The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 3

Targeted Maximum Likelihood Estimation of Natural Direct Effects

Wenjing Zheng, *University of California, Berkeley*
Mark J. van der Laan, *University of California, Berkeley*

Recommended Citation:
Zheng, Wenjing and van der Laan, Mark J. (2012) "Targeted Maximum Likelihood Estimation of Natural Direct Effects," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 3.
DOI: 10.2202/1557-4679.1361

Targeted Maximum Likelihood Estimation of Natural Direct Effects

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 (2001) formalized the definition of two types of direct effects (natural and controlled) under the counterfactual framework. The efficient scores (under a nonparametric model) for the various natural effect parameters and their general robustness conditions, as well as an estimating equation based estimator using the efficient score, are provided in Tchetgen Tchetgen and Shpitser (2011b). In this article, we apply the targeted maximum likelihood framework of van der Laan and Rubin (2006) and van der Laan and Rose (2011) to construct a semiparametric efficient, multiply robust, substitution estimator for the natural direct effect which satisfies the efficient score equation derived in Tchetgen Tchetgen and Shpitser (2011b). We note that the robustness conditions in Tchetgen Tchetgen and Shpitser (2011b) may be weakened, thereby placing less reliance on the estimation of the mediator density. More precisely, the proposed estimator is asymptotically unbiased if either one of the following holds: i) the conditional mean outcome given exposure, mediator, and confounders, and the mediated mean outcome difference are consistently estimated; (ii) the exposure mechanism given confounders, and the conditional mean outcome are consistently estimated; or (iii) the exposure mechanism and the mediator density, or the exposure mechanism and the conditional distribution of the exposure given confounders and mediator, are consistently estimated. If all three conditions hold, then the effect estimate is asymptotically efficient. Extensions to the natural indirect effect are also discussed.

KEYWORDS: natural direct effects, natural indirect effects, mediation analysis, mediation formula, mediator, direct effects, asymptotic efficiency, robust, double robust, asymptotic linearity, canonical gradient, efficient influence curve, efficient score, loss-based learning, targeted maximum likelihood estimator, targeted learning, parametric working submodels

Author Notes: We thank the anonymous reviewers for the very helpful comments and suggestions.

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 (2001) 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 its would-be value under a reference (null) exposure level. 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 (2001), Robins (2003), van der Laan and Petersen (2004), Petersen, Sinisi, and van der Laan (2006), Hafeman and VanderWeele (2010), Imai, Keele, and Yamamoto (2010), Robins and Richardson (2010), and Pearl (2011).

Prior to the formal frameworks developed by Robins and Greenland (1992) and Pearl (2001), 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 the 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, Maclehose, and Kaufman (2004). The developments of Robins and Greenland (1992) and Pearl (2001), and the identifiability studies that followed suit, address definition and identification of direct and indirect effects in causal models that do not put restrictions on the distribution of the observed data, 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 can be identified as a function of the marginal covariate distribution, the conditional mediator density, and the conditional mean outcome (e.g. Robins and Greenland (1992), Pearl (2001), Robins (2003) and van der Laan and Petersen (2004), Petersen et al. (2006)). 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 compo-

nents is inconsistent, the effect estimate will be biased. VanderWeele and Vansteelandt (2010) illustrated how this approach can 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 effect estimation is based on the estimating equation methodology developed by Robins (1999), Robins and Rotnitzky (2001) and van der Laan and Robins (2003). Under this approach, a score is expressed as a function of the parameter of interest ψ and a nuisance parameter η (whenever such representation is possible); if the resulting estimating equation, as an equation in the variable ψ, has a unique solution, the parameter estimate is given as the root to this equation. For most parameters arising from causal inference, the efficient score under a nonparametric model is a robust estimating function (i.e. unbiased against mis-specification of specific components of the likelihood), 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-supplied 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 therefore 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. VanderWeele (2009) discussed more fully the use of marginal structural models with inverse probability weighting for estimation of the natural direct effect parameter. 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 given estimators of relevant components of the likelihood P, one iteratively maximizes the likelihood (or minimize a loss) along a least favorable submodel through the initial estimators. The parameter estimate is given by evaluating the parameter map at the final estimator of the likelihood, thus providing a substitution estimator of the parameter of interest. By construction, the final estimate of the likelihood satisfies the efficient score equation in the variable P. Therefore, the effect estimate also shares the robustness properties of the efficient score. In addition, the substitution principle incorporates global constraints of the statistical model that do not affect the form of the efficient score; this allows for potential improvement in finite sample performance. van der Laan and Petersen

(2008) also applied the targeted MLE procedure to their generalized class of direct effect parameters. Both the estimating equation 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 mean 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-supplied conditional mediator density in the method of van der Laan and Petersen (2008) must be correct. The use of propensity score matching in causal effect estimation was introduced in Rosenbaum and Rubin (1983). Application of propensity score in mediation analysis has also been proposed (e.g. Jo, Stuart, MacKinnon, and Vinokur (2011)).

Most recently, Tchetgen Tchetgen and Shpitser (2011b) derived the efficient scores (under a nonparametric model) for the various natural effect parameters, and established their general robustness properties and their implications on efficiency bounds. They also proposed semiparametric efficient, multiply robust estimators based on the estimating equation methodology using the efficient score equation. We also refer the reader to that work for presentation of a sensitivity analysis framework to assess the impact of the ignorability assumption of the mediator variable on inference. In Tchetgen Tchetgen and Shpitser (2011a), the authors extended the theory to the case where one specifies a parametric model for the natural direct (indirect) effect conditional on a subset of baseline covariates.

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 proposed estimator satisfies the efficient score equation derived in Tchetgen Tchetgen and Shpitser (2011b). However, we note that the robustness conditions in Tchetgen Tchetgen and Shpitser (2011b) may be weakened (lemma 1), thereby placing less reliance on the estimation of the mediator density. This weaker version of robustness conditions is of particular interest when the mediator is high-dimensional, since it allows one to replace estimation of the conditional mediator density with objects that are easier (or at least with more available tools) to estimate. More precisely, the proposed estimator is asymptotically unbiased if either one of the following holds: i) the conditional mean outcome given exposure, mediator, and confounders, and the mediated mean outcome difference are consistently estimated; (ii) the exposure mechanism given confounders, and the conditional mean outcome are consistently estimated; or (iii) the exposure mechanism and the mediator density, or the exposure mechanism and the conditional distribution of the exposure given confounders and mediator, are consistently estimated. If all three conditions hold, then the effect estimate is asymptotically efficient. We also extend the results to the estimation of natural indirect effects. In addition, we discuss in detail conditions needed to ensure asymptotic linearity of

the resulting estimator. These conditions should provide a guideline for situations where an influence curve based variance estimate is realistic.

