediate variables (*mediator*). In many causal inference problems, one is interested in the *direct* effect of such exposure on the outcome, not mediated by the effect of the intermediate variables. Robins and Greenland (1992) and Pearl (2000) defined two types of direct effects under the counterfactual framework. The *controlled* direct effect refers to the effect of the exposure on the outcome under an idealized experiment where the mediator is set to a given constant value, whereas the *natural* (or *pure*) direct effect pertains to an experiment where the mediator is set to be distributed according to the null exposure level and the individual's covariates. The definition of these causal effects are based on counterfactual outcomes that are not fully observed, therefore they are not always identifiable from the observed data. Identifiability conditions are studied extensively in Robins and Greenland (1992), Pearl (2000), Robins (2003), van der Laan and Petersen (2004), Hafeman and VanderWeele (2010), Imai et al. (2010), and Pearl (2011).

Prior to the formal frameworks developed by Robins and Greenland (1992) and Pearl (2000), the social science literature had proposed the use of parametric linear structural equations in mediation analysis (e.g. Baron and Kenny (1986)), where the outcome response and mediator response are each modeled using linear main term regression on their parent nodes, and the direct and indirect effects are defined and estimated in terms of coefficients in these regression equations. The limited causal validity of this parameter due to its dependence on model specification (e.g. no-interactions and linearity assumptions) is discussed in Kaufman et al. (2004). The developments of Robins and Greenland (1992) and Pearl (2000), and the identifiability studies that followed suit, address definition and identification of direct and indirect effects in a framework that is detached from statistical model specifications, allowing one to separate the identification problem from the estimation problem.

Several approaches to the estimation problem are available in the current literature. A likelihood-based estimator approach (the g-computation formula) builds upon the identifiability results using a substitution estimator plugging in maximum likelihood based estimates of the relevant components of the data generating distribution. The natural direct effect may be identified as a function of the marginal covariate distribution, the conditional mediator distribution, conditioned on null exposure and individual covariates, and the conditional outcome distribution, conditioned on exposure, mediator and individual covariates (Robins and Greenland (1992), Pearl (2000), Robins (2003) and van der Laan and Petersen (2004)). When all of these components of the data generating distribution are estimated consistently, the resulting g-computation estimate is unbiased and efficient. However, if either of these components is inconsistent, the effect estimate will be biased. VanderWeele and Vansteelandt (2010) illustrated how this approach may be applied to the estimation of natural direct effect odds ratio of rare outcomes. The use of (sequential) g-computation in structural nested models for estimation of controlled direct effects is proposed in Vansteelandt (2009). A second approach to causal effects estimation is based on the estimating function methodology developed by Robins (1999), Robins

and Rotnitzky (2001) and van der Laan and Robins (2003), where the root to a score equation is used as the effect estimate. For most parameters arising from missing data problems (including causal effect parameters), the efficient score under a nonparametric model is a robust estimating function (i.e. unbiased against mis-specification of the missingness mechanism or mis-specification of the full data model), therefore the resulting effect estimate shares the same robustness properties. In van der Laan and Petersen (2008), an application of this approach to a generalized class of direct effects using marginal structural models was discussed. The parameter studied in that work is a population mean of a subject-specific average controlled direct effect, averaged with respect to a user-specified conditional mediator density given null exposure and individual covariates. If the supplied conditional mediator density is the true conditional mediator density of the data generating process, then the parameter of van der Laan and Petersen (2008) evaluates to the same value as the natural direct effect parameter. However, even in such case, these two parameters are not the same maps on the model since the former is a map indexed by the supplied mediator density and is a function of the outcome expectation and marginal covariate distribution alone. As a consequence, the efficient score of the parameter of van der Laan and Petersen (2008) is not the same as the efficient score of the natural direct effect parameter we study in this article. VanderWeele (2009) discussed more fully the use of marginal structural models with inverse probability weighting for estimation of the natural direct effect parameter. Most recently, Tchetgen Tchetgen and Shpitser (2011) developed the application of the estimating function methodology to natural direct effect estimation using the efficient score equation, as well as a sensitivity analysis framework for the assumption of ignorability of the mediator variable. We also refer the interested reader to their work for discussion on semiparametric efficiency bounds for the nonparametric model. A third approach to causal effect estimation is the targeted maximum likelihood framework of van der Laan and Rubin (2006) and van der Laan and Rose (2011). For each relevant component of the data generating distribution, one obtains a loss-based estimate that would solve the corresponding component of the efficient score equation. These estimates are then used to obtain a substitution estimator of the parameter of interest. The resulting estimator solves the efficient score equation, therefore also shares its robustness properties. In addition, the substitution principle allows for estimation of the parameter range providing additional information gain, and preserves properties of the parameter as a map on the model. van der Laan and Petersen (2008) also applied the targeted MLE procedure to their generalized class of direct effect parameters. Both the estimating function approach and the targeted MLE approach in van der Laan and Petersen (2008) are robust (with respect to its parameter of interest) against mis-specification of the conditional expected outcome or mis-specification of the treatment mechanism. However, since its parameter of interest is indexed by the user-supplied conditional mediator density, if one is interested in the natural direct effect, then the user-specified conditional mediator density in the method of van der Laan and Petersen (2008) must be correct. The use of propensity score matching in estimation of causal effects from observational studies was introduced in Rosenbaum and Rubin (1983). Application of propensity score in mediation analysis has also been proposed (e.g. Jo et al. (2011)).

In this article, we apply the targeted MLE framework of van der Laan and Rubin (2006) and van der Laan and Rose (2011) to the estimation of the natural direct effect of a binary exposure. The identifiability results in Robins and Greenland (1992), Pearl (2000), Robins (2003) and van der Laan and Petersen (2004) imply in particular that the natural direct effect of a binary treatment may be estimated as the marginal mean (over strata of confounders) of the mediated mean outcome difference, where the mediated mean outcome difference is the conditional expectation of the difference in outcome under two different exposure levels, conditioned with respect to the mediator given null exposure and confounders. We propose a semiparametric efficient and robust estimator which, given initial estimators of the exposure mechanism, conditional mediator density and conditional outcome expectation, targetedly modifies the estimates of the conditional outcome expectation and the mediated mean outcome difference using a set of parametric working submodels. These resulting targeted components are then used to produce a plug-in estimator for the parameter of interest. The procedure systematically incorporates estimation of the boundary of the parameter domain. The set of parametric working submodels are defined such that the resulting estimator solve the efficient score equation, and hence inherits its robustness properties. The proposed estimator is asymptotically unbiased if either one of the following holds: i) the conditional outcome expectation, and the mediated mean outcome difference are consistently estimated; (ii) the exposure mechanism given confounders, and the conditional outcome expectation are consistently estimated; or (iii) the exposure mechanism given confounders, and a ratio of conditional mediator densities are consistently estimated. If all three conditions hold, then the effect estimate is asymptotically efficient.

This article is organized as follows: In section 2 we define formally the natural direct causal effect of a binary treatment on an outcome using the Non-Parametric Structural Equations Model framework of Pearl (2009), and summarize its identifiability conditions. Based on the identifiability result, one may consider the natural direct effect parameter as a map from the model to the parameter space. We study this map in greater detail in section 2.3. In particular, the robustness properties of its efficient score under a nonparametric model are summarized in lemma 1 of that section. Section 3 begins with a general description of the targeted MLE estimation framework of van der Laan and Rubin (2006), and then presents a step by step construction of the targeted MLE estimator for the natural direct effect of a binary treatment. Asymptotic properties of this estimator are summarized in section 3.2 and proved in the Appendix A. The estimation procedure in section 3 focuses on the targeted estimation of the conditional outcome expectation and the mediated mean outcome difference, as described above. An alternative procedure focusing on the conditional outcome expectation and the conditional mediator density is described in Appendix B. This alternative estimator shares the same asymptotic properties as the one proposed in section 3. Section 4 describes in greater detail two alternative estimators under the estimation equation framework of Robins (1999), and the maximum likelihood based g-computation framework. In section 5, we illustrate with simulations the robustness of the targeted MLE estimator against model mis-specifications. We will also explore the performance of the various estimators in the presence of data

sparsity. This article concludes with a summary.

2 Natural Direct Effect of a binary treatment

2.1 Causal Parameter

Consider n i.i.d observations of $O = (W, A, Z, Y)$, where W represents baseline covariates, A a binary treatment, Z represents the mediator between the treatment and the outcome of interest Y. Let P_0 denote the distribution of O. We apply here the Non-Parametric Structural Equations Model of Pearl (2009) to encode the causal relations of interest. The NPSEM on an unit consists of a set of exogenous random variables U which are determined by factors outside the model, a set of endogenous variables X which are determined by variables inside the system $(U \cup X)$, and a set of unspecified deterministic functions $\{f_x : x \in X\}$ which encodes for each $x \in X$ the variables that have direct influence on x. More specifically, in the present situation the causal relations are encoded by the NPSEM

$$\begin{aligned} U &= (U_W, U_A, U_Z, U_Y) \sim P_U \\ W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Z &= f_Z(W, A, U_Z) \\ Y &= f_Y(W, A, Z, U_Y), \end{aligned}$$

where $X = (W, A, Z, Y)$ is the endogenous variable, and $U = (U_W, U_A, U_Z, U_Y)$ is the unobserved exogenous variable. This model defines a random variable (U, X) on the unit of observation, we denote its distribution by $P_{U,X}$.

One may define a submodel of the NPSEM by intervening on a subset of the equations. The counterfactual variables or potential outcomes in the Rubin Causal Model (Rubin (1978), Rosenbaum and Rubin (1983) and Holland (1986)) may then be interpreted as variables in the post-intervention submodel. For instance, the counterfactual $Z(a)$ is defined as the random variable Z in a system where one intervened to set $Z = f_Z(W, a, U_Z)$, and may be interpreted as the mediator variable that the unit would have had if the exposure had been a. Similarly, $Y(a', Z(a))$ is the counterfactual outcome that results from setting $Y = f_Y(W, a', Z(a), U_Y)$, and may be interpreted as the response that one may have had if the exposure had been a' while the mediator variable had been identical to the one under exposure a.

Under the NPSEM, a causal parameter of interest may be defined as a function of the distribution $P_{U,X}$. More specifically, the *natural direct causal effect* is defined as

$$\Psi(P_{U,X}) = E\left[Y(1, Z(0)) - Y(0, Z(0))\right].$$

This causal parameter may be interpreted from the following hypothetical randomized trial: one randomly assigns each subject to treatment or control, while setting the subject's mediator variable to be distributed as if treatment was absent, and then take the difference in mean outcome between the treated and control cohort.

2.2 Identifiability

Under experimental or observational studies, for each unit the investigator only observes the outcome and mediating response under the unit's actual exposure, that is, $O = (W, A, Z(A), Y(A, Z(A))$. Hence, the causal parameter $\Psi(P_{U,X})$ is not always identifiable from the observed data.

Conditions under which the natural direct effect will be identifiable are addressed extensively in Robins and Greenland (1992), Pearl (2000), Robins (2003), van der Laan and Petersen (2004), Hafeman and VanderWeele (2010), Imai et al. (2010), and Pearl (2011). In particular, if the randomization assumptions

1. For all values (a, z), (A, Z) is independent of $Y(a, z)$, given W;

2. For all values of a, A is independent of $Z(a)$, given W;

and the conditional independence assumption

3. For all z, $E\left(Y(1, z) - Y(0, z)|Z(0) = z, W\right) = E\left(Y(1, z) - Y(0, z)|W\right)$

are satisfied, then the causal effect $\Psi(P_{U,X})$ may be expressed as a function of the observed data generating distribution P_0:

$$\Psi(P_0) = E_W \left\{ \sum_z \left[E(Y|W, A = 1, Z = z) - E(Y|W, A = 0, Z = z)\right] p(z|W, A = 0) \right\}. \quad (1)$$

In the following sections, we will focus on the estimation of this statistical parameter.

The randomization assumptions 1 and 2 ensure that sufficient covariates are measured to control for confounding of the effects of treatment on outcome, treatment on mediator, and mediator on outcome. As a result, the counterfactual elements $Y(a, z)$ and $Z(a)$ will be identifiable within covariate stratum. Under these randomization assumptions alone, the statistical parameter (1) equals the population mean of a subject-specific average controlled direct effect $\sum_z (Y(1, z) - Y(0, z))P(Z(0) = z|W)$ (van der Laan and Petersen (2008)). Therefore, in the absence of the conditional independence assumption 3 the statistical parameter (1) still offers a causal interpretation.

2.3 The Natural Direct Effect parameter

Let $\mathcal{M}$ denote a model containing the true data generating distribution P_0. For any $P \in \mathcal{M}$, the likelihood decomposes into

$$P(O) = P_W(W)P_A(A|W)P_Z(Z|W, A)P_Y(Y|W, A, Z).$$

For later convenience, we adopt the notations $g(A|W, Z) = P_A(A|W, Z)$, $Q_W(W) = P_W(W)$, $Q_Z(Z|W, A) = P_Z(Z|W, A)$, and $\bar{Q}_Y(W, A, Z) = E(Y|W, A, Z)$. Moreover, let $Q = (Q_W, Q_Z, \bar{Q}_Y)$. The notations Q_0 and g_0 are reserved for the corresponding components of the true data generating distribution P_0. For a function $f(O)$, we will use Pf to denote the expectation of $f(O)$ under the probability distribution $P \in \mathcal{M}$.