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 and its efficient score in greater detail in section 2.3. Section 3 describes how to construct a 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. 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 estimation methodologies: 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. Section 6 extends analogously the discussions on identifiability, robustness, and estimation to the case of natural indirect effect. This article concludes with a summary and a few remarks.

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 a mediator of interest 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 (NPSEM) of Pearl (2009) to encode the causal relations under consideration. The NPSEM on a 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 encode for each $x \in X$ the variables that have direct influence on x. More specifi-

DOI: 10.2202/1557-4679.1361

cally, in the present situation the causal relations are described 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 $P_{U,X}$.

The counterfactual variables or potential outcomes in the Rubin Causal Model (Rubin (1978), Rosenbaum and Rubin (1983) and Holland (1986)) can be represented as restrictions on the input of the functions f_x. For instance, the counterfactual $Z(a)$ is defined as the random variable $Z(a) \equiv f_Z(W, A = a, U_Z)$, and can be interpreted as the mediator variable that the unit would have had if the exposure had been a. In particular, $Z(a)$ is a random variable through U_W and U_Z. Similarly, $Y(a', Z(a))$ is the counterfactual outcome that results from setting $Y(a', Z(a)) \equiv f_Y(W, A = a', Z(a), U_Y)$, and can be interpreted as the individual's response if the exposure had been a' while the mediator variable had been identical to the one under exposure a. $Y(a', Z(a))$ is a random variable through U_W, U_Z and U_Y.

Under the NPSEM, a causal parameter of interest is 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 can be interpreted from the following hypothetical experiment: one randomly assigns each subject to treatment or control, while always setting the subject's mediator variable to its value under no treatment, and then takes the difference in mean outcome between the treated and control cohort.

2.2 Identifiability

We will also use the notation $Z(A)$ to denote the unintervened $Z = f_Z(W, A, U_Z)$, which is random through U_W, U_A, U_Z. Similarly, the unintervened $Y(A, Z(A)) \equiv f_Y(W, A, Z(A), U_Y)$ is random through U_W, U_A, U_Z, U_Y. Under experimental or observational studies, for each unit, the investigator only observes the outcome and mediator response under the unit's actual exposure. In other words, the observation is in fact $O = (W, A, Z(A), Y(A, Z(A)))$. Hence, the causal parameter $\Psi(P_{U,X})$ is not always identifiable from the observed data.

Conditions under which the natural direct effect (or natural effects in general) will be identifiable were addressed extensively in Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen et al. (2006), Hafeman and VanderWeele (2010), Imai et al. (2010), Robins and Richardson (2010) and Pearl (2011). In particular, Pearl (2001) gave the following identifiability conditions: If randomization assumptions

A1. For all values (a,z), $Y(a,z)$ given W is identifiable,
A2. For all values of a, $Z(a)$ given W is identifiable,

and the conditional independence assumption

A3. For all $a \neq a', z$, $Y(a',z)$ is independent of $Z(a)$ given W

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

$$\begin{aligned}\Psi(P_{U,X}) &\overset{A1,A2,A3}{=} \Psi(P_0)\\ &\equiv E_W\left\{\sum_z \left[E(Y|W,A=1,Z=z)-E(Y|W,A=0,Z=z)\right] p(z|W,A=0)\right\}. \qquad (1)\end{aligned}$$

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

Many of these previous authors have established that the randomization assumptions A1 and A2 can be satisfied by requiring that (A,Z) is independent of $Y(a,z)$, given W, and A is independent of $Z(a)$, given W. These can be ensured by measuring sufficient covariates to control for confounding of the effects of treatment on outcome, treatment on mediator, and mediator on outcome. As a result, the distributions of $Y(a,z)$ and $Z(a)$ will be identifiable within covariate stratum.

Petersen et al. (2006) showed that A3 can be weakened to a conditional mean independence $E(Y(1,z)-Y(0,z)|W) = E(Y(1,z)-Y(0,z)|W,Z(0)=z)$. Still, it was recognized in Pearl (2001) that the conditional counterfactual independence is in general difficult to interpret. Imai et al. (2010) offered a stronger version of assumption A3 which is more interpretable: $Y(a',z)$ is independent of Z given W and $A=a$. This new version implies assumption A3, but the converse is not necessarily true. Robins and Richardson (2010) established that in general condition A3 cannot be enforced by randomized experiments, which implies that the natural effects are in general not identifiable by randomized experiments. In such cases, what kind of causal interpretations can the statistical parameter in (1) still offer? Note that

under the randomization assumptions A1 and A2 alone, the statistical parameter (1) equals (e.g. Pearl (2001), van der Laan and Petersen (2008)):

$$\Psi(P_0) \overset{A1,A2}{=} E_W\left(\sum_z E(Y(1,z)-Y(0,z)|W)P(Z(0)=z|W)\right).$$

The quantity in the right hand side is the population mean of an average of subject-specific controlled direct effect $E(Y(1,z)-Y(0,z)|W)$, weighted by $P(Z(0)=z|W)$. However, while this quantity serves to provide a causal interpretation for the statistical parameter (1) in the absence of condition A3, it is certainly not the natural direct causal effect; therefore one should be cautious about putting it into the context of the traditional total effect decomposition.

2.3 The Natural Direct Effect Parameter

Let $\mathscr{M}$ denote a model containing the true data generating distribution P_0. For any $P \in \mathscr{M}$, 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 $g(A|W) \equiv P_A(A|W)$, $Q_W(W) \equiv P_W(W)$, $Q_Z(Z|W,A) \equiv P_Z(Z|W,A)$, and $\bar{Q}_Y(W,A,Z) \equiv E(Y|W,A,Z)$. Moreover, let $Q \equiv (Q_W, Q_Z, \bar{Q}_Y)$. 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 $P \in \mathscr{M}$. For instance, $P_0 f \equiv \sum_{o\in\mathscr{O}} f(o)dP_0(o)$ denotes the expectation of f under the true data generating distribution, while $P_n f \equiv \frac{1}{n}\sum_{i=1}^n f(o_i)$ denotes the empirical mean of f.

One may consider the natural direct effect parameter Ψ in (1) as a map

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

We refer to the inner expectation above as the *(null level) mediated mean outcome difference*, and denote it by the map $Q \mapsto \psi_Z(Q)$, where

$$\psi_Z(Q)(W) \equiv \psi_Z(Q_Z,\bar{Q}_Y)(W) \equiv E_{Q_Z}\left(\bar{Q}_Y(W,1,Z)-\bar{Q}_Y(W,0,Z)|W,A=0\right). \quad (2)$$

This way, $\Psi(Q) = \Psi(Q_W, \psi_Z(Q)) = E_{Q_W}(\psi_Z(Q)(W))$. The parameter of interest (1) is this map evaluated at the true data generating distribution:

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

2.3.1 Effcient score

Under a nonparametric model $\mathscr{M}$, for any $P \in \mathscr{M}$, the *efficient score* (*efficient influence curve*, or *canonical gradient*) of Ψ at P, as derived in Tchetgen Tchetgen and Shpitser (2011b), 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 can also be derived by first considering $\Psi(P)$ as a function of $P = (Pf : f \in \mathscr{F})$, where $\mathscr{F}$ is a class of indicator functions $\mathscr{F} = \{I(w,a,z,y), I(w,a,z), I(w,a), I(w) : w \in \mathscr{W}, a \in \mathscr{A}, z \in \mathscr{Z}, y \in \mathscr{Y}\}$. For any given "vector" $h = (h(f) : f \in \mathscr{F})$, one can consider a directional derivative $\frac{d}{d\varepsilon}\Psi(P + \varepsilon h)|_{\varepsilon=0}$. The efficient score is given by the directional derivative applied to the direction of $h = (f(O) - Pf : f \in \mathscr{F})$. In other words, it is given by $\sum_{f \in \mathscr{F}} \frac{\partial \Psi(P)}{\partial Pf}(f(O) - Pf)$. A more detailed exposition can be found in van der Laan and Rose (2011).

2.3.2 Robustness of the efficient score

The general robustness conditions of the efficient score were given in Tchetgen Tchetgen and Shpitser (2011b): (i) the mediator density $Q_Z(Z|W,A)$ and the conditional mean outcome $\bar{Q}_Y(W,A,Z)$ are both correct; (ii) the conditional mean outcome and the exposure mechanism $g(A|W)$ are both correct; or (iii) the exposure mechanism and the mediator density are both correct. We note below that conditions (i) and (iii) may be weakened to accommodate difficulties in estimation of the mediator density. In fact, the estimation of Q_Z may be avoided with the use of data-adaptive estimators. This is particularly appealing when Z is high dimensional. We summarize these in the following lemma and its subsequent remarks. The proof of this lemma is straightforward from the form of the efficient score, and we refer the interested reader to appendix App1.

Lemma 1. ***Robustness of the efficient score***
Suppose there exists constants $1 > \delta, \delta' > 0$ *such that* $g(A = 1|W) < 1 - \delta$ *and* $Q_Z(Z|W,1) < 1 - \delta'$ *a.e. over the support of* W *and* Z*. 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 mean outcome $\bar{Q}_Y = E(Y|W,A,Z)$*, and the mediated mean outcome difference* $\psi_Z(Q) = E_{Q_Z}(\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)|W,0)$ *are correct.*
(ii) The exposure mechanism $g(A|W)$*, and the conditional mean outcome are correct.*
(iii) The exposure mechanism and conditional mediator density $Q_Z(Z|W,A)$*, or the exposure mechanism and the conditional distribution of treatment given mediator and covariates* $p(A|W,Z)$*, are correct.*

Condition (i) follows from the fact that, given $\bar{Q}_Y$, we only need a conditional expectation of $\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)$ under $Q_Z(Z|W,0)$. Therefore, consistent estimation of $Q_{Z,0}$ per se is not necessary to obtain consistent estimator of $\psi_Z(Q_0)$, as long as one has a consistent estimator $\hat{\bar{Q}}_{Y,n}$ of $\bar{Q}_{Y,0}$ and an optimal procedure to regress the difference $\hat{\bar{Q}}_{Y,n}(W,1,Z) - \hat{\bar{Q}}_{Y,n}(W,0,Z)$ on W among the control observations. Condition (iii) is a consequence of the fact when g is correct, dependence on consistent estimation of Q_Z is only through $\frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)}$, which can be consistently estimated using either Q_Z or combining ratios of $g(A|W)$ and $p(A|W,Z)$.

When Z is high-dimensional, few tools are available to estimate the conditional mediator density $Q_Z(Z|W,A)$. On the other hand, there is abundant literature addressing estimation of conditional means. This can be used to estimate $\psi_Z(Q)$, and conditional probabilities of a categorical A. Lemma 1 implies in particular that estimation of $Q_{Z,0}$ may be replaced by estimations of $g_0(A|W)$, $p_0(A|W,Z)$, and the conditional expectation $\psi_Z(Q_0)$,

3 Targeted Maximum Likelihood Estimation 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 MLE (TMLE) estimator of a parameter of interest $\Psi(P_0) = \Psi(Q_0)$

calls for two sets of ingredients. For each component $Q_j(P)$ of $Q(P)$, one defines a uniformly bounded (w.r.t. the supremum norm) *loss function* $L_j : \mathscr{Q}_j \to \mathscr{L}^\infty(K)$ satisfying

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

where $\mathscr{L}^\infty(K)$ 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* $\{Q_j(P)(\varepsilon_j) : \varepsilon_j\} \subset \mathscr{M}$ passing through $Q_j(P)$ at $\varepsilon_j = 0$ with score $D_j^*(P)$ at $\varepsilon_j = 0$ that satisfies

$$\langle \frac{d}{d\varepsilon_j} L_j(Q_j(P)(\varepsilon_j)) \mid_{\varepsilon_j=0} \rangle \supset \langle D_j^*(P) \rangle,$$

where $\langle h \rangle$ denotes the linear span of a vector h. These result in a least favorable parametric submodel $Q(\varepsilon)$ through Q. For given initial estimator $(\hat{Q}, \hat{g})$ of (Q_0, g_0), the fluctuation parameter ε is fitted to minimize the empirical risk of $\hat{Q}(\varepsilon)$, providing an updated estimator $\hat{Q}(\hat{\varepsilon})$. This updating process is repeated until $\hat{\varepsilon} \approx 0$. The final estimator $\hat{Q}^*$ of Q_0 is then used to obtain a substitution estimator $\Psi(\hat{Q}^*)$ of $\Psi(Q_0)$. By its construction, the estimator $\hat{Q}^*$ satisfies the efficient score equation $P_n D^*(\hat{Q}^*, \hat{g}, \Psi(\hat{Q}^*)) = 0$.

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 $\psi_Z(Q)$ as defined in (2). Secondly, the empirical marginal distribution $\hat{Q}_{W,n}$ of W is a consistent estimator of $Q_{W,0}$ that readily solves the equation $P_n D_W^*(\psi_Z(Q), \hat{Q}_{W,n}) = 0$ for any $\psi_Z(Q)$. Hence, the proposed estimator will focus on targeted estimation of $\bar{Q}_{Y,0}(W,A,Z)$, and $\psi_Z(Q_0)(W)$.