For instance, $P_0 f \equiv \sum_{o \in \mathcal{O}} f(o) dP_0(o)$ denotes the expectation of f under the true data generating distribution, while $P_n f \equiv \frac{1}{n} \sum_{i=1}^{n} f(o_i)$ is the empirical mean of f.

One may consider the natural direct effect parameter Ψ as a map

$$\begin{aligned} \Psi: \quad \mathcal{M} &\to \quad \mathbb{R} \\ P &\mapsto \quad \Psi(P) = \Psi(Q) \equiv E_{Q_W}\left[E_{Q_Z}\left(\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)|W, A=0\right)\right]. \end{aligned}$$

The parameter of interest in (1) is thus

$$\psi_0 \equiv \Psi(P_0) = E_{Q_{W,0}}\left[E_{Q_{Z,0}}\left(\bar{Q}_{Y,0}(W,1,Z) - \bar{Q}_{Y,0}(W,0,Z)|W, A=0\right)\right].$$

We refer to the inner expectation above as the *(null level) mediated mean outcome difference*, and denote it by

$$E_{Q_Z}(\bar{Q}_Y|W,0) \equiv \sum_z \left(\bar{Q}_Y(W,1,z) - \bar{Q}_Y(W,0,z)\right) Q_Z(z|W,0). \tag{2}$$

This mediated difference is a function of W alone. For convenience, we may abuse the notation $E_{Q_Z}(\bar{Q}_Y) \equiv E_{Q_Z}(\bar{Q}_Y|W,0)$ when referring to the function of W. This way, $\Psi(P) = \Psi\left(Q_W, E_{Q_Z}(\bar{Q}_Y)\right)$.

Effcient score

Under a nonparametric model $\mathcal{M}$, for any $P \in \mathcal{M}$, the *efficient score* (*efficient influence curve*, or *canonical gradient*) of Ψ at P is given by

$$\begin{aligned} D^*(Q,g,\Psi(Q)) &= \left\{\frac{I(A=1)}{g(1|W)}\frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)} - \frac{I(A=0)}{g(0|W)}\right\}\left(Y - \bar{Q}_Y(W,A,Z)\right) \\ &+ \frac{I(A=0)}{g(0|W)}\left\{\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z) - E_{Q_Z}\left(\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)|W,0\right)\right\} \\ &+ E_{Q_Z}\left(\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)|W,0\right) - \Psi(Q) \\ &= D_Y^* + D_Z^* + D_W^*. \end{aligned}$$

Note that the components D_Y^*, D_Z^*, D_W^* are respectively the projection of D^* onto the tangent subspaces corresponding to the components $P(Y|W,A,Z)$, $P(Z|W,A)$, $P(W)$ of the likelihood.

This efficient score for a nonparametric model may be derived by first considering $\Psi(P)$ as a function of $P = (Pf : f \in \mathcal{F})$, where $\mathcal{F}$ is a class of indicator functions $\mathcal{F} = \{I(w,a,z,y), I(w,a,z), I(w,a), I(w) : w \in \mathcal{W}, a \in \{0,1\}, z \in \mathcal{Z}, y \in \mathcal{Y}\}$. For any given "vector" $h = (h(f) : f \in \mathcal{F})$, one may consider a directional derivative $\frac{d}{d\epsilon}\Psi(P + \epsilon h)|_{\epsilon=0}$. The efficient score is then given by the directional derivative applied to the direction of $h = (f(O) - Pf : f \in \mathcal{F})$. In other words, it is given by $\sum_{f \in \mathcal{F}} \frac{\partial \Psi(P)}{\partial P_f}(f(O) - Pf)$. A more detail exposition may be found in van der Laan and Rose (2011).

Lemma 1. ***Robustness of the efficient score***
Suppose there exists $1 > \delta > 0$ *such that* $g(A=1|W) < 1 - \delta$ *a.e. over the support of* W. *The efficient score is a robust estimating function for the parameter at* P_0, *in the sense that*

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

if either of the following holds:

(i) The conditional outcome expectation $\bar{Q}_Y = E(Y|W,A,Z)$, and the mediated mean outcome difference $E_{Q_Z}\left(\bar{Q}_{Y,0}(W,1,Z) - \bar{Q}_{Y,0}(W,0,Z) \mid W,0\right)$ are correct.

(ii) The treatment mechanism $g = p(A|W)$, and the conditional outcome expectation $\bar{Q}_Y = E(Y|W,A,Z)$ are correct.

(iii) The treatment mechanism $g = p(A|W)$, and the conditional mediator density ratio $Q_Z(Z|W,0)/Q_Z(Z|W,1)$ are correct.

(iv) The treatment mechanism $g = p(A|W)$, and the conditional distribution of treatment given mediator and covariates $p(A|W,Z)$ are correct.

The proof of this lemma is straightforward, and we refer the interested reader to Appendix A. A noteworthy observation is that in case (i), it was not necessary that $Q_Z = Q_{Z,0}$. But rather, any function $E_Z(\bar{Q}_{Y,0}|W,0)$ which captures the dependence of the true outcome difference on the null exposure and confounder, and equals the true mediated mean difference $E_{Q_{Z,0}}(\bar{Q}_{Y,0}|W,0)$ will yield the desired result. This suggests that in the case where the outcome expectation can be correctly estimated while the treatment mechanism and the mediator density are difficult to ascertain, one may still obtain unbiasedness using a consistent data-adaptive estimator that regresses the correctly predicted outcome difference on W among the control observations. On the contrary, in cases when g is correct, robustness does not impose any requirement on $E_{Q_Z}(\bar{Q}_Y|W,0)$. In fact, the cancelation in the proof shows that it may be any function of W. We illustrate this last observation in the simulation section (implemented as TMLE 2). Case (iv) is a simple consequence of case (iii). In situations when Z is high dimensional, consistent estimation of $p(A|W,Z)$ may prove more attainable than consistent estimation of $Q_Z(Z|W,A)$.

It is worthwhile to note that in the case where only $\bar{Q}_Y$ is correctly specified, the solution ψ_1 to the equation $P_0 D^*(\bar{Q}_{Y,0}, Q_Z, g, \psi) = 0$ corresponds to an alternative effect parameter of the form

$$\begin{aligned}
\psi_1 &= E_{W,0}\Bigg\{\sum_z \left(\bar{Q}_{Y,0}(W,1,z) - \bar{Q}_{Y,0}(W,0,z)\right) \times \\
&\quad \left\{\left(Q_{Z,0}(z|W,0) - Q_Z(z|W,0)\right)\frac{g_0(0|W)}{g(0|W)} + Q_Z(z|W,0)\right\}\Bigg\} \\
&= \psi_0 \\
&+ E_{W,0}\Bigg\{\sum_z \left(\bar{Q}_{Y,0}(W,1,z) - \bar{Q}_{Y,0}(W,0,z)\right) \times \\
&\quad \left\{\left(Q_{Z,0}(z|W,0) - Q_Z(z|W,0)\right)\left(\frac{g_0(0|W)}{g(0|W)} - 1\right)\right\}\Bigg\}
\end{aligned}$$

3 Targeted MLE for the Natural Direct Effect of a binary treatment

In general, under the framework of van der Laan and Rubin (2006) the construction of a targeted estimator of a parameter of interest $\Psi(P_0)$ calls for two sets of ingredients. For each component Q_j of Q, one defines a uniformly bounded (w.r.t. the supremum norm) *loss function* $L_j : \mathcal{Q}_j \to \mathcal{L}^\infty(K)$ satisfying

$$Q_{j,0} = \arg \min_{Q_j \in \mathcal{Q}_j} P_0 L_j(Q_j),$$

where $\mathcal{L}^\infty(K)$ is the class of functions of O with bounded supremum norm over a set of K containing the support of O under P_0. Given the loss function L_j, one defines a one-dimensional *parametric working submodel* $\{Q_j(Q, g, \epsilon) : \epsilon\} \subset \mathcal{M}$ passing through Q at $\epsilon = 0$ with score $D_j^*(Q, g)$ at $\epsilon = 0$ that satisfies

$$\langle \frac{d}{d\epsilon} L_j \left(Q_j(Q, g, \epsilon) \right) |_{\epsilon=0} \rangle \supset \langle D_j^*(Q, g) \rangle. \tag{3}$$

The acceptable functional forms of the submodel are thus ruled by the loss function L_j and the functional form of D_j^*. The loss functions and their respective parametric submodels satisfying (3) allow for loss-based estimates of each components of Q that would also solve the corresponding component of the efficient score equation.

To specialize to the natural direct effect, we first note that the parameter of interest and the components D_Z^* and D_W^* of the efficient score depend on Q_Z only through the mediated mean outcome difference $E_{Q_Z}(\bar{Q}_Y)$ as denoted in (2). Secondly, the empirical marginal distribution $\hat{Q}_W$ of W is a consistent estimator of $Q_{W,0}$ that readily solves the equation $P_n D_W^*(E_{Q_Z}(\bar{Q}_Y), \hat{Q}_W) = 0$ for any $E_{Q_Z}(\bar{Q}_Y)$. Hence, the proposed estimator will focus on targeted estimation of $\bar{Q}_{Y,0}(W, A, Z)$, and $E_{Q_{Z,0}}(\bar{Q}_{Y,0}|W, 0)$.

An alternative targeted estimation to the one proposed above is to targetedly estimate the conditional mediator density $Q_{Z,0}$ instead of the mediated mean outcome difference $E_{Q_{Z,0}}(\bar{Q}_{Y,0}|W, 0)$. We refer the interested reader to Appendix B for this alternative approach. The proposed and the alternative targeted procedures both require an initial estimator of the conditional mediator density $Q_{Z,0}$. Their key difference lies in that the former defines a loss function and parametric working submodel for the mediated mean outcome difference $E_{Q_Z}(\bar{Q}_Y|W, 0)$, whereas the latter defines a loss function and parametric working submodel for the conditional mediator density Q_Z and then estimates the mediated mean outcome difference plugging in the targeted mediator density and the targeted $\bar{Q}_Y$. The bias-variance trade-off in the targeting step of the first approach is more optimal for estimating the ultimate component of interest, which is the mediated mean outcome difference.

3.1 Construction of the targeted MLE

Loss functions

Suppose Y is binary or continuous and bounded. In the latter case, without loss of generality we may assume that Y is bounded in $(0, 1)$. In this case, a valid loss

function for $\bar{Q}_Y$ is the minus-loglikelihood

$$L_Y(\bar{Q}_Y)(O) = -\left\{Y \log \bar{Q}_Y(W,A,Z) + (1-Y)\log(1-\bar{Q}_Y(W,A,Z))\right\}. \quad (4)$$

For a given $\bar{Q}_Y(\cdot)$, suppose $\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)$ is also bounded. Without loss of generality, we may also assume it's bounded between $(0,1)$. Let the loss function for $E_{Q_Z}(\bar{Q}_Y|W,0)$ be

$$\begin{aligned} &L_Z(E_{Q_Z}(\bar{Q}_Y))(O) = \\ &- I(A=0)\times \\ &\Big\{\bar{Q}_Y(W,A,Z)\log E_{Q_Z}(\bar{Q}_Y|W,0) + (1-\bar{Q}_Y(W,A,Z))\log(1-E_{Q_Z}(\bar{Q}_Y|W,0))\Big\}. \end{aligned} \quad (5)$$

Linear transformations onto the unit interval may be needed in order to use loss functions L_Y and L_Z. However, since the parameter of interest and the components of the efficient score are linear in $\bar{Q}_Y$ and $E_{Q_Z}(\bar{Q}_Y)$, the necessary linear transformations and their inverse maps do not affect the properties of the estimators.

In a more general setting one may instead use the squared error loss functions

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

and

$$L_Z(E_{Q_Z}(\bar{Q}_Y))(O) = \left(\bar{Q}_Y(W,A,Z) - E_{Q_Z}(\bar{Q}_Y|W,0)\right)^2 I(A=0).$$

Parametric working submodels.

Under the loss function (4) for $\bar{Q}_Y$, consider the logistic working submodel

$$\bar{Q}_Y(Q_Z,g)(\epsilon_1) \equiv expit\left\{logit(\bar{Q}_Y) + \epsilon_1 C_Y(Q_Z,g)\right\},$$

where $C_Y(Q_Z,g)(O) = \left\{\frac{I(A=1)}{g(1|W)}\frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)} - \frac{I(A=0)}{g(0|W)}\right\}$. This submodel satisfies

$$\frac{d}{d\epsilon_1} L_Y\left(\bar{Q}_Y(Q_Z,g)(\epsilon_1)\right)|_{\epsilon_1=0} = D_Y^*(\bar{Q}_Y,Q_Z,g). \quad (6)$$

Similarly, for a given $E_{Q_Z}(\bar{Q}_Y|W,0)$, under the loss function (5) the logistic working submodel

$$E_{Q_Z}(\bar{Q}_Y)(g)(\epsilon_2) \equiv expit\left\{logit(E_{Q_Z}(\bar{Q}_Y)) + \epsilon_2 C_Z(g)\right\},$$

with $C_Z(g)(O) = \frac{1}{g(0|W)}$, satisfies

$$\frac{d}{d\epsilon_2} L_Z\left(E_{Q_Z}(\bar{Q}_Y)(g)(\epsilon_2)\right)|_{\epsilon_2=0} = D_Z^*(E_{Q_Z}(\bar{Q}_Y),\bar{Q}_Y,g). \quad (7)$$

In the case where the squared error loss function is used, the parametric working submodels are of the form

$$\bar{Q}_Y(Q_Z,g)(\epsilon_1) = \bar{Q}_Y + \epsilon_1 C_Y(Q_Z,g)$$

and

$$E_{Q_Z}(\bar{Q}_Y)(g)(\epsilon_2) = E_{Q_Z}(\bar{Q}_Y) + \epsilon_2 C_Z(g).$$

Implementation

Let P_n denote the empirical distribution of n i.d.d observations of O. Let $\hat{\bar{Q}}_Y$, $\hat{Q}_Z$, and $\hat{g}$ be respectively initial estimators of $\bar{Q}_{Y,0}$, $Q_{Z,0}$ and g_0. Let

$$\hat{\epsilon_1}^* = \arg\min_{\epsilon} P_n L_Y \left(\hat{\bar{Q}}_Y(\hat{Q}_Z, \hat{g})(\epsilon_1)\right)$$

be the optimal ϵ_1 which minimizes the empirical risk. The update

$$\hat{\bar{Q}}_Y^* \equiv \hat{\bar{Q}}_Y(\hat{Q}_Z, \hat{g})(\hat{\epsilon}^*) \tag{8}$$

is the *targeted MLE estimator of* $\bar{Q}_{Y,0}$.

Next, let $\hat{E}_Z(\hat{\bar{Q}}_Y^*|W, 0)$ be an initial estimator of the mediated mean outcome difference $E_{Q_{Z,0}}\Big(\hat{\bar{Q}}_Y^*(W, 1, Z) - \hat{\bar{Q}}_Y^*(W, 0, Z) \mid W, A = 0\Big)$. This may be constructed using a plug-in estimator with $\hat{Q}_Z$ and $\hat{\bar{Q}}_Y^*$. The optimal ϵ_2 is given by

$$\hat{\epsilon_2}^* = \arg\min_{\epsilon} P_n L_Z \left(\hat{E}_Z(\hat{\bar{Q}}_Y^*)(\hat{g})(\epsilon_2)\right).$$

The update

$$\hat{E}_Z^*(\hat{\bar{Q}}_Y^*) \equiv \hat{E}_Z(\hat{\bar{Q}}_Y^*)(\hat{g})(\epsilon_2^*) \tag{9}$$

is the *targeted MLE estimator of* $E_{Q_{Z,0}}\left(\hat{\bar{Q}}_Y^*(W, 1, Z) - \hat{\bar{Q}}_Y^*(W, 0, Z) \mid W, A = 0\right)$. Moreover, it is a function of W alone, i.e. $\hat{E}_Z^*(\hat{\bar{Q}}_Y^*|W, 0) = \hat{E}_Z^*(\hat{\bar{Q}}_Y^*)(W)$. The *targeted MLE estimator of* ψ_0 is thus given by

$$\hat{\psi}^* = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_Z^*(\hat{\bar{Q}}_Y^*)(W_i). \tag{10}$$

It follows from (6) that $P_n D_Y^*(\hat{\bar{Q}}_Y^*, \hat{Q}_Z, \hat{g}) = 0$ and it follows from (7) that $P_n D_Z^*\left(\hat{E}_Z^*(\hat{\bar{Q}}_Y^*), \hat{\bar{Q}}_Y^*, \hat{g}\right) = 0$. Moreover, the empirical distribution $\hat{Q}_W$ of W solves $P_n D_W^*(\hat{E}_Z^*(\hat{\bar{Q}}_Y^*), \hat{Q}_W) = 0$. Therefore the resulting targeted estimator solves the efficient score equation.

Remarks on implementation

When Z is high dimensional, consistent estimation of $p(A|W, Z)$ may be more attainable than consistent estimation of $Q_Z(Z|W, A)$. In such case, instead of using an estimator of Q_Z to estimate the ratio $Q_Z(Z|W, 0)/Q_Z(Z|W, 1)$ in the targeting step of $\bar{Q}_Y$, one may use an estimator $\frac{\hat{p}(A=0|W,z)}{\hat{g}(A=0|w)}\frac{\hat{g}(A=1|W)}{\hat{p}(A=1|W,z)}$.

In the step of targeting the mediated mean outcome difference, we mentioned the plug-in estimator using the initial mediator density and the targeted outcome predictor as an initial estimator. However, the initial estimator may be any function $E_Z(\hat{\bar{Q}}_Y^*|W, 0)$ of W which regresses the predicted outcome difference given by $\hat{\bar{Q}}_Y^*$ on W among control observations. From lemma 1, we see that when the treatment mechanism is correct, i.e. in cases (ii), (iii), and (iv), the specification of the mediated mean outcome difference per se does not affect robustness (we illustrate this observation using implementation TMLE 2 in the simulations section).

3.2 Asymptotic Properties of the Targeted MLE

Since the proposed targeted MLE estimator solves the efficient score equation, lemma 1 implies in particular that the estimator is asymptotically unbiased if either of the following is true: (i) The conditional outcome expectation $\hat{Q}_Y^*$ and the mediated mean outcome difference $\hat{E}_Z^*(\hat{\bar{Q}}_Y^*|W,0)$ are consistent; (ii) the treatment mechanism $\hat{g}$ and the conditional outcome expectation $\hat{\bar{Q}}_Y^*$ are consistent; (iii) the treatment mechanism g, and the conditional mediator density ratio $Q_Z(Z|W,0)/Q_Z(Z|W,1)$ are consistently estimated. These properties are illustrated in the simulations section below.

Under certain empirical conditions, an estimator that solves an estimating equation will be asymptotically linear with influence curve given by the estimating function (e.g.. van der Vaart (1998), van der Laan and Robins (2003), Tsiatis (2006), Kosorok (2008)). In such case, central limit theorem implies that one may obtain an asymptotic variance estimate of the said estimator using the variance estimate of its influence curve. We detail conditions for asymptotic linearity of the targeted MLE estimator in the theorem 1 in Appendix A.

Empirical process conditions are often necessary for the asymptotic linearity of an Z-estimator as they restrict the size of the class of functions containing the influence curve (we refer to the CV-TMLE framework in Zheng and van der Laan (2010) and Zheng and van der Laan (2011) for an alternative targeted estimator which avoids such conditions through the use of cross-validation). The conclusion (18) of theorem 1 shows that under certain empirical process conditions, the estimator behaves as an empirical mean of mean zero i.i.d. random variables (which converges to a normal distribution by CLT), plus specified second order remainders from which one may infer the conditions needed for asymptotic linearity.

When true treatment mechanism g_0 is used in the estimation procedure (e.g. in an RCT), the remainders concerning estimation of g_0 vanish, leaving only a second order remainder term which concerns the speed at which the targeted outcome expectation estimator $\hat{\bar{Q}}_Y^*$ and the initial mediator density $\hat{Q}_Z$ converge to their respective limits, and two first order remainder terms concerning the difference between these limits and the truth. Asymptotic linearity requires firstly that the second order term be $o_P(1/\sqrt{n})$ (condition (19) of theorem 1). If both $\hat{\bar{Q}}_Y^*$ and $\hat{Q}_Z$ are consistent (and satisfies this speed condition), then the estimator is asymptotically linear. Moreover, if the plug-in initial estimator for the mediated mean outcome difference is used in this case, it follows that the estimator is in fact asymptotically efficient. Otherwise, asymptotic efficiency follows only if the initial estimator of the mediated mean outcome difference is also consistent. In the case that one of the two components, $\bar{Q}_Y$ or Q_Z, is inconsistently estimated, the resulting first order remainder will have to satisfy an asymptotically linear condition ((23) or (21)). This implies in particular that if one uses a data-adaptive estimator for the outcome $\bar{Q}_{Y,0}$, then the estimator $\hat{Q}_{Z,n}$ for the mediator density needs to converge fast enough so that second order condition (19) is satisfied. Even though this compromise may cause the mediator density estimator to be inconsistent, as long as the outcome estimator $\hat{Q}_{Y,n}^*$ is consistent and satisfies the asymptotic linearity condition of (21), then the effect estimate will still

be asymptotically linear. On the other hand, if one chooses to use a data-adaptive estimator for the mediator density, it may come at the expensive of a smaller model for the outcome so that (19) is met. If this smaller model for the outcome is not correct, then the mediator density estimator will need to be consistent and satisfy the asymptotically linear condition of (23).

When the true treatment mechanism g_0 is not used, one is confronted with 3 second order remainders that concern the speeds at which the pairs $\left(\hat{\bar{Q}}^*_Y, \hat{Q}_Z\right)$, $\left(\hat{\bar{Q}}^*_Y, \hat{g}\right)$, and $\left(\hat{g}, \hat{E}^*_Z(\hat{\bar{Q}}^*_Y)\right)$ converge to their respective limits, and $3 \times 2 = 6$ first order remainders that concern the difference between these limits and the truth. Asymptotic linearity requires firstly that the 3 second order remainders are $o_P(1/\sqrt{n})$ (conditions (19), (25) and (26)). This impose restrictions on how large a model one may use to estimate each component. If g_0 is contained in a correctly specified parametric model (e.g. it only depends on a discrete covariate and one uses a saturated model), then rate conditions (25) and (26) are satisfied for reasonable estimators of $\bar{Q}_{Y,0}$ and $Q_{Z,0}$. However, if g_0 is contained in a large semiparametric model, the estimators for outcome and mediator density will need to both converge fast enough so that conditions (25) and (26) are satisfied, which severely restrict their data-adaptiveness. If all the components are consistently estimated, then the effect estimator is asymptotically efficient. However, if one of the components is inconsistent, then one is confronted with two first order remainders and asymptotic linearity conditions on these terms are needed to ensure asymptotic linearity of the resulting effect estimate.

In short, asymptotic linearity requires that (a) estimators of each component converge to their respective limits at a reasonable speed; (b) at most one component may be inconsistently estimated, in which case the consistent estimators of the remaining components must meet stricter asymptotic linearity conditions.

More generally, many conditions which concern estimation of the mediator density are in fact conditions on estimation of the mediator density ratio (as we see in theorem 1). Therefore, if one decides to make use of case (iv) in lemma 1, the estimation of $p(A|W, Z)$ ought to be such that the corresponding speed conditions and first-order linearity conditions are satisfied for the resulting mediator density ratio estimator.

4 Some existing estimation methodologies

In this section, we describe how the estimating equation approach and the g-computation approach may be applied to the natural direct effect of a binary exposure, and contrast their theoretical properties with those of the proposed targeted estimator.

4.1 Estimating equation approach

Under the estimating equation based approach (Robins (1999), Robins and Rotnitzky (2001), van der Laan and Robins (2003)), one may use the efficient score under a nonparametric model as the estimating function. An estimate of the parameter is given by a root of the efficient score equation. In missing data and causal inference

applications, where the observed data is regarded as a 'censored' version of a full data structure, the efficient score of a parameter of the observed data distribution will involve inverse weighting of the treatment (censoring) mechanism. Moreover, the efficient score is an unbiased estimating function if either the relationship between outcome and covariates under the full data model or the treatment mechanism is correct. Therefore, this approach is also known as doubly robust inverse probability treatment (or censoring) weighting (DR-IPTW).

Under this framework, an estimate for the natural direct effect is given by solving for the root of the equation given by the efficient score in section 2.3. We refer to Tchetgen Tchetgen and Shpitser (2011) for detailed study of this estimator. For given estimators $\hat{Q}_Y$, $\hat{Q}_Z$, $\hat{E}_Z(\hat{\bar{Q}}_Y)$ and $\hat{g}$, the natural direct effect estimate is given by

$$\hat{\psi}_{driptw} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\Bigg\{\frac{I(A_i=1)}{\hat{g}(1|W_i)}\frac{\hat{Q}_Z(Z_i|W_i,0)}{\hat{Q}_Z(Z_i|W_i,1)} - \frac{I(A_i=0)}{\hat{g}(0|W_i)}\Bigg\}\left(Y_i - \hat{\bar{Q}}_Y(W_i,A_i,Z_i)\right)$$
$$+ \frac{I(A_i=0)}{\hat{g}(0|W_i)}\left\{\hat{\bar{Q}}_Y(W_i,1,Z_i) - \hat{\bar{Q}}_Y(W_i,0,Z_i) - \hat{E}_Z\left(\hat{\bar{Q}}_Y(W_i,1,Z_i) - \hat{\bar{Q}}_Y(W_i,0,Z_i)|W,0\right)\right\}$$
$$+ \hat{E}_Z\left(\hat{\bar{Q}}_Y(W_i,1,Z_i) - \hat{\bar{Q}}_Y(W_i,0,Z_i)|W,0\right)\Bigg\}$$

By design, this estimator solves the efficient score equation

$$P_n D^*\left(\hat{Q}_Y, \hat{Q}_Z, \hat{E}_Z(\hat{\bar{Q}}_Y), \hat{g}, \hat{\psi}_{driptw}\right) = 0.$$

Therefore, the DR-IPTW estimator and the proposed targeted MLE estimator share the same asymptotic properties that are inherited from the efficient score. By the same token, they are both sensitive to extreme values of the treatment model, such as in the case of near positivity violations. This was demonstrated in Kang and Schafer (2007). Indeed, in the case of natural direct effect, when $\hat{g}(A_i|W_i)$ is small for some observations, the estimated D_Y^* component of the efficient score will be large; this problem is exacerbated if $A_i = 0$, in which case the estimated D_Z^* is also large.