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 $\psi_Z(Q_0)$. We refer the interested reader to Appendix B for this alternative approach. The key difference between the proposed and the alternative targeting procedures lies in that the former defines a loss function and parametric working submodel for the mediated mean outcome difference $\psi_Z(Q)$, 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 $\psi_Z(Q_0)$ by plugging in the targeted mediator density and the targeted $\bar{Q}_Y$. We note that the bias variance trade-off in the proposed targeting procedure is more optimal over the alternative procedure for estimating the ultimate component of interest, which is the mediated mean outcome difference.

DOI: 10.2202/1557-4679.1361

3.1 Construction of the Targeted MLE

3.1.1 Loss functions and parametric working submodels

Suppose for now that Y is binary or continuous and bounded. In the latter case, without loss of generality we may assume that Y is bounded in $(0,1)$. We consider the minus-loglikelihood loss function for $\bar{Q}_Y$:

$$L_Y(\bar{Q}_Y)(O) = -\log\left(\bar{Q}_Y(W,A,Z)^Y(1-\bar{Q}_Y(W,A,Z))^{(1-Y)}\right). \tag{3}$$

Under this loss function, consider the logistic working submodel

$$\bar{Q}_Y(\varepsilon_1) \equiv expit\left(logit(\bar{Q}_Y) + \varepsilon_1 C_Y(Q_Z,g)\right),$$

where $C_Y(Q_Z,g)(O) = \left\{\frac{I(A=1)}{g(1|W)}\frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)} - \frac{I(A=0)}{g(0|W)}\right\}$. Note that this submodel $\bar{Q}_Y(\varepsilon_1)$ depends on the components Q_Z and g, but we suppress that in the notation. This submodel satisfies

$$\frac{d}{d\varepsilon_1} L_Y\left(\bar{Q}_Y(\varepsilon_1)\right)|_{\varepsilon_1=0} = D_Y^*(\bar{Q}_Y, Q_Z, g). \tag{4}$$

For a given $\bar{Q}_Y$, the difference $\bar{Q}_Y(W,Z) \equiv \bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)$ is also bounded. Without loss of generality, we may also assume it is bounded between $(0,1)$. Let the loss function for $\psi_Z(Q)$ be

$$L_Z(\psi_Z(Q))(O) = \\ -I(A=0)\log\Big((\psi_Z(Q)(W))^{\bar{Q}_Y(W,Z)}(1-\psi_Z(Q)(W))^{1-\bar{Q}_Y(W,Z)}\Big).$$

Under this loss function, the logistic working submodel

$$\psi_Z(Q)(\varepsilon_2) \equiv expit\left(logit\left(\psi_Z(Q)\right) + \varepsilon_2 C_Z(g)\right),$$

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

$$\frac{d}{d\varepsilon_2} L_Z\left(\psi_Z(Q)(\varepsilon_2)\right)|_{\varepsilon_2=0} = D_Z^*(\psi_Z(Q), \bar{Q}_Y, g). \tag{5}$$

The dependence of $\psi_Z(Q)(\varepsilon_2)$ on g is again suppressed in our notation.

Note that linear transformations onto the unit interval may be needed in order to use the 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 $\psi_Z(Q)$, the necessary linear transformations and their inverse maps do not affect the properties of the estimators.

In settings where Y is not bounded, 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

$$L_Z(\psi_Z(Q))(O) = I(A=0)\left(\bar{Q}_Y(W,Z) - \psi_Z(Q)(W)\right)^2;$$

and corresponding parametric working submodels

$$\bar{Q}_Y(\varepsilon_1) = \bar{Q}_Y + \varepsilon_1 C_Y(Q_Z, g)$$

and

$$\psi_Z(Q)(\varepsilon_2) = \psi_Z(Q) + \varepsilon_2 C_Z(g).$$

However, compared to the minus loglikelihood losses, this choice of loss functions and the corresponding parametric working submodels may result in estimators that are more sensitive to near positivity violations (Gruber and van der Laan (2010), Gruber and van der Laan (2011)). Therefore, in such situations it would be more sensible to bound Y by the range of the observed data, and apply the minus loglikelihood losses above.

3.1.2 Implementation

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

$$\hat{\varepsilon}_1^* = \arg\min_{\varepsilon} P_n L_Y\left(\hat{\bar{Q}}_{Y,n}(\varepsilon_1)\right)$$

be the optimal ε_1 which minimizes the empirical risk. We are reminded that, though not shown in the notation, the estimators $(\hat{Q}_{Z,n}, \hat{g}_n)$ are used in constructing $\hat{\bar{Q}}_{Y,n}(\varepsilon_1)$. The update

$$\hat{\bar{Q}}_{Y,n}^* \equiv \hat{\bar{Q}}_{Y,n}(\hat{\varepsilon}_1^*) \tag{6}$$

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

Next, let $\hat{\psi}_Z(P_n)(\cdot)$ be an estimating procedure for $\psi_Z(Q_0)$. That is, for given observations P_n, $\hat{\psi}_{Z,n} \equiv \hat{\psi}_Z(P_n)$ is a function which maps an estimator $\hat{\bar{Q}}_{Y,n}$ of $\bar{Q}_{Y,0}$ to an estimator $\hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n})$ of $\psi_Z(Q_{Z,0}, \bar{Q}_{Y,0})$. This function $\hat{\psi}_{Z,n}$ depends on the estimation procedure $\hat{\psi}_Z$, and the observed data P_n. This estimating procedure can be plug-in or regression-based. For a plug-in estimator, $\hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n}) \equiv \psi_Z(\hat{Q}_{Z,n}, \hat{\bar{Q}}_{Y,n})$. For a regression-based estimator, $\hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n})$ regresses the difference $\hat{\bar{Q}}_{Y,n}(W,1,Z) - \hat{\bar{Q}}_{Y,n}(W,0,Z)$ on W among control observations. In this latter

case, $\hat{\psi}_{Z,n}$ encodes what this regression procedure consists of, and the observed data on which it is carried out.

Given the targeted MLE $\hat{\bar{Q}}^*_{Y,n}$ of the mean outcome, $\hat{\psi}_{Z,n}(\hat{\bar{Q}}^*_{Y,n})$ is an initial estimator of the mediated mean outcome difference $\psi_Z(Q_{Z,0}, Q_0)$. The optimal ε_2 is given by

$$\hat{\varepsilon}^*_2 = \arg\min_{\varepsilon} P_n L_Z\left(\hat{\psi}_{Z,n}(\hat{\bar{Q}}^*_{Y,n})(\varepsilon_2)\right).$$

We are reminded that, though not shown in the notation, the estimator $\hat{g}_n$ is used in constructing $\hat{\psi}_{Z,n}(\hat{\bar{Q}}^*_{Y,n})(\varepsilon_2)$. The update

$$\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}) \equiv \hat{\psi}_{Z,n}(\hat{\bar{Q}}^*_{Y,n})(\hat{\varepsilon}^*_2) \tag{7}$$

is the *targeted MLE estimator of* $\psi_Z(Q_{Z,0}, Q_0)$. The *targeted MLE estimator of* $\psi_0 = E_{W,0}(\psi_Z(Q_{Z,0}, Q_0)(W))$ is thus given by

$$\hat{\psi}^*_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})(W_i). \tag{8}$$

It follows from (4) that $P_n D^*_Y(\hat{\bar{Q}}^*_{Y,n}, \hat{Q}_{Z,n}, \hat{g}_n) = 0$ and it follows from (5) that $P_n D^*_Z\left(\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}), \hat{\bar{Q}}^*_{Y,n}, \hat{g}_n\right) = 0$. Moreover, the empirical distribution $\hat{Q}_{W,n}$ of W solves $P_n D^*_W(\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n}), \hat{Q}_{W,n}) = 0$. Therefore the resulting targeted estimator $\hat{\psi}^*_n$ solves the efficient score equation.

Remarks on implementation: *When Z is high-dimensional, and A is categorical, 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 can use an estimator $\frac{\hat{p}_n(A=0|W,Z)}{\hat{g}_n(A=0|W)}\frac{\hat{g}_n(A=1|W)}{\hat{p}_n(A=1|W,Z)}$. Similarly, the estimating procedure $\hat{\psi}_{Z,n}(\cdot)$ does not need to use $\hat{Q}_{Z,n}$ and can be any procedure which regresses $\hat{\bar{Q}}^*_{Y,n}(W,1,Z) - \hat{\bar{Q}}^*_{Y,n}(W,0,Z)$ on W among control observations. Therefore, when Z is high dimensional, estimation of Q_Z may be avoided if one has available optimal estimators $\hat{g}_n$ and $\hat{p}_n(A|W,Z)$, and a regression-based estimator $\hat{\psi}_{Z,n}(\cdot)$. From lemma 1, we see that this still allows for robust estimation.*

3.2 Asymptotic Properties of the Targeted MLE

Since the proposed targeted MLE estimator satisfies 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 $\hat{\bar{Q}}^*_{Y,n}$ and

the mediated mean outcome difference $\hat{\psi}_Z^*(\hat{\bar{Q}}_{Y,n}^*)$ are consistent; (ii) the treatment mechanism $\hat{g}_n$ and the conditional outcome expectation $\hat{\bar{Q}}_{Y,n}^*$ are consistent; (iii) the treatment mechanism $\hat{g}_n$ and the conditional mediator density $\hat{Q}_{Z,n}(Z|W,A)$, or the treatment mechanism and $\hat{p}_n(A|W,Z)$, are consistent. These properties are illustrated in the simulations section below.

Under certain empirical conditions, an estimator that satisfies a given estimating equation will be asymptotically linear with influence curve given by the estimating function (e.g. Bickel, Klaassen, Ritov, and Wellner (1997), van der Vaart (1998), van der Laan and Robins (2003), Tsiatis (2006), Kosorok (2008)). In this case, the central limit theorem implies that one can obtain an asymptotic variance estimate of the said estimator using the variance estimate of its influence curve. Otherwise, bootstrap procedures can be used to obtain variance estimates for the estimator. We detail conditions for asymptotic linearity of the targeted MLE estimator in theorem 1 below. These conditions state that in general, asymptotic linearity requires that: 1) estimators of the likelihood converge to their respective limits at a reasonable speed (second-order conditions), and 2) if there is a component that is not consistently estimated, the remaining consistent components must be estimated in a specific asymptotically linear fashion (first-order conditions). These conditions provide a guideline for situations where influence curve based variance estimates are realistic. Note that these conditions stem from the properties of the efficient score, and therefore can be easily modified to apply to any estimator which satisfy the efficient score equation (e.g. Tchetgen Tchetgen and Shpitser (2011b)). We also refer the readers to Zheng and van der Laan (2010) and Zheng and van der Laan (2011) for an alternative targeted estimation procedure which weaken the empirical process conditions through the use of cross-validation.

We use the following notations in the theorem: Let $\hat{Q}_{Z,n}$, $\hat{g}_n$ be estimators of $Q_{Z,0}$ and g_0; and let $\hat{\bar{Q}}_{Y,n}^*$, $\hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*)$ be the TMLE estimators of $\bar{Q}_{Y,0}$ and $\psi_Z(Q_0)$, as defined in (6) and (7). The TMLE estimator $\hat{\psi}_n^*$ of ψ_0 is defined in (8). Let Q_Z, g, $\bar{Q}_Y^*$ be limits of $\hat{Q}_{Z,n}$, $\hat{g}_n$, $\hat{\bar{Q}}_{Y,n}^*$. Note that these limits are not necessarily the true data generating components. Similarly, for the procedure $\hat{\psi}_{Z,n}^*(\cdot)$ which, for a given $\hat{\bar{Q}}_{Y,n}^*$, provides a targeted estimator $\hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*)$ of the conditional mean $\psi_Z(Q_{Z,0}, \hat{\bar{Q}}_{Y,n}^*)$, let $\psi_Z^*(\cdot)$ denote its limit. In other words, $\psi_Z^*(\hat{\bar{Q}}_{Y,n}^*)$ estimates $\psi_Z(Q_{Z,0}, \hat{\bar{Q}}_{Y,n}^*)$ using an infinite population. The limit of $\hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*)$ is given by $\psi_Z^*(\bar{Q}_Y^*)$.

DOI: 10.2202/1557-4679.1361

Theorem 1. *Firstly, the TMLE estimator $\hat{\psi}_n^*$ defined in (8) satisfies*

$$\begin{aligned}
&\hat{\psi}_n^* - \psi_0 = (P_n - P_0) D^* \left(\hat{\bar{Q}}_{Y,n}^*, \hat{Q}_{Z,n}, \hat{g}_n, \hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*) \right) \\
&+ P_{W,0} \sum_z Q_{Z,0}(z|W,1) \left(\bar{Q}_{Y,0}(W,1,z) - \hat{\bar{Q}}_{Y,n}^*(W,1,z) \right) \left(\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)} - \frac{Q_{Z,0}(z|W,0)}{Q_{Z,0}(z|W,1)} \right) \\
&+ P_0 \left(C_Y(\hat{g}_n, \hat{Q}_{Z,n}) - C_Y(g_0, \hat{Q}_{Z,n}) \right) \left(\bar{Q}_{Y,0} - \hat{\bar{Q}}_{Y,n}^* \right) \\
&+ P_0 \left(\frac{I(A=0)}{\hat{g}_n(0|W)} - \frac{I(A=0)}{g_0(0|W)} \right) \left(\psi_Z(Q_{Z,0}, \hat{\bar{Q}}_{Y,n}^*) - \hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*) \right). \qquad (9)
\end{aligned}$$