When near positivity violation is present, the estimating equation estimator may yield estimates that are out of the bounds of the parameter. For instance, in the case of binary outcome Ψ is the mean difference of two probabilities and hence bounded between -1 and 1. But under extreme values of $P_n\hat{D}_Y^*$ and $P_n\hat{D}_Z^*$, the DR-IPTW may yield estimates that are out of these bounds. The proposed targeted estimator using a logistic working submodel (introduced in Gruber and van der Laan (2010)) aims to provide more stable estimates through the combination of a unit linear transformation, which estimates implicitly the boundary of the parameter domain, and the virtue of a substitution estimator, which effectively translates domain boundary into bounds of the parameter range.

4.2 G-computation approach

The sensitivity to near positivity violation of the targeted estimator and the DR-IPTW estimator stem from the use of inverse probability weightings in the efficient score. A g-computation approach based on the identifiability result in (1) avoids this inverse weighting. More specifically, for $\hat{Q}_Y$ and $\hat{Q}_Z$ likelihood based estimators of

the outcome expectation and mediator density, respectively, consider a g-computation estimator given by:

$$\hat{\psi}_{gcomp} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}_Y(W_i,1,Z_i) - \hat{Q}_Y(W_i,0,Z_i)\right)\hat{Q}_Z(Z_i|W_i,0).$$

Unlike the robustness of the targeted estimator and the DR-IPTW estimator, the consistency of the g-computation estimator relies on correct specification of both the outcome expectation and mediator density. In the case of these likelihood-based estimates being correct, the resulting $\hat{\psi}_{gcomp}$ is more efficient than the two robust estimators. However, even though this g-computation estimator does not use inverse probability weighting explicitly, it may still be affected by the data sparsity, since quality of the outcome expectation estimate (even under the correct model) is sensitive to the overlap between the empirical covariate distribution of the treated and the empirical covariate distribution of the control.

5 Simulation Study

In this section we evaluate the performance of the targeted estimator, the DR-IPTW estimator, and the g-computation estimator under model mis-specification and data sparsity. From lemma 1, one expects to see that in the absence of positivity violations, the TMLE and DR-IPTW be robust against model mis-specifications.

5.1 Simulation schemes

The following three data generating schemes are used. The mediator variable Z is discrete with three categories, i.e. $Z \in \{0,1,2\}$. Each scheme has a version with a binary outcome Y and a version with a continuous and bounded outcome Y. Simulations 2 and 3 differ from simulation 1 in their mediator density and treatment mechanism, respectively.

1. **Simulation 1**: no positivity violations.

$$W \sim U(0,2)$$
$$A \sim Bern\left(expit(-1+2W-0.08W^2)\right)$$
$$Z \sim Multinom\Bigg(p(Z=0) = expit(-0.2+0.5A+0.3A\times W+0.7W-1.5W^2),$$
$$p(Z=1|Z\neq 0) = expit(-0.2+0.4A+.8A\times W+0.4W-2.5W^2)\Bigg)$$

version a:

$$Y \sim Bern\Bigg(expit(-2+A-W+W^2+Z+0.8A\times W-A\times W^2$$
$$-0.5A\times Z+0.7A\times Z^2)\Bigg)$$

version b:

$$Y \sim -0.1+0.5A-0.2W+0.1W^2+0.2Z+0.4A\times W-0.5A\times W^2$$
$$-0.3A\times Z+0.5A\times Z^2+N(0,1)$$

Probability of receiving treatment give covariate, $g_A(A=1|w)$, is bounded in $(0.26, 0.94)$. Probability of a particular mediator value z given $A=1$ and $W=w$, $Q_Z(z|A=1,w)$, is bounded between $(0.0005, 0.9753)$, whereas the ratio $Q_Z(z|A=0,w)/Q_Z(z|A=1,w)$ for a particular z and w is bounded in $(0.1376, 2.0103)$. In version b with continuous outcome, the expected value $E(Y|W,A,Z)$ is bounded in $(-0.8, 2.25)$.

The parameters of interest are $\psi_0 = 0.2585079$ for the binary version, and $\psi_0 = 1.158052$ for the continuous version. The semiparametric efficiency bounds are $var(D^*(P_0)) = 1.157$ for the binary version, and $var(D^*(P_0)) = 7.967$ for the continuous version.

2. **Simulation 2**: larger effect of treatment on the distribution of mediator.

$$Z \sim Multinom\Bigg(p(Z=0) = expit(-2-2A-0.5A\times W+3W-W^2),$$
$$p(Z=1|Z\neq 0) = expit(1-4A-A\times W+W+W^2)\Bigg).$$

Conditional distribution for W, A, Y are the same as simulation 1. The conditional probability of a particular mediator value z given $A=1$ and $W=w$, $Q_Z(z|A=1,w)$, range in $(0.017, 0.081)$ for $Z=0$, $(0.046, 0.697)$ for $Z=1$ and $(0.256, 0.936)$ for $Z=2$. The ratio of conditional mediator density $Q_Z(z|A=0,w)/Q_Z(z|A=1,w)$ range in $(6.583, 10.543)$ for $Z=0$, $(0.717, 13.826)$ for $Z=1$ and $(0.0018, 0.253)$ for $Z=2$.

The parameters of interest are $\psi_0 = 0.12556476$ for the binary version, and $\psi_0 = 0.4183004$ for the continuous version. The semiparametric efficiency bounds are $var(D^*(P_0)) = 3.721905$ for the binary version, and $var(D^*(P_0)) = 17.53054$ for the continuous version.

3. **Simulation 3**: near positivity violation the treatment mechanism.

$$A \sim Bern\left(expit(-2-3W+5W^2)\right).$$

Conditional distributions for W, Z, Y are the same as simulation 1, therefore the values of the parameters of interest also remain the same. The treatment mechanism is bounded in $g_A(A=1|W) \in (0.0794, 0.999994)$. Moreover, $g_A(A=1|W) > 0.99$ for $W > 1.5$.

5.2 Estimators

For each data generating distribution, initial maximum likelihood based estimators of the outcome expectation $\bar{Q}_{Y,0}$, treatment mechanism $g_{A,0}$ and mediator density $Q_{Z,0}$ will be obtained according to each of the three cases of model mis-specification in lemma 1, as well as the case where all models are correct. The model mis-specifications considered are as follows:

- Mis-specified outcome model is $Y \sim A + W + Z + A \times Z$, with gaussian family for Continuous outcome, and binomial family (with logit link) for binary Y.
- Mis-specified mediator density is multinomial with $p(Z=0|A,W) \sim A$ and $p(Z=1|A,W,Z\neq 0) \sim A$, both from a binomial family with logit link.
- Mis-specified treatment mechanism is $A \sim W^2$ for simulations 1 and 2, and $A \sim W$ for simulation 3, both from a binomial family with logit link.

The estimators $\hat{\psi}_{gcomp}$ and $\hat{\psi}_{driptw}$ will be implemented using these likelihood-based estimators as described in section 4.

The targeted estimator $\hat{\psi}^*$ will be constructed using these initial estimators under logistic working submodels. Firstly, in the case of continuous outcome, linear transformation T_1 is performed on Y and the initial estimator $\hat{\bar{Q}}_Y$, using bounds given by the range of the observed outcomes and the predicted outcomes under $\hat{\bar{Q}}_Y$. After obtaining the targeted estimator $\hat{\bar{Q}}^*_Y$ on unit scale using logistic working submodel, we perform a second linear transformation T_2 to bound the difference $\hat{\bar{Q}}^*_Y(W,1,Z) - \hat{\bar{Q}}^*_Y(W,0,Z)$ in the unit interval, and obtain the targeted estimator $\hat{E}^*_Z(\hat{\bar{Q}}^*_Y|W,0)$ using logistic working submodel. Finally, we apply the inverse map T_2^{-1} to $\hat{E}^*_Z(\hat{\bar{Q}}^*_Y|W,0)$ and T_1^{-1} to the final effect estimate and estimate of $\bar{Q}_Y$.

We will consider two implementations of TMLE which differ in their initial estimator of the mediated mean outcome difference. In TMLE 1, that initial estimator is given by a plug-in estimator $E_{\hat{Q}_Z}(\hat{\bar{Q}}^*_Y|W,0)$ using $\hat{Q}_Z$ and the updated $\hat{\bar{Q}}^*_Y$. In TMLE 2, that initial estimate is obtained by performing a main term regression $(\hat{\bar{Q}}^*_{Y,n}(W,1,Z) - \hat{\bar{Q}}^*_{Y,n}(W,0,Z)) \sim W$ among the observations with $A=0$. With the data generating distributions under consideration, the initial estimate in TMLE 2 is incorrect regardless of the consistency of $\bar{Q}_Y$ or Q_Z. However, from lemma 1, we expect that TMLE 2 to be consistent in the cases (ii) and (iii), in the absence of positivity violation.

5.3 Results

For each data generating distribution, 1000 samples of each size $n = 500$ and $n = 5000$ are generated. Bias, variance and mse for each sample size are estimated over the 1000 samples.

5.3.1 Simulation 1: No positivity violation

Recall that the parameters of interest are $\psi_0 = 0.2585079$ for the binary version, and $\psi_0 = 1.158052$ for the continuous version, and the semiparametric efficiency bounds are $var(D^*(P_0)) = 1.157$ for the binary version, and $var(D^*(P_0)) = 7.967$ for the continuous version. Therefore, $var(D^*(P_0))/n \approx 2.314e-03$ and $2.314e-04$ for $n = 500$ and 5000, respectively, in the case of the binary outcome, and $var(D^*(P_0))/n \approx 1.593e-02$ and $1.593e-03$ in the case of continuous Y.

The results are detailed in tables 1 and 2. When the outcome expectation and the mediator density are correctly specified, the robust estimators TMLE and DR-IPTW provide little advantage over the g-computation estimator in terms of bias or efficiency. However, when either the outcome expectation or the mediator density are mis-specified, TMLE and DR-IPTW using a correct treatment mechanism provide substantial bias correction so that MSE is reducing at rate $1/n$. The two robust estimators behave similarly. Moreover, as predicted by lemma 1, TMLE 2, which utilizes a mis-specified initial estimator of the mediated mean outcome difference, behaves as well as TMLE 1 when the treatment mechanism is correct.

Table 1: Simulation 1: Binary outcome, no positivity violations

	Bias		Var		MSE	
n	500	5000	500	5000	500	5000
$\bar{Q}_Y$ correct, Q_Z correct						
gcomp: qy.c, qz.c	6.350e-04	5.837e-04	2.452e-03	2.261e-04	2.452e-03	2.264e-04
tmle 1: qy.c, qz.c, ga.c	2.394e-04	5.223e-04	2.499e-03	2.287e-04	2.499e-03	2.290e-04
tmle 2: qy.c, qz.c, ga.c	3.104e-04	5.647e-04	2.525e-03	2.295e-04	2.525e-03	2.298e-04
driptw: qy.c, qz.c, ga.c	2.005e-04	5.227e-04	2.501e-03	2.287e-04	2.501e-03	2.289e-04
tmle: qy.c, qz.c, ga.m	4.453e-04	4.694e-04	2.627e-03	2.373e-04	2.627e-03	2.375e-04
driptw: qy.c, qz.c, ga.m	7.288e-04	4.583e-04	2.754e-03	2.447e-04	2.754e-03	2.449e-04
$\bar{Q}_Y$ correct, g_A correct						
gcomp: qy.c, qz.m	4.260e-02	4.075e-02	3.017e-03	2.771e-04	4.832e-03	1.937e-03
tmle 1: qy.c, qz.m, ga.c	2.221e-04	5.691e-04	2.478e-03	2.279e-04	2.478e-03	2.282e-04
tmle 2: qy.c, qz.m, ga.c	2.004e-04	6.232e-04	2.495e-03	2.286e-04	2.495e-03	2.289e-04
driptw: qy.c, qz.m, ga.c	2.714e-04	5.474e-04	2.494e-03	2.289e-04	2.494e-03	2.292e-04
Q_Z correct, g_A correct						
gcomp: qy.m, qz.c	2.834e-02	2.825e-02	2.434e-03	2.258e-04	3.238e-03	1.024e-03
tmle 1: qy.m, qz.c, ga.c	2.072e-04	5.450e-04	2.530e-03	2.288e-04	2.530e-03	2.291e-04
tmle 2: qy.m, qz.c, ga.c	4.050e-04	5.664e-04	2.543e-03	2.296e-04	2.543e-03	2.299e-04
driptw: qy.m, qz.c, ga.c	3.716e-04	5.493e-04	2.532e-03	2.292e-04	2.532e-03	2.295e-04

Table 2: Simulation 1: Continuous outcome, no positivity violations

	Bias		Var		MSE	
n	500	5000	500	5000	500	5000
Q_Y correct, Q_Z correct						
gcomp: qy.c, qz.c	4.786e-04	5.049e-04	1.597e-02	1.663e-03	1.597e-02	1.663e-03
tmle 1: qy.c, qz.c, ga.c	5.390e-04	4.571e-04	1.654e-02	1.704e-03	1.654e-02	1.704e-03
tmle 2: qy.c, qz.c, ga.c	2.140e-03	4.496e-04	1.686e-02	1.719e-03	1.686e-02	1.720e-03
driptw: qy.c, qz.c, ga.c	4.788e-04	4.569e-04	1.653e-02	1.703e-03	1.653e-02	1.704e-03
tmle: qy.c, qz.c, ga.m	7.706e-04	8.787e-04	1.737e-02	1.797e-03	1.737e-02	1.797e-03
driptw: qy.c, qz.c, ga.m	1.142e-03	9.824e-04	1.844e-02	1.886e-03	1.844e-02	1.887e-03
Q_Y correct, g_A correct						
gcomp: qy.c, qz.m	2.150e-01	2.143e-01	1.778e-02	1.759e-03	6.402e-02	4.767e-02
tmle 1: qy.c, qz.m, ga.c	9.824e-04	5.641e-04	1.666e-02	1.692e-03	1.666e-02	1.692e-03
tmle 2: qy.c, qz.m, ga.c	1.334e-03	5.689e-04	1.679e-02	1.706e-03	1.679e-02	1.706e-03
driptw: qy.c, qz.m, ga.c	6.694e-04	5.908e-04	1.652e-02	1.695e-03	1.652e-02	1.696e-03
Q_Z correct, g_A correct						
gcomp: qy.m, qz.c	7.574e-02	7.435e-02	1.364e-02	1.457e-03	1.938e-02	6.984e-03
tmle 1: qy.m, qz.c, ga.c	7.186e-04	4.839e-04	1.656e-02	1.705e-03	1.656e-02	1.706e-03
tmle 2: qy.m, qz.c, ga.c	1.272e-03	4.591e-04	1.675e-02	1.710e-03	1.675e-02	1.710e-03
driptw: qy.m, qz.c, ga.c	6.413e-04	4.597e-04	1.673e-02	1.707e-03	1.673e-02	1.707e-03

5.3.2 Simulation 2: larger effect of treatment on mediator

Under this simulation scheme, the parameters of interest are $\psi_0 = 0.12556476$ for the binary version, and $\psi_0 = 0.4183004$ for the continuous version. The efficiency bounds are $var(D^*(P_0)) = 3.721905$ for the binary version, and $var(D^*(P_0)) = 17.53054$ for the continuous version. Therefore, $var(D^*(P_0))/n \approx 7.444e-03$ and $7.444e-04$ for $n = 500$ and 5000, respectively, in the case of the binary outcome, and $var(D^*(P_0))/n \approx 3.506e-02$ and $3.506e-03$ in the case of continuous Y.

In this simulation, the treatment has a moderately large effect on the mediator distribution. Compared to simulation 1, this simulation scheme has a larger ratio of $Q_Z(z|0,w)/Q_Z(z|1,w)$ for categories of $Z = 0, 1$ over a region of the sample space of W (details are explained previously). We see that in this case all estimators behave as expected as in the previous simulation. When implemented using the correct treatment mechanism, they provide bias reduction over g-computation estimator in the cases when either the mediator density or the outcome model are mis-specified. When the outcome model and mediator density are both correct, then g-computation is consistent. In this case the TMLE and DR-IPTW are also consistent but less efficient. In all cases, TMLE and DR-IPTW behave similarly. We observe again that when the treatment mechanism is correct, TMLE 2, which utilizes a mis-specified initial estimator of the mediated mean outcome difference, behaves as well as TMLE 1.

Table 3: Simulation 2: Binary outcome, larger effect of treatment on mediator

	Bias		Var		MSE	
n	500	5000	500	5000	500	5000
Q_Y correct, Q_Z correct						
gcomp: qy.c, qz.c	1.993e-03	3.457e-04	6.090e-03	5.743e-04	6.094e-03	5.744e-04
tmle1 : qy.c, qz.c, ga.c	5.457e-03	5.824e-04	8.710e-03	7.873e-04	8.740e-03	7.877e-04
tmle 2: qy.c, qz.c, ga.c	5.226e-03	5.029e-04	8.733e-03	7.889e-04	8.761e-03	7.892e-04
driptw: qy.c, qz.c, ga.c	6.046e-03	5.692e-04	8.973e-03	7.862e-04	9.009e-03	7.865e-04
tmle: qy.c, qz.c, ga.m	5.124e-03	6.550e-04	8.076e-03	7.339e-04	8.102e-03	7.343e-04
driptw: qy.c, qz.c, ga.m	5.140e-03	6.736e-04	8.330e-03	7.693e-04	8.357e-03	7.697e-04
Q_Y correct, g_A correct						
gcomp: qy.c, qz.m	1.200e-02	1.308e-02	5.907e-03	5.674e-04	6.050e-03	7.384e-04
tmle 1: qy.c, qz.m, ga.c	3.042e-03	4.958e-04	6.233e-03	5.812e-04	6.242e-03	5.814e-04
tmle 2: qy.c, qz.m, ga.c	2.854e-03	4.200e-04	6.245e-03	5.833e-04	6.253e-03	5.835e-04
driptw: qy.c, qz.m, ga.c	2.891e-03	4.714e-04	6.194e-03	5.788e-04	6.203e-03	5.791e-04
Q_Z correct, g_A correct						
gcomp: qy.m, qz.c	8.807e-03	1.350e-02	5.736e-03	5.824e-04	5.813e-03	7.648e-04
tmle 1: qy.m, qz.c, ga.c	7.602e-03	5.844e-04	8.903e-03	7.961e-04	8.961e-03	7.964e-04
tmle 2: qy.m, qz.c, ga.c	7.810e-03	6.202e-04	8.902e-03	7.947e-04	8.963e-03	7.951e-04
driptw: qy.m, qz.c, ga.c	6.843e-03	5.093e-04	8.931e-03	7.918e-04	8.978e-03	7.921e-04

Table 4: Simulation 2: Continuous outcome, larger effect of treatment on mediator

	Bias		Var		MSE	
n	500	5000	500	5000	500	5000
Q_Y correct, Q_Z correct						
gcomp: qy.c, qz.c	1.090e-02	4.189e-04	2.494e-02	2.392e-03	2.506e-02	2.392e-03
tmle 1: qy.c, qz.c, ga.c	1.203e-02	2.325e-03	4.245e-02	3.498e-03	4.260e-02	3.504e-03
tmle 2: qy.c, qz.c, ga.c	1.105e-02	2.488e-03	4.236e-02	3.507e-03	4.248e-02	3.513e-03
driptw: qy.c, qz.c, ga.c	1.023e-02	2.373e-03	4.295e-02	3.493e-03	4.305e-02	3.499e-03
tmle: qy.c, qz.c, ga.m	1.244e-02	1.670e-03	3.908e-02	3.094e-03	3.924e-02	3.096e-03
driptw: qy.c, qz.c, ga.m	1.134e-02	1.834e-03	3.991e-02	3.253e-03	4.004e-02	3.257e-03
$\bar{Q}_Y$ correct, g_A correct						
gcomp: qy.c, qz.m	5.763e-02	6.780e-02	2.317e-02	2.244e-03	2.649e-02	6.841e-03
tmle 1: qy.c, qz.m, ga.c	1.276e-02	2.737e-04	2.624e-02	2.418e-03	2.640e-02	2.418e-03
tmle 2: qy.c, qz.m, ga.c	1.149e-02	4.602e-04	2.626e-02	2.426e-03	2.639e-02	2.426e-03
driptw: qy.c, qz.m, ga.c	1.219e-02	3.249e-04	2.598e-02	2.405e-03	2.613e-02	2.405e-03
Q_Z correct, g_A correct						
gcomp: qy.m, qz.c	2.742e-02	4.450e-02	2.947e-02	2.816e-03	3.022e-02	4.796e-03
tmle 1: qy.m, qz.c, ga.c	1.134e-02	2.905e-03	4.632e-02	3.546e-03	4.645e-02	3.555e-03
tmle 2: qy.m, qz.c, ga.c	1.217e-02	2.793e-03	4.613e-02	3.529e-03	4.628e-02	3.537e-03
driptw: qy.m, qz.c, ga.c	5.395e-03	2.925e-03	4.125e-02	3.552e-03	4.128e-02	3.561e-03

5.3.3 Simulation 3: Near positivity violation.

The parameters of interest are the same as in simulation 1: $\psi_0 = 0.2585079$ for the binary version, and $\psi_0 = 1.158052$ for the continuous version. Probability of treatment given covariate W is bounded between $(0.0794, 0.999994)$, with treatment probability > 1.99 for $W > 1.5$. Estimators using a truncated version of the correct treatment mechanism with an a-priori specified bound of (0.025, 0.975) were also considered ('ga.tr').

In the presence of data sparsity, the robustness results of lemma 1 no longer apply when the treatment model values are extreme. We observe here that the MSE of TMLE and DR-IPTW in the case of mis-specification of outcome model or mediator density cease to reduce at a rate proportional to sample size. However, when both of the outcome model and mediator density are correct, TMLE and DR-IPTW with an incorrect treatment mechanism (either through truncation or incorrect modeling) yields MSE that are proportional to sample size. This last result is predicted by the robustness result (i) of lemma 1 and the fact that the mis-specified treatment models is bounded away from 1.

We observe also that in the case of near positivity violation, TMLE 2 is less favorable than TMLE 1 across all cases. This may suggest that under data sparsity, the use of plug-in estimator for the mediated mean outcome difference is more beneficial than considerations such as the rate at which it is estimated. Interestingly, in table 5, which pertains to a binary outcome, we observe an increase in MSE (driven by the increase in variance) as one moves away from the use of substitution principle (with TMLE 1 being the one which uses substitution estimators in all its steps, TMLE 2 which does not use substitution estimator in the initial estimate of the mediated mean outcome difference but uses substitution in the final effect estimate, and DR-IPTW which does not use substitution at all). This may suggest that in the case of positivity violation, when strict bounds exist on the parameter, the degree at which each step of the estimation procedure respects the bounds affects the stability of the resulting estimate. Nonetheless, rigorous analysis is needed to provide more valid insights.

Unlike in previous two cases, we observe that TMLE and DR-IPTW behave differently in some cases. We first consider the version with binary outcome. Since the parameter is an average of probability differences, for a given dataset one would like the effect estimates to be bounded between -1 and 1. However, when using a correctly specified treatment mechanism, the DR-IPTW estimator exhibits estimates that are out of bound (of magnitude larger than 3 in some cases, and of magnitude 11 and 14 in one dataset). The bias, variance and mse of each estimator are detailed in table 5. When outcome model and mediator density are correct, the g-computation is still consistent despite the positivity violation. Nonetheless, the effect of data-sparsity on g-comp is apparent when comparing this g-comp estimator with its counterpart in the case of no positivity violation (table 1, line 1). On the other hand, under correct outcome model and mediator density, TMLE and DR-IPTW have poor variance when implemented with an untruncated correct treatment mechanism ('qy.c, qz.c, ga.c'). However, their performances are improved when implemented with a truncated or mis-specified treatment ('qy.c, qz.c, ga.tr' and 'qy.c, qz.c, ga.m'). We also observe

that in the case of all models correct ('qy.c, qz.c, ga.c'), TMLE and DR-IPTW have a different bias-variance trade-off, with TMLE having smaller variance but larger bias, with respect to DR-IPTW (which has a larger variance but smaller bias). This difference in relative bias and variance is also present in the case of mis-specified mediator density but correct outcome and treatment ('qy.c, qz.m, ga.c'): we observe that using the untruncated correct treatment, TMLE has larger bias and smaller variance than DR-IPTW; but when the truncated treatment mechanism is used, the two robust estimators behave similarly and provide bias reduction over the g-computation estimator. When the outcome model is mis-specified, TMLE and DR-IPTW provide similar bias reduction over g-computation estimator. However, in this case TMLE has a smaller variance (than DR-IPTW) when the untruncated treatment mechanism is used, while the opposite is true when the truncated treatment mechanism is used.

Consider now the case of continuous outcome (table 6). When the outcome model and mediator density are correct, the g-computation is consistent, though converging at a slower rate than its counterpart in the no-sparsity case (table 2, line 1) due to the larger variances. We also observe that when using an untruncated correct treatment mechanism ('qy.c, qz.c, ga.c'), the TMLE 1 has a larger bias but substantially smaller variance than the DR-IPTW in smaller sample size. This is likely due to some large effect estimates in DR-IPTW in the dataset with smaller sample size. The variance of DR-IPTW decreases substantially when sample size increases. On the other hand, when the treatment mechanism is truncated ('qy.c, gz.c, ga.tr'). DR-IPTW has now a smaller variance but larger bias than TMLE 1. When a mis-specified treatment mechanism is used, the two robust estimators behave similarly, but still have larger variance than the g-computation estimator. In the case of incorrect mediator density, when the untruncated treatment mechanism is used, we observe again that DR-IPTW has much smaller bias than TMLE 1, but substantially larger variance in finite sample (for the same reason mentioned above). This difference largely disappears when sample size increases. But when the treatment is truncated, we observe again that TMLE has smaller bias but larger variance than DR-IPTW. In the case when the outcome model is incorrect: when the treatment is not truncated, TMLE 1 has larger bias and smaller variance than DR-IPTW, and that relation is reversed when truncation is applied.