Suppose the following assumption holds:

$$(P_n - P_0) \left\{ D^* \left(\hat{\bar{Q}}_{Y,n}^*, \hat{Q}_{Z,n}, \hat{g}_n, \hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*) \right) - D^* \left(\bar{Q}_Y^*, Q_Z, g, \psi_Z^*(\bar{Q}_Y^*) \right) \right\} = o_P(1\sqrt{n}). \quad (10)$$

We proceed now under the assumption (10) and the following assumptions regarding speed of convergence:

$$\begin{aligned}
&\sqrt{P_{W,0} E_{Q_{Z,0}} \left(\left(\bar{Q}_Y^*(W,1,Z) - \hat{\bar{Q}}_{Y,n}^*(W,1,Z) \right)^2 | W, A=1 \right)} \times \\
&\sqrt{P_{W,0} E_{Q_{Z,0}} \left(\left(\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)} - \frac{Q_Z(z|W,0)}{Q_Z(z|W,1)} \right)^2 | W, A=1 \right)} \\
&= o_P(1\sqrt{n}), \qquad (11)
\end{aligned}$$

$$\sqrt{P_0 \left(C_Y(\hat{g}_n, \hat{Q}_{Z,n}) - C_Y(g, \hat{Q}_{Z,n}) \right)^2} \sqrt{P_0 \left(\bar{Q}_Y^* - \hat{\bar{Q}}_{Y,n}^* \right)^2} = o_P(1/\sqrt{n}), \qquad (12)$$

and

$$\sqrt{P_0 \left(\frac{I(A=0)}{\hat{g}_n(0|W)} - \frac{I(A=0)}{g(0|W)} \right)^2} \sqrt{P_0 \left(\psi_Z^*(\hat{\bar{Q}}_{Y,n}^*) - \hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*) \right)^2} = o_P(1/\sqrt{n}). \qquad (13)$$

If $g = g_0$, $\bar{Q}_Y^* = \bar{Q}_{Y,0}$, $Q_Z = Q_{Z,0}$ and $\psi_Z^*(\cdot) = \psi_Z(Q_{Z,0}, \cdot)$, then (10), (11), (12) and (13) imply that $\hat{\psi}_n^*$ is asymptotically linear. Moreover, it also follows from these conditions that $\psi_Z^*(\bar{Q}_Y^*) = \psi_Z(Q_{Z,0}, \bar{Q}_{Y,0})$, therefore $\hat{\psi}_n^*$ is in fact asymptotically efficient.

Suppose $\bar{Q}_Y^* = \bar{Q}_{Y,0}$, $\psi_Z^*(\cdot) = \psi_Z(Q_{Z,0},\cdot)$, *but* $g \neq g_0$, *and* $Q_Z \neq Q_{Z,0}$. *If there exist mean zero functions* $IC_g(O)$ *and* $IC_g'(O)$ *such that*

$$P_0\left(C_Y(g,\hat{Q}_{Z,n}) - C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0} - \hat{\bar{Q}}_{Y,n}^*\right) = (P_n - P_0)IC_g + o_P(1\sqrt{n}) \tag{14}$$

and

$$\begin{aligned} &P_0\left(\frac{I(A=0)}{g(0|W)} - \frac{I(A=0)}{g_0(0|W)}\right)\left(\psi_Z(Q_{Z,0},\hat{\bar{Q}}_{Y,n}^*) - \hat{\psi}_{Z,n}^*(\hat{\bar{Q}}_{Y,n}^*)\right) \\ &= (P_n - P_0)IC_g' + o_P(1\sqrt{n}), \end{aligned} \tag{15}$$

and there exists a mean zero function $IC_{Q_Z}(O)$ *satisfying*

$$\begin{aligned} &P_{W,0}\sum_z Q_{Z,0}(z|W,1)\left(\bar{Q}_{Y,0}(W,1,z) - \hat{\bar{Q}}_{Y,n}^*(W,1,z)\right)\left(\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)} - \frac{Q_{Z,0}(z|W,0)}{Q_{Z,0}(z|W,1)}\right) \\ &= (P_n - P_0)IC_{Q_Z} + o_P(1\sqrt{n}), \end{aligned} \tag{16}$$

then (10), (11), (12), (13), (14), (15) and (16) imply that $\hat{\psi}_n^*$ *is asymptotically linear:*

$$\begin{aligned} \hat{\psi}_n^* - \psi_0 = (P_n - P_0)\left\{D^*\left(\bar{Q}_{Y,0}, Q_Z, g, \psi_Z(Q_{Z,0},\bar{Q}_{Y,0})\right) + IC_g + IC_g' + IC_{Q_Z}\right\} \\ + o_P(1\sqrt{n}). \end{aligned}$$

If $Q_Z = Q_{Z,0}$, *then the condition (16) is trivially true with* $IC_{Q_Z} \equiv 0$.

On the other hand, consider the case of $g = g_0$ *and* $\bar{Q}_Y^* = \bar{Q}_{Y,0}$, *but* $\psi_Z^*(\cdot) \neq \psi_Z(Q_{Z,0},\cdot)$ *and* $Q_Z \neq Q_{Z,0}$. *Suppose that there exists a mean zero function* $IC_{\psi_Z}(O)$ *such that*

$$\begin{aligned} &P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)} - \frac{I(A=0)}{g_0(0|W)}\right)\left(\psi_Z(Q_{Z,0},\hat{\bar{Q}}_{Y,n}^*) - \psi_Z^*(\hat{\bar{Q}}_{Y,n}^*)\right) \\ &= (P_n - P_0)IC_{\psi_Z} + o_P(1\sqrt{n}). \end{aligned} \tag{17}$$

Then (10), (11), (12), (13), (16), and (17) imply that $\hat{\psi}_n^*$ *is asymptotically linear:*

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

If $Q_Z = Q_{Z,0}$, *then the condition (16) is trivially true with* $IC_{Q_Z} \equiv 0$. *Similarly, if* $\psi_Z^*(\hat{\bar{Q}}_{Y,n}^*) = \psi_Z(Q_{Z,0},\hat{\bar{Q}}_{Y,n}^*)$, *then (17) is vacuously true with* $IC_{\psi_Z} \equiv 0$.

Lastly, suppose $g = g_0$, $Q_Z = Q_{Z,0}$, *but* $\bar{Q}_Y^* \neq \bar{Q}_{Y,0}$ *and* $\psi_Z^*(\cdot) \neq \psi_Z(Q_{Z,0},\cdot)$. *Suppose there exists mean zero functions* $IC_Y(O)$ *and* $IC_Y'(O)$ *such that*

$$P_{W,0}\sum_z Q_{Z,0}(z|W,1)\left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}_Y^*(W,1,z)\right)\left(\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-\frac{Q_{Z,0}(z|W,0)}{Q_{Z,0}(z|W,1)}\right)$$
$$=(P_n-P_0)IC_Y+o_P(1\sqrt{n}), \tag{18}$$

and

$$P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\bar{Q}_Y^*\right)=(P_n-P_0)IC_Y'+o_P(1/\sqrt{n}). \tag{19}$$

Then (10), (11), (12), (13), (17), (18) and (19) imply that $\hat{\psi}_n^*$ *is asymptotically linear:*

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

If $\psi_Z^*(\hat{\bar{Q}}_{Y,n}^*)=\psi_Z(Q_{Z,0},\hat{\bar{Q}}_{Y,n}^*)$, *then (17) is vacuously true with* $IC_{\psi_Z}\equiv 0$.

We refer the reader to appendix App2 for the proof. We also note that conditions regarding convergence of Q_Z in fact only involve the ratio $\frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)}$, therefore can be expressed in terms of $g(A|W)$ and $p(A|W,Z)$.

4 Some Existing Estimation Methodologies

In this section, we describe how the estimating equation and the g-computation approaches can 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 (EE) based approach (Robins (1999), Robins and Rotnitzky (2001), van der Laan and Robins (2003)), one may use the efficient score $D^*(P)$ under a nonparametric model as an estimating function of ψ, if i) $D^*(P)$ can be expressed as a function of ψ and some nuisance parameter η, i.e. $D^*(P)=D(\psi(P),\eta(P))$, for some function D, and ii) the solution to the resulting equation in the variable ψ is unique. When these requirements hold, an estimate of the parameter is given by the root of the resulting estimating equation, i.e. $\hat{\psi}$ is defined as the solution to the equation $P_nD^*(\hat{\eta}(P_n),\hat{\psi})=0$.

An estimator of the natural direct effect under this framework is provided in Tchetgen Tchetgen and Shpitser (2011b). For given estimators $\hat{Q}_{Y,n}$, $\hat{Q}_{Z,n}$, $\hat{g}_n$, and

an estimating procedure $\hat{\psi}_{Z,n}(\cdot)$ for $\psi_Z(Q_0)$, the EE estimator for the natural direct effect is given by

$$\hat{\psi}_{ee} \equiv \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\left(\frac{I(A_i=1)}{\hat{g}_n(1|W_i)}\frac{\hat{Q}_{Z,n}(Z_i|W_i,0)}{\hat{Q}_{Z,n}(Z_i|W_i,1)} - \frac{I(A_i=0)}{\hat{g}_n(0|W_i)}\right)\left(Y_i - \hat{\bar{Q}}_{Y,n}(W_i,A_i,Z_i)\right)$$
$$+\frac{I(A_i=0)}{\hat{g}_n(0|W_i)}\left(\hat{\bar{Q}}_{Y,n}(W_i,1,Z_i) - \hat{\bar{Q}}_{Y,n}(W_i,0,Z_i) - \hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n})\right)$$
$$+\hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n})\Bigg\}$$

We remind the reader again that in the present paper, $\hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n})$ may not need to use $\hat{Q}_{Z,n}$, but will surely make use of $\hat{\bar{Q}}_{Y,n}$.

By definition, this EE estimator solves the efficient score equation

$$P_n D^*\left(\hat{\bar{Q}}_{Y,n}, \hat{Q}_{Z,n}, \hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n}), \hat{g}_n, \hat{\psi}_{ee}\right) = 0.$$

Therefore, the $\hat{\psi}_{ee}$ estimator and the proposed TMLE 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 $\hat{g}_n(A_i|W_i)$ 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 EE estimator may yield estimates that are out of the bounds of the parameter, since constraints such as bounds of the parameter are not reflected in the functional form of the efficient score. 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 root $\hat{\psi}_{ee}$ 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 implicitly estimates the boundary of the parameter domain, and the virtue of the substitution principle.

4.2 G-computation Approach

The sensitivity to near positivity violation of the TMLE estimator and the $\hat{\psi}_{ee}$ estimator stems 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 $\hat{\bar{Q}}_{Y,n}$ and $\hat{Q}_{Z,n}$ likelihood based estimators of the outcome expectation and mediator density, respectively, consider a g-computation estimator given by:

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

This estimator can be similarly defined using a regression-based $\hat{\psi}_{Z,n}(\hat{\bar{Q}}_{Y,n})$ which does not use Q_Z. Unlike the robust TMLE and $\hat{\psi}_{ee}$ estimators, the consistency of the g-computation estimator relies on correct specification of both the outcome expectation, and mediator density (or the regression procedure for the mediated mean outcome difference). 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 can still be affected by data sparsity, since the quality of the mean outcome estimate (even under the correct specification) is sensitive to the overlap between the empirical covariate distribution of the treated cohort and the empirical covariate distribution of the control cohort.

5 Simulation Study

In this section we evaluate the performance of the targeted estimator, the $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$ are 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: $Z \in \{0,1,2\}$. 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.

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

version a:

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

version b:

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

The treatment probability $g_A(A=1|w)$, is bounded in $(0.26,0.94)$. The conditional density $Q_Z(z|A=1,w)$ is bounded between $(0.0005,0.9753)$ for any z and w, whereas the ratio $Q_Z(z|A=0,w)/Q_Z(z|A=1,w)$ is bounded in $(0.13,2.02)$. 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 $var(D^*(P_0)) \approx 1.157$ for the binary version, and $var(D^*(P_0)) \approx 7.967$ for the continuous version.

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

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

Conditional distributions for W,A,Y are the same as simulation 1. The conditional mediator density $Q_Z(z|w,A=1)$ ranges in $(0.017,0.081)$ for $Z=0$, ranges in $(0.046,0.697)$ for $Z=1$ and ranges in $(0.256,0.936)$ for $Z=2$. The ratio $\frac{Q_Z(z|w,A=0)}{Q_Z(z|w,A=1)}$ ranges in $(6.583,10.543)$ for $Z=0$, ranges in $(0.717,13.826)$ for $Z=1$ and ranges in $(0.0018,0.253)$ 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 $var(D^*(P_0)) \approx 3.721905$ for the binary version, and $var(D^*(P_0)) \approx 17.53054$ 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 $\hat{\psi}_{ee}$ will be implemented using these likelihood-based estimators as described in section 4.

The TMLE 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 $\hat{\bar{Q}}_{Y,n}$, using bounds given by the range of the observed outcomes and the predicted outcomes under $\hat{\bar{Q}}_{Y,n}$. After obtaining the targeted estimator $\hat{\bar{Q}}^*_{Y,n}$ on unit scale using logistic working submodel, we perform a second linear transformation T_2 to bound the difference $\hat{\bar{Q}}^*_{Y,n}(W,1,Z) - \hat{\bar{Q}}^*_{Y,n}(W,0,Z)$ in the unit interval, and obtain the targeted estimator $\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})$ using logistic working submodel. Finally, we apply the inverse map T_2^{-1} to $\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})$ and then T_1^{-1} to the final effect estimate.