Table 5: Simulation 3: Binary outcome, positivity violations in $p(A|W)$

	Bias		Var		MSE	
n	500	5000	500	5000	500	5000
$\bar{Q}_Y$ correct, Q_Z correct						
gcomp: qy.c, qz.c	2.352e-02	2.019e-03	1.092e-02	1.145e-03	1.147e-02	1.149e-03
tmle 1: qy.c, qz.c, ga.c	5.681e-02	3.592e-02	3.450e-02	1.556e-02	3.773e-02	1.685e-02
tmle 2: qy.c, qz.c, ga.c	4.660e-02	7.505e-02	5.915e-02	2.513e-02	6.132e-02	3.076e-02
driptw: qy.c, qz.c, ga.c	1.846e-02	3.097e-04	4.691e-02	4.824e-02	4.725e-02	4.824e-02
tmle 1: qy.c, gz.c, ga.tr	2.586e-02	2.088e-03	1.555e-02	1.591e-03	1.622e-02	1.596e-03
driptw: qy.c, gz.c, ga.tr	2.393e-02	1.815e-03	1.235e-02	1.248e-03	1.292e-02	1.252e-03
tmle 1: qy.c, qz.c, ga.m	2.324e-02	2.792e-03	1.338e-02	1.381e-03	1.392e-02	1.388e-03
driptw: qy.c, qz.c, ga.m	2.635e-02	2.223e-03	1.837e-02	1.570e-03	1.907e-02	1.575e-03
$\bar{Q}_Y$ correct, g_A correct						
gcomp: qy.c, qz.m	5.017e-02	5.847e-02	1.063e-02	1.355e-03	1.315e-02	4.773e-03
tmle 1: qy.c, qz.m, ga.c	1.434e-01	1.129e-01	1.770e-02	6.660e-03	3.825e-02	1.940e-02
tmle 2: qy.c, qz.m, ga.c	4.655e-02	7.698e-02	5.442e-02	2.105e-02	5.658e-02	2.697e-02
driptw: qy.c, qz.m, ga.c	5.417e-03	7.108e-03	1.768e-01	5.231e-02	1.768e-01	5.236e-02
tmle 1: qy.c, gz.m, ga.tr	3.359e-02	1.655e-02	1.526e-02	1.798e-03	1.638e-02	2.072e-03
driptw: qy.c, gz.m, ga.tr	2.893e-02	3.711e-02	1.391e-02	1.605e-03	1.475e-02	2.982e-03
Q_Z correct, g_A correct						
gcomp: qy.m, qz.c	8.195e-02	8.263e-02	4.271e-03	4.561e-04	1.099e-02	7.284e-03
tmle 1: qy.m, qz.c, ga.c	4.855e-02	9.406e-03	3.555e-02	1.585e-02	3.791e-02	1.594e-02
tmle 2: qy.m, qz.c, ga.c	1.087e-03	6.615e-02	6.191e-02	2.847e-02	6.191e-02	3.285e-02
driptw: qy.m, qz.c, ga.c	3.791e-02	1.157e-02	2.738e-01	1.149e-01	2.753e-01	1.151e-01
tmle 1: qy.m, gz.c, ga.tr	6.252e-02	5.530e-02	1.367e-02	1.342e-03	1.758e-02	4.401e-03
driptw: qy.m, gz.c, ga.tr	7.356e-02	7.080e-02	6.202e-03	6.226e-04	1.161e-02	5.635e-03

Table 6: Simulation 3: Continuous outcome, positivity violations in $p(A|W)$

	Bias		Var		MSE	
n	500	5000	500	5000	500	5000
Q_Y correct, Q_Z correct						
gcomp: qy.c, qz.c	2.390e-03	3.603e-03	7.999e-02	8.030e-03	8.000e-02	8.043e-03
tmle 1: qy.c, qz.c, ga.c	6.235e-02	4.228e-02	7.509e-01	4.091e-01	7.548e-01	4.109e-01
tmle 2: qy.c, qz.c, ga.c	2.556e-01	4.214e-01	1.080e+00	6.355e-01	1.145e+00	8.130e-01
driptw: qy.c, qz.c, ga.c	1.847e-02	2.185e-02	1.836e+00	2.474e-01	1.836e+00	2.479e-01
tmle 1: qy.c, gz.c, ga.tr	2.895e-03	1.652e-03	1.227e-01	1.087e-02	1.227e-01	1.087e-02
driptw: qy.c, gz.c, ga.tr	2.733e-03	2.608e-03	8.762e-02	8.473e-03	8.763e-02	8.479e-03
tmle 1: qy.c, qz.c, ga.m	3.104e-04	4.806e-03	1.231e-01	1.209e-02	1.231e-01	1.212e-02
driptw: qy.c, qz.c, ga.m	6.349e-03	4.447e-03	1.497e-01	1.228e-02	1.497e-01	1.230e-02
Q_Y correct, g_A correct						
gcomp: qy.c, qz.m	2.927e-01	2.996e-01	8.383e-02	8.112e-03	1.695e-01	9.787e-02
tmle 1: qy.c, qz.m, ga.c	5.792e-01	4.894e-01	2.332e-01	1.429e-01	5.687e-01	3.824e-01
tmle 2: qy.c, qz.m, ga.c	2.114e-01	4.413e-01	9.927e-01	5.920e-01	1.037e+00	7.867e-01
driptw: qy.c, qz.m, ga.c	4.033e-02	6.585e-02	8.779e+00	1.899e-01	8.781e+00	1.943e-01
tmle 1: qy.c, gz.m, ga.tr	1.077e-01	8.515e-02	1.030e-01	1.046e-02	1.147e-01	1.771e-02
driptw: qy.c, gz.m, ga.tr	1.795e-01	1.873e-01	9.681e-02	9.235e-03	1.290e-01	4.433e-02
Q_Z correct, g_A correct						
gcomp: qy.m, qz.c	1.553e-01	1.616e-01	2.087e-02	2.142e-03	4.499e-02	2.825e-02
tmle 1: qy.m, qz.c, ga.c	2.451e-02	2.284e-01	7.689e-01	4.513e-01	7.695e-01	5.035e-01
tmle 2: qy.m, qz.c, ga.c	7.633e-02	2.932e-01	1.051e+00	6.325e-01	1.057e+00	7.185e-01
driptw: qy.m, qz.c, ga.c	4.949e-02	9.666e-03	8.180e-01	7.365e-01	8.205e-01	7.366e-01
tmle 1: qy.m, gz.c, ga.tr	1.017e-01	1.108e-01	8.538e-02	6.351e-03	9.573e-02	1.862e-02
driptw: qy.m, gz.c, ga.tr	1.323e-01	1.361e-01	3.437e-02	3.049e-03	5.189e-02	2.157e-02

6 Summary

Using the framework of van der Laan and Rubin (2006), we have proposed a semiparametric efficient, multiply robust substitution estimator for the natural direct effect of a binary exposure in a nonparametric model. The estimation procedure consists of targetedly modifying the conditional outcome expectation and the mediated mean outcome difference, in that order, and then obtaining the effect estimate as the marginal mean of the targeted mediated mean outcome difference. This estimator is asymptotically unbiased if either one of the following holds: i) the conditional outcome expectation given exposure, mediator, and confounders, and the mediated mean outcome difference are consistently estimated; (ii) the exposure mechanism given confounders, and the conditional outcome expectation are consistently estimated; or (iii) the exposure mechanism given confounders, and the conditional mediator density ratio are consistently estimated. If all three conditions hold, then the effect estimate is asymptotically efficient.

In applications, the components that are difficult to estimate are often times the outcome model or the mediator density. Case (iii) implies in particular, that one may still obtain unbiased effect estimates without correct estimation of either of these components. More specifically, if the conditional distribution of treatment given confounders, and the conditional distribution of treatment given confounders and mediator are correct, then the targeted estimator will be asymptotically unbiased. Case (i) implies that if one can only consistently estimate the outcome model, but not the mediator density or treatment mechanism, it is still possible to obtain unbiased estimates if one has available a consistent initial estimator for the mediated mean outcome difference itself (e.g. a data-adaptive estimator which regresses the predicted outcome difference on the confounders, among control observations).

We have also described general conditions for the estimator to be asymptotically linear. More specifically, (a) estimators of each component must converge to their respective limits at a reasonable speed; (b) at most one component may be inconsistently estimated, in which case the consistent estimators of the remaining components must meet stricter asymptotic linearity conditions. These conditions provide a guide for situations where influence curve based variance estimates are realistic.

Estimators which make use of the efficient score are robust, but are generally sensitive to practical positivity violations. We refer to Petersen et al. (2010) for methods of diagnosing and responding to violations of the positivity assumption. The substitution principle and the logistic working submodels in the targeted estimation procedure aims to provide more stable estimates in such situations. However, identification of the parameter depends ultimately on the information available in the given finite sample. A way to improve finite sample robustness is the Collaborative TMLE framework of van der Laan and Gruber (2010), where, instead of estimating the true treatment mechanism, for a given initial estimator of the Q component one estimates a conditional distribution of the treatment, conditioned only on confounders which explain the residual bias of the estimator of Q. We aim to investigate applications of Collaborative TMLE to the effect mediation problem.

References

R.M. Baron and D.A. Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Penalty and Social Psychology*, 51(6):1173–1182, 1986.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6, 2010.

D.M. Hafeman and T.J. VanderWeele. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, 2010.

P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

K. Imai, L. Keel, and T. Yamamoto. Identication, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

B. Jo, E. Stuart, and D. MacKinnon. The use of propensity scores in mediation analysis. *Multivariabe behavior research*, 46:425–452, 2011.

J. Kang and J. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:523–39, 2007.

J.S. Kaufman, R.F. Maclehose, and S. Kaufman. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives & Innovations*, page 1:4, 2004.

M. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

J. Pearl. Direct and indirect effects. In M. Kaufmann, editor, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2000.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

J. Pearl. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, and L. Bernardinelli, editors, *Causality: Statistical Perspectives and Applications*. 2011.

M. Petersen, K. Porter, S.Gruber, Y. Wang, and M.J. van der Laan. Diagnosing and responding to violations in the positivity assumption. Technical report 269, Division of Biostatistics, University of California, Berkeley, 2010. URL `http://www.bepress.com/ucbbiostat/paper269`.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology: the environment and clinical trials*, pages 95–134. Springer-Verlag, 1999.

J.M. Robins. Semantics of causal dag models and the identification of direct and indirect effects. In N. Hjort P. Green and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 70–81. Oxford University Press, Oxford, 2003.

J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(0):143–155, 1992.

J.M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11 (4):920–936, 2001.

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

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

E.J. Tchetgen Tchetgen and I. Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. Technical report 130, Biostatistics, Harvard University, June 2011.

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

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6, 2010.

M.J. van der Laan and M. Petersen. Estimation of direct and indirect causal effects in longitudinal studies. Technical report 155, Division of Biostatistics, University of California, Berkeley, August 2004.

M.J. van der Laan and M.L. Petersen. Direct effect models. *The International Journal of Biostatistics*, 4(1), 2008.

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

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

M.J. van der Laan and D.B. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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

T.J. VanderWeele and S. Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172:1339–1348, 2010.

S. Vansteelandt. Estimating direct effects in cohort and case control studies. *Epidemiology*, 20:851–860, 2009.

W. Zheng and M.J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. Technical report 273, Division of Biostatistics, University of California, Berkeley, November 2010.

W. Zheng and M.J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Appendix A

A1. Proof of lemma 1

$$P_0 D^*(Q,g,\psi_0) = P_{W,0}\left\{\frac{g_0(1|W)}{g(1|W)}\sum_z Q_{Z,0}(z|W,1)\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)}\left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}_Y(W,1,z)\right)\right\} \tag{11}$$

$$- P_{W,0}\left\{\frac{g_0(0|W)}{g(0|W)}\sum_z Q_{Z,0}(z|W,0)\left(\bar{Q}_{Y,0}(W,0,z)-\bar{Q}_Y(W,0,z)\right)\right\} \tag{12}$$

$$+ P_{W,0}\left\{\frac{g_0(0|W)}{g(0|W)}\sum_z Q_{Z,0}(z|W,0)\Big(\bar{Q}_Y(W,1,z)-\bar{Q}_Y(W,0,z)\Big)\right\} \tag{13}$$

$$- P_{W,0}\Big\{\frac{g_0(0|W)}{g(0|W)}E_{Q_Z}(\bar{Q}_Y|W,0)\Big\} \tag{14}$$

$$+ P_{W,0}\Big\{E_{Q_Z}(\bar{Q}_Y|W,0)\Big\} - \psi_0 \tag{15}$$

Suppose (i) holds, i.e. $\bar{Q}_Y = \bar{Q}_{Y,0}$ and $E_{Q_Z}(\bar{Q}_{Y,0}|W,0) = E_{Q_{Z,0}}(\bar{Q}_{Y,0}|W,0)$. Then (11) and (12) are each exactly 0; the expectation in (13) and (14) are the same exactly; and $P_{W,0}\Big\{E_{Q_Z}(\bar{Q}_Y|W,0)\Big\} = P_{W,0}E_{Q_{Z,0}}(\bar{Q}_{Y,0}|W,0) = \psi_0$. Notice that in this case, it was not necessary that $Q_Z = Q_{Z,0}$. But rather, any function $E_Z(\bar{Q}_{Y,0}|W,0)$ which equals the true mediated mean difference $E_{Q_{Z,0}}(\bar{Q}_{Y,0}|W,0)$ will yield the desired result.

Suppose now that (ii) holds. Then (11) and (12) are each exactly 0. The expression in (14) equals $P_{W,0}\Big\{E_{Q_Z}(\bar{Q}_Y|W,0)\Big\}$, and the expression in (13) equals ψ_0. Therefore the mean is zero.

Suppose that (iii) holds. Then, rearranging (11) and (12) we may rewrite the

above expectation as

$$\begin{aligned} P_0 D^*(Q,g,\psi_0) &= P_{W,0}\left\{\sum_z Q_{Z,0}(z|W,0)\left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}_{Y,0}(W,0,z)\right)\right\} \\ &- P_{W,0}\left\{\sum_z Q_{Z,0}(z|W,0)\left(\bar{Q}_{Y}(W,1,z)-\bar{Q}_{Y}(W,0,z)\right)\right\} \\ &+ P_{W,0}\left\{\sum_z Q_{Z,0}(z|W,0)\Big(\bar{Q}_{Y}(W,1,z)-\bar{Q}_{Y}(W,0,z)\Big)\right\} \\ &- P_{W,0}E_{Q_Z}(\bar{Q}_Y|W,0)+P_{W,0}\Big\{E_{Q_Z}(\bar{Q}_Y|W,0)\Big\}-\psi_0 \\ &= 0 \end{aligned}$$

Moreover, contrary to scenario (i), we see that when g is correct, robustness does not impose any requirement on $E_{Q_Z}(\bar{Q}_Y|W,0)$. In fact the cancelation suggests that it may be any function of W.

A2. Asymptotic linearity of the targeted MLE

Theorem 1. *Let $\hat{Q}_{Z,n}$, $\hat{g}_n$ be estimators of $Q_{Z,0}$ and g_0, and $\hat{\bar{Q}}^*_{Y,n}$, $\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})$ be the TMLE estimators, as defined in (8) and (9), of $\bar{Q}_{Y,0}$ and $E_{Q_{Z,0}}(\bar{Q}_{Y,0})$.*

The TMLE estimator ψ^*_n defined in (10) satisfies

$$\begin{aligned} &\psi^*_n-\psi_0=(P_n-P_0)\,D^*\left(\hat{\bar{Q}}^*_{Y,n},\hat{Q}_{Z,n},\hat{g}_n,\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) \\ &+P_{W,0}\sum_z\left(Q_{Y,0}(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right)\left(Q_{Z,0}(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_{Z,0}(z|W,0)\right) \\ &+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(Q_{Y,0}-\hat{\bar{Q}}^*_{Y,n}\right) \\ &+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(E_{Q_{Z,0}}(\hat{\bar{Q}}^*_{Y,n}|W,0)-\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right) \end{aligned} \tag{16}$$

Suppose the estimators $\hat{Q}_{Z,n}$, $\hat{g}_n$, $\hat{\bar{Q}}^*_{Y,n}$ and $\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})$ have limits Q_Z, g, $\bar{Q}^*_Y$, and $E^*_Z(\bar{Q}^*_Y)$ such that

$$(P_n-P_0)\left\{D^*\left(\hat{\bar{Q}}^*_{Y,n},\hat{Q}_{Z,n},\hat{g}_n,\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right)-D^*\left(\bar{Q}^*_Y,Q_Z,g,E^*_Z(\bar{Q}^*_Y)\right)\right\}=o_P(1\sqrt{n}), \tag{17}$$

then

$$\begin{aligned} &\psi^*_n-\psi_0=(P_n-P_0)\,D^*\left(\bar{Q}^*_Y,Q_Z,g,E^*_Z(\bar{Q}^*_Y)\right) \\ &+P_{W,0}\sum_z\left(\bar{Q}_{Y,0}(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right)\left(Q_{Z,0}(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_{Z,0}(z|W,0)\right) \\ &+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\hat{\bar{Q}}^*_{Y,n}\right) \\ &+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(\left\{E_{Q_{Z,0}}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right\}-\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right) \\ &+o_P(1\sqrt{n}) \end{aligned} \tag{18}$$

We proceed now under the assumption of (17) and the assumption that $\hat{\bar{Q}}^*_{Y,n}$ *and* $\hat{Q}_{Z,n}$ *converge to their respective limits at a rate satisfying:*

$$\sqrt{P_{W,0}\left(\bar{Q}^*_Y(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right)^2}\sqrt{P_{W,0}\left(Q_Z(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_Z(z|W,0)\right)^2}$$
$$\leq o_P(1\sqrt{n}) \text{ a.e. over the support of } Z, \tag{19}$$

Consider firstly the case where the true treatment mechanism is given, i.e. $\hat{g}_n = g_0$*. If* $\bar{Q}^*_Y = \bar{Q}_{Y,0}$ *and* $Q_Z = Q_{Z,0}$*, then it follows from (17) and (19) that* ψ^*_n *is asymptotically linear:*

$$\psi^*_n - \psi_0 = (P_n - P_0)\, D^*\left(\bar{Q}_{Y,0}, Q_{Z,0}, g_0, E^*_Z(\bar{Q}_{Y,0})\right) + o_P(1\sqrt{n}). \tag{20}$$

Moreover, if $E^*_Z(\bar{Q}_{Y,0}) = E_{Q_{Z,0}}(\bar{Q}_{Y,0})$*, then* ψ^*_n *is asymptotically efficient. On the other hand, suppose* $\bar{Q}^*_Y = \bar{Q}_{Y,0}$ *but* $Q_Z \neq Q_{Z,0}$*. If there exists a mean zero function* $IC_Z(O)$ *satisfying*

$$\begin{aligned}
&P_{W,0}\sum_z\left(\bar{Q}_{Y,0}(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right)\\
&\times\left\{\sum_{a=0,1}\left(Q_{Z,0}(z|W,a)-Q_Z(z|W,a)\right)\left(a\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-(1-a)\right)\right\}\\
&=(P_n-P_0)IC_Z+o_P(1\sqrt{n}),
\end{aligned} \tag{21}$$

then (17), (19) and (21) imply that ψ^*_n *is asymptotically linear:*

$$\psi^*_n - \psi_0 = (P_n - P_0)\, D^*\left(\bar{Q}_{Y,0}, Q_Z, g_0, E^*_Z(\bar{Q}_{Y,0})\right) + IC_Z + o_P(1\sqrt{n}). \tag{22}$$

Analogously, if $Q_Z = Q_{Z,0}$ *but* $\bar{Q}^*_Y \neq \bar{Q}_{Y,0}$*, and there exists a mean zero function* $IC_Y(O)$ *such that*

$$\begin{aligned}
&P_{W,0}\sum_z\left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}^*_Y(W,1,z)\right)\left(Q_{Z,0}(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_{Z,0}(z|W,0)\right)\\
&=(P_n-P_0)IC_Y+o_P(1\sqrt{n}),
\end{aligned} \tag{23}$$

then (17), (19) and (23) imply that ψ^*_n *is asymptotically linear:*

$$\psi^*_n - \psi_0 = (P_n - P_0)\, D^*\left(\bar{Q}^*_Y, Q_{Z,0}, g_0, E^*_Z(\bar{Q}^*_Y)\right) + IC_Y + o_P(1\sqrt{n}). \tag{24}$$

More generally, consider the case when the treatment mechanism is not given. Assume in addition to the rate condition of (19), the following rate conditions:

$$\sqrt{P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g,\hat{Q}_{Z,n})\right)^2}\sqrt{P_0\left(\bar{Q}^*_Y-\hat{\bar{Q}}^*_{Y,n}\right)^2}\leq o_P(1/\sqrt{n}), \tag{25}$$

and

$$\sqrt{P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g(0|W)}\right)^2}\sqrt{P_0\left(E^*_Z(\hat{\bar{Q}}^*_{Y,n}|W,0)-\hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right)^2}\leq o_P(1/\sqrt{n}). \tag{26}$$

If $g = g_0$, $\bar{Q}_Y^* = \bar{Q}_{Y,0}$, $Q_Z = Q_{Z,0}$ and $E_Z^*(\hat{\bar{Q}}_{Y,n}^*) = E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*)$, then (17), (19), (25) and (26) imply ψ_n^* is asymptotically linear, as in (20). Moreover, it follows from these conditions that $E_Z^*(\bar{Q}_Y^*) = E_{Q_{Z,0}}(\bar{Q}_{Y,0})$, therefore ψ_n^* is in fact asymptotically efficient. Suppose $g = g_0$, $\bar{Q}_Y^* = \bar{Q}_{Y,0}$, $Q_Z \neq Q_{Z,0}$ but $E_Z^*(\hat{\bar{Q}}_{Y,n}^*) = E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*)$, then (17), (19), (25), (26), (21) imply the asymptotic linearity of ψ_n^* as in (22). However, if $Q_Z \neq Q_{Z,0}$ and $E_Z^*(\hat{\bar{Q}}_{Y,n}^*) \neq E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*)$, in addition to the asymptotically linear condition of (21), assume there exists another mean zero function $IC_Z'(O)$ such that

$$P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)} - \frac{I(A=0)}{g_0(0|W)}\right)\left(\left\{E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*|W,0)\right\} - E_Z^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right) = (P_n - P_0)IC_Z' + o_P(1\sqrt{n}), \tag{27}$$

then (17), (19), (25), (26), (21) and (27) imply that ψ_n^* is asymptotically linear:

$$\psi_n^* - \psi_0 = (P_n - P_0)\, D^*\left(\bar{Q}_{Y,0}, Q_Z, g_0, E_Z^*(\bar{Q}_{Y,0})\right) + (IC_Z + IC_Z') + o_P(1\sqrt{n}). \tag{28}$$

Analogously, suppose only $g = g_0$, $Q_Z = Q_{Z,0}$, and $E_Z^*(\bar{Q}_Y^*) = E_{Q_{Z,0}}(\bar{Q}_Y^*)$, but $\bar{Q}_Y^* \neq \bar{Q}_{Y,0}$. If the condition (23) holds and there exists a mean zero function $IC_Y'(O)$ such that

$$P_0\left(\bar{Q}_{Y,0} - \bar{Q}_Y^*\right)\left(C_Y(\hat{g}_n, \hat{Q}_{Z,n}) - C_Y(g_0, \hat{Q}_{Z,n})\right) = (P_n - P_0)IC_Y' + o_P(1/\sqrt{n}), \tag{29}$$

then (17), (19), (25), (26), (23) and (29) imply that ψ_n^* is asymptotically linear:

$$\psi_n^* - \psi_0 = (P_n - P_0)\, D^*\left(\bar{Q}_Y^*, Q_{Z,0}, g_0, E_{Q_{Z,0}}^*(\bar{Q}_Y^*)\right) + (IC_Y + IC_Y') + o_P(1\sqrt{n}). \tag{30}$$

Lastly, consider the case where $\bar{Q}_Y^* = \bar{Q}_{Y,0}$, $Q_Z = Q_{Z,0}$ and $E_Z^*(\hat{\bar{Q}}_{Y,n}^*) = E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*)$, but $g \neq g_0$. If there exist mean zero function $IC_g(O)$ and $IC_g'(O)$ such that

$$P_0\left(\bar{Q}_{Y,0} - \hat{\bar{Q}}_{Y,n}^*\right)\left(C_Y(g, \hat{Q}_{Z,n}) - C_Y(g_0, \hat{Q}_{Z,n})\right) = (P_n - P_0)IC_g + o_P(1\sqrt{n}) \tag{31}$$

and

$$P_0\left(\frac{I(A=0)}{g(0|W)} - \frac{I(A=0)}{g_0(0|W)}\right)\left(\left\{E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*|W,0)\right\} - \hat{E}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right) = (P_n - P_0)IC_g' + o_P(1\sqrt{n}), \tag{32}$$

then (17), (19), (25), (26), (31) and (32) imply that ψ_n^* is asymptotically linear:

$$\psi_n^* - \psi_0 = (P_n - P_0)\, D^*\left(\bar{Q}_{Y,0}, Q_{Z,0}, g, E_{Q_{Z,0}}(\bar{Q}_{Y,0})\right) + (IC_g + IC_g') + o_P(1\sqrt{n}). \tag{33}$$

Proof of theorem 1

To see (16) we note firstly that for any Q and ψ

$$\begin{aligned}
&P_0 D^*\left(\bar{Q}_Y, Q_Z, g_0, \psi\right) = E_{W,0}E_{Q_{Z,0}}\left(\bar{Q}_{Y,0}|W,0\right) - \psi \\
&+ P_{W,0}\sum_z \left(\bar{Q}_{Y,0}(W,1,z) - \bar{Q}_Y(W,1,z)\right)\left(Q_{Z,0}(z|W,1)\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)} - Q_{Z,0}(z|W,0)\right) \\
&= \psi_0 - \psi \\
&+ P_{W,0}\sum_z \left(\bar{Q}_{Y,0}(W,1,z) - \bar{Q}_Y(W,1,z)\right)\left(Q_{Z,0}(z|W,1)\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)} - Q_{Z,0}(z|W,0)\right)
\end{aligned}$$

On the other hand, $P_n D^* \left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, \hat{g}_n, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) = 0$ by design of the TMLE estimator. Combining these two results, we may express

$$\begin{aligned}
\hat{\psi}^*_n - \psi_0 &= (P_n - P_0)\, D^* \left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, \hat{g}_n, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) \\
&+ P_{W,0} \sum_z \left(\bar{Q}_{Y,0}(W,1,z) - \hat{\bar{Q}}^*_{Y,n}(W,1,z)\right) \left(Q_{Z,0}(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)} - Q_{Z,0}(z|W,0)\right) \\
&+ P_0 \left\{D^* \left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, \hat{g}_n, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) - D^* \left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, g_0, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right)\right\},
\end{aligned}$$

where the last summand may be rewritten as

$$\begin{aligned}
&P_0 \left\{D^* \left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, \hat{g}_n, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) - D^* \left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, g_0, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right)\right\} = \\
&+P_0 \left(C_Y(\hat{g}_n, \hat{Q}_{Z,n}) - C_Y(g_0, \hat{Q}_{Z,n})\right) \left(\bar{Q}_{Y,0} - \hat{\bar{Q}}^*_{Y,n}\right) \\
&+P_0 \left(\frac{I(A=0)}{\hat{g}_n(0|W)} - \frac{I(A=0)}{g_0(0|W)}\right) \left(E_{Q_{Z,0}}(\hat{\bar{Q}}^*_{Y,n}|W,0) - \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right).
\end{aligned} \tag{34}$$

Result (16) thus follows. Moreover, the Donsker class condition in (17) yields (18).

The conditions for asymptotic linearity can now be ascertained from the second order terms of (18), namely,

$$\begin{aligned}
&P_{W,0} \sum_z \left(\bar{Q}_{Y,0}(W,1,z) - \hat{\bar{Q}}^*_{Y,n}(W,1,z)\right) \left(Q_{Z,0}(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)} - Q_{Z,0}(z|W,0)\right) \\
&+ P_0 \left(C_Y(\hat{g}_n, \hat{Q}_{Z,n}) - C_Y(g_0, \hat{Q}_{Z,n})\right) \left(\bar{Q}_{Y,0} - \hat{\bar{Q}}^*_{Y,n}\right) \\
&+ P_0 \left(\frac{I(A=0)}{\hat{g}_n(0|W)} - \frac{I(A=0)}{g_0(0|W)}\right) \left(\left\{E_{Q_{Z,0}}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right\} - \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}|W,0)\right),
\end{aligned}$$

by a straightforward expansion:

$$
\begin{aligned}
&P_{W,0}\sum_z\left(\bar{Q}_Y^*(W,1,z)-\hat{\bar{Q}}_{Y,n}^*(W,1,z)\right)\left(Q_Z(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_Z(z|W,0)\right) && (35)\\
&+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g,\hat{Q}_{Z,n})\right)\left(\bar{Q}_Y^*-\hat{\bar{Q}}_{Y,n}^*\right) && (36)\\
&+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g(0|W)}\right)\left(\left\{E_Z^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right\}-\hat{E}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right) && (37)\\
&+P_{W,0}\sum_z\left(\bar{Q}_Y^*(W,1,z)-\hat{\bar{Q}}_{Y,n}^*(W,1,z)\right)\\
&\times\left\{\sum_{a=0,1}\left(Q_{Z,0}(z|W,a)-Q_Z(z|W,a)\right)\left(a\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-(1-a)\right)\right\} && (38)\\
&+P_{W,0}\sum_z\left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}_Y^*(W,1,z)\right)\left(Q_Z(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_Z(z|W,0)\right) && (39)\\
&+P_0\left(C_Y(g,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_Y^*-\hat{\bar{Q}}_{Y,n}^*\right) && (40)\\
&+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\bar{Q}_Y^*\right) && (41)\\
&+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g(0|W)}\right)\left(\left\{E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*|W,0)-E_Z^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right\}\right) && (42)\\
&+P_0\left(\frac{I(A=0)}{g(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(\left\{E_Z^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right\}-\hat{E}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right) && (43)\\
&+P_{W,0}\sum_z\left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}_Y^*(W,1,z)\right)\\
&\times\left\{\sum_{a=0,1}\left(Q_{Z,0}(z|W,a)-Q_Z(z|W,a)\right)\left(a\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-(1-a)\right)\right\}\\
&+P_0\left(C_Y(g,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\bar{Q}_Y^*\right)\\
&+P_0\left(\frac{I(A=0)}{g(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*|W,0)-E_Z^*(\hat{\bar{Q}}_{Y,n}^*|W,0)\right).
\end{aligned}
$$

In this theorem we study situations pertaining to (i) $g=g_0$ and $\bar{Q}_Y^*=\bar{Q}_{Y,0}$; (ii) $g=g_0$, $Q_Z=Q_{Z,0}$; or (iii) $\bar{Q}_Y^*=\bar{Q}_{Y,0}$, $Q_Z=Q_{Z,0}$ and $E_Z^*(\hat{\bar{Q}}_{Y,n}^*|W,0)=E_{Q_{Z,0}}(\hat{\bar{Q}}_{Y,n}^*|W,0)$. Under either case, the last three unlabeled summands above are exactly zero. Therefore, we only need to focus on the first order ((38), (39),(40), (41), (42), (43)) and second order ((35), (36), (37)) remainders.

Under condition (19), the second order term in (35) is bounded by

$$
\begin{aligned}
&\left|\sum_z P_{W,0}\left(\bar{Q}_Y^*(W,1,z)-\hat{\bar{Q}}_{Y,n}^*(W,1,z)\right)\left(Q_Z(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_Z(z|W,0)\right)\right|\\
&\le\sum_z\left|P_{W,0}\left(\bar{Q}_Y^*(W,1,z)-\hat{\bar{Q}}_{Y,n}^*(W,1,z)\right)\left(Q_Z(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_Z(z|W,0)\right)\right|\\
&\le\sum_z\sqrt{P_{W,0}\left(\bar{Q}_Y^*(W,1,z)-\hat{\bar{Q}}_{Y,n}^*(W,1,z)\right)^2P_{W,0}\left(Q_Z(z|W,1)\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-Q_Z(z|W,0)\right)^2}\\
&\le o_P(1/\sqrt{n}).
\end{aligned}
$$

If g_0 is given (e.g. in a randomized controlled trial), i.e. $\hat{g}_n = g_0$, then the term

$$P_0\left\{D^*\left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, \hat{g}_n, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) - D^*\left(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, g_0, \hat{E}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right)\right\} = 0,$$

which implies that (36), (37), (40)- (43) are all exactly 0. When $\bar{Q}^*_Y = \bar{Q}_{Y,0}$ and $Q_Z = Q_{Z,0}$, the remainders in (38) and (39) vanish. Therefore (20) holds. If $\bar{Q}^*_Y = \bar{Q}_{Y,0}$ and $Q_Z \neq Q_{Z,0}$, the condition (21) ensures the asymptotic linearity of the first order remainder (38). Therefore, (22) holds. When $\bar{Q}^*_Y \neq \bar{Q}_{Y,0}$ and $Q_Z = Q_{Z,0}$, a similar argument applying condition (23) on (39)implies (24).

Now, consider the case where $\hat{g}_n$ is an estimator of g_0 which converges to some g. The rate condition (19) bounds the second order term in (35), as mentioned before. The rate conditions (25) and (26) ensure that the second order terms (36) and (37) are also $o_P(1/\sqrt{n})$. If $g = g_0$, $\bar{Q}^*_Y = \bar{Q}_{Y,0}$ and $Q_Z = Q_{Z,0}$ and $E^*_Z(\hat{\bar{Q}}^*_{Y,n}) = E_{Q_{Z,0}}\hat{\bar{Q}}^*_{Y,n})$, then the first order remainders (38), (39), (40), (41), (42) and (43) all vanish. Moreover, the linear term in (18) is given by $D^*(\bar{Q}_{Y,0}, Q_{Z,0}, g_0)$, which implies asymptotic efficiency. Suppose on the other hand that $Q_Z \neq Q_{Z,0}$ but $E^*_Z(\hat{\bar{Q}}^*_{Y,n}) = E_{Q_{Z,0}}(\hat{\bar{Q}}^*_{Y,n})$. Then the term (42) vanishes and condition (21) ensures asymptotic linearity of the first order remainder (38), which implies (22). If $Q_Z \neq Q_{Z,0}$ and $E^*_Z(\hat{\bar{Q}}^*_{Y,n}) \neq E_{Q_{Z,0}}(\hat{\bar{Q}}^*_{Y,n})$, then (21) and (27) ensure that the first order remainders (38) and (42) are both asymptotically linear, which implies (28). An analogous argument shows that conditions (23) and (29) on remainders (39) and (41) imply (30) in the case $\bar{Q}^*_Y \neq \bar{Q}_{Y,0}$. Lastly, if $g \neq g_0$, then conditions (31) and (32) ensure asymptotic linearity of the first order remainders (40) and (43). This implies (33).

Appendix B

In this section, we describe an alternative targeted estimator for the natural direct effect by targeting on the conditional outcome expectation and the mediator density. The key difference between the estimator proposed in the main section and the estimator in this appendix lies in that the former defines a loss function and parametric working submodel for the mediated mean outcome difference $E_{Q_Z}(\bar{Q}_Y|W, 0)$, whereas the latter defines a loss function and parametric working submodel for the conditional mediator density Q_Z and then estimates the mediated mean outcome difference plugging in the targeted mediator density and the targeted $\bar{Q}_Y$.

The loss function L_Y for Q_Y remains the same as in the main section. That is, we consider the loglikelihood loss when Y is binary or bounded in the unit interval, or the squared error loss otherwise. Consequently, the parametric submodels for Q_Y remain the same as in the main section.

We make the assumption that the mediator Z is discrete with $K + 1$ levels, i.e. $Z \in \{0, 1, \ldots, K\}$. Let the variable Z_k denote the indicator $I(Z = k)$, and $Q_{Z_k} \equiv P(Z_k|Z_0, \ldots, Z_{k-1}, W, A)$, for $k = 0, \ldots, K - 1$. Then, Z has a binary representation $Z = (Z_k : k = 0, \ldots, K - 1)$, and $Q_Z = \prod_{k=0}^{K-1} Q_{Z_k}$. For notational

convenience, we will sometimes write $Q_{Z_k}(1|W,A)$ for the conditional probability $P(Z_k = 1|Z_0,\ldots,Z_{k-1},W,A)$, and $\mathbf{Z_{k-1}}$ for the vector $(Z_0,\ldots,Z_{k-1})$.

Define for Q_Z the loglikelihood loss function

$$L_Z(Q_Z) = \sum_{k=0}^{K-1} Z_k \log Q_{Z_k}(1|W,A) + (1-Z_k)\log Q_{Z_k}(0|W,A).$$

We wish to find a logistic parameter working submodel $Q_Z(\epsilon)$ satisfying

$$\frac{d}{d\epsilon} L_Z\left(Q_Z(\epsilon)\mid_{\epsilon=0}\right) = D_Z(Q_Z,g,\bar{Q}_Y). \tag{44}$$

For that purpose, we first decompose D_Z orthogonally as $D_Z = \sum_{k=0}^{K-1} D_{Z_k}$, where

$$D_{Z_k} = \frac{I(A=0)}{g(0|W)}\left\{E\left(D_Z|Z_k=1,\mathbf{Z_{k-1}},W,A\right) - E\left(D_Z|Z_k=0,\mathbf{Z_{k-1}},W,A\right)\right\}\left(Z_k - Q_{Z_k}(1|W,A)\right).$$

A parametric submodel for $Q_Z = \prod_{k=0}^{K-1} Q_{Z_k}$ may defined in terms of each component:

$$logitQ_{Z_k}(g,\bar{Q}_Y)(\epsilon)(1|W,A) = logitQ_{Z_k}(1|W,A) + \epsilon C_{Z_k}(g,\bar{Q}_Y)(W,A),$$

where

$$\begin{aligned}
&C_{Z_k}(g,\bar{Q}_Y)(W,A) = \\
&\frac{I(\mathbf{Z_{k-1}}=0,A=0)}{g(0|W)}\left\{E\left(\bar{Q}_Y(W,Z)|Z_k=1,\mathbf{Z_{k-1}},W,A\right) - E\left(\bar{Q}_Y(W,Z)|Z_k=0,\mathbf{Z_{k-1}},W,A\right)\right\} \\
&= I(\mathbf{Z_{k-1}}=0)\frac{I(A=0)}{g(0|W)}\left\{\bar{Q}_Y(W,k) - \sum_{l>k}\bar{Q}_Y(W,l)\left\{\prod_{m=k+1}^{l-1} Q_{Z_m}(0|W,A)\right\}Q_{Z_l}(1|W,A)\right\}.
\end{aligned}$$

This way, the working submodel $Q_Z(g,\bar{Q}_Y)(\epsilon) = \prod_{k=0}^{K-1} Q_{Z_k}(g,\bar{Q}_Y)(\epsilon)$ satisfies (44).

Given initial estimators of $\bar{Q}_{Y,0}$, $Q_{Z,0}$, and g_0, a targeted MLE estimator for $\hat{\bar{Q}}^*_Y$ for $Q_{Y,0}$ is constructed as in (8). Using this updated $\hat{\bar{Q}}^*_Y$, the optimal ϵ for the submodel of Q_Z is given by

$$\hat{\epsilon}^* = \arg\min_{\epsilon} P_n L_Z\left(\hat{Q}_Z(\hat{g},\hat{\bar{Q}}^*_Y)(\epsilon)\right),$$

and the targeted estimator of the mediator density is given by $\hat{Q}_Z(\hat{g},\hat{\bar{Q}}^*_Y)(\hat{\epsilon}^*)$, we denote this by $\hat{Q}^*_Z$ for convenience. Finally, the targeted MLE estimator of ψ_0 is the substitution estimator plugging in these two updated components:

$$\hat{\psi}^* = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\bar{Q}}^*_Y(W_i,1,Z_i) - \hat{\bar{Q}}^*_Y(W_i,0,Z_i)\right\}\hat{Q}^*_Z(Z=Z_i|W_i,A=0).$$

It follows from (6) that $P_n D^*_Y(\hat{\bar{Q}}^*_Y,\hat{Q}_Z,\hat{g}) = 0$, and it follows from (44) that $P_n D^*_Z\left(\hat{\bar{Q}}^*_Y,\hat{Q}^*_Z,\hat{g}\right) = 0$. Moreover, the empirical distribution $\hat{Q}_W$ of W solves the score equation $P_n D^*_W(\hat{\bar{Q}}^*_Y,\hat{Q}^*_Z,\hat{Q}_W) = 0$. Therefore the resulting targeted estimator also solves the efficient score equation.