We will consider two implementations of TMLE which differ in their initial estimator of the mediated mean outcome difference $\psi_Z(Q_{Z,0}, \bar{Q}_{Y,0})$. In TMLE 1, the initial estimator is given by a plug-in estimator $\hat{\psi}_{Z,n}(\hat{\bar{Q}}^*_{Y,n}) \equiv \psi_Z(\hat{Q}_{Z,n}, \hat{\bar{Q}}^*_{Y,n})$, using $\hat{Q}_{Z,n}$ and the updated $\hat{\bar{Q}}^*_{Y,n}$. In TMLE 2, the initial estimator $\hat{\psi}_{Z,n}(\hat{\bar{Q}}^*_{Y,n})(W)$ 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, this initial estimator in TMLE 2 is incorrect regardless of the consistency of $\hat{\bar{Q}}_{Y,n}$. However, from lemma 1, we expect TMLE 2 to be consistent in the cases (ii) and (iii) of lemma 1, 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. In the tables below, notations for model specifications are as follows:

notation	model specifications
qy.c, qz.c, ga.c	correct $\bar{Q}_Y$, correct Q_Z, correct g
qy.c, qz.c, ga.m	correct $\bar{Q}_Y$, correct Q_Z, mis-specified g
qy.c, qz.m, ga.c	correct $\bar{Q}_Y$, mis-specified Q_Z, correct g
qy.m, qz.c, ga.c	mis-specified $\bar{Q}_Y$, correct Q_Z, correct g
qy.c, qz.c, ga.tr	correct $\bar{Q}_Y$, correct Q_Z, truncated g
qy.c, qz.m, ga.tr	correct $\bar{Q}_Y$, mis-specified Q_Z, truncated g
qy.m, qz.c, ga.tr	mis-specified $\bar{Q}_Y$, correct Q_Z, truncated g

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 $var(D^*(P_0)) \approx 1.157$ for the binary version, and $var(D^*(P_0)) \approx 7.967$ 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 $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$ 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
ee: 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
ee: 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
ee: 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
ee: 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
$\bar{Q}_Y$ 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
ee: 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
ee: qy.c, qz.c, ga.m	1.142e-03	9.824e-04	1.844e-02	1.886e-03	1.844e-02	1.887e-03
$\bar{Q}_Y$ 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
ee: 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
ee: 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 $var(D^*(P_0)) \approx 3.721905$ for the binary version, and $var(D^*(P_0)) \approx 17.53054$ for the continuous version. Therefore, $var(D^*(P_0)/n$ are approximately $7.444e-03$ 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 larger 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 $\hat{\psi}_{ee}$ are also consistent but less efficient. In all cases, TMLE and $\hat{\psi}_{ee}$ 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
$\bar{Q}_Y$ 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
ee: 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
ee: qy.c, qz.c, ga.m	5.140e-03	6.736e-04	8.330e-03	7.693e-04	8.357e-03	7.697e-04
$\bar{Q}_Y$ 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
ee: 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
ee: 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
$\bar{Q}_Y$ 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
ee: 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
ee: 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
ee: 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
ee: 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 > 0.99 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').

When the treatment model values are extreme, the robustness results of lemma 1 no longer apply. We observe here that the MSE of TMLE and $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$ with an incorrect treatment mechanism (either through truncation or incorrect modeling) yield MSE that are proportional to sample size. This last result is predicted by the robustness result (i) of lemma 1 since the mis-specified treatment models is bounded away from 1. We observe also that in this simulation scheme, 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 $\hat{\psi}_{ee}$ 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.

In this simulation, we observe that TMLE and $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$ have a different bias-variance trade-off, with TMLE having smaller variance but larger bias, with respect to $\hat{\psi}_{ee}$ (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 $\hat{\psi}_{ee}$; 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 $\hat{\psi}_{ee}$ provide similar bias reduction over g-computation estimator; but TMLE has a smaller variance than $\hat{\psi}_{ee}$ when the untruncated treatment mechanism is used, while the opposite is true with the truncated treatment mechanism.

In 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

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
ee: 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
ee: 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
ee: 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
ee: 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
ee: 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
ee: 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
ee: qy.m, gz.c, ga.tr	7.356e-02	7.080e-02	6.202e-03	6.226e-04	1.161e-02	5.635e-03

the larger variances. We also observe that in smaller sample size, when using an untruncated correct treatment mechanism, the TMLE 1 has a larger bias but substantially smaller variance than the $\hat{\psi}_{ee}$. This is likely due to some large effect estimates in $\hat{\psi}_{ee}$ in the dataset with smaller sample size. The variance of $\hat{\psi}_{ee}$ decreases substantially when sample size increases. On the other hand, under the truncated treatment mechanism, $\hat{\psi}_{ee}$ 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, under untruncated treatment mechanism, we observe again that $\hat{\psi}_{ee}$ 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 $\hat{\psi}_{ee}$. If the outcome model is incorrect: when the treatment is not truncated, TMLE 1 has larger bias and smaller variance than $\hat{\psi}_{ee}$, and that relation is reversed under truncation.

Table 6: Simulation 3: Continuous 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	2.390e-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
ee: 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
ee: 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
ee: qy.c, qz.c, ga.m	6.349e-03	4.447e-03	1.497e-01	1.228e-02	1.497e-01	1.230e-02
$\bar{Q}_Y$ 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
ee: 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
ee: 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
ee: 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
ee: 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 Extension to Natural Indirect Effect

In this section, we extend the above discussions in an analogous fashion to address the natural indirect effect.

In the context of natural effects, the total effect of A on Y can be decomposed into natural indirect and direct effects (Robins and Greenland (1992), Pearl (2001), Robins (2003)):

$$\begin{aligned} &E(Y(1)-Y(0)) \\ &= [E(Y(1,Z(1))-E(Y(1,Z(0))] + [E(Y(1,Z(0))-E(Y(0,Z(0))], \end{aligned}$$

where $Y(a)$ represents the restriction to set $Y(a) \equiv f_Y(W, A=a, Z=Z(a), U_Y)$ on the NPSEM. This decomposition formalizes the concept that the total effect of the exposure on the outcome is a combination of its indirect effect through a mediator Z, and its direct effect not mediated by Z. The quantity $E(Y(1,Z(1))-E(Y(1,Z(0))$ is referred to as the *additive natural indirect effect*. Its identification is studied in the same body of literature (Robins and Greenland (1992), Pearl (2001), Robins

(2003), Petersen et al. (2006), Hafeman and VanderWeele (2010), Imai et al. (2010), Robins and Richardson (2010) and Pearl (2011)). More specifically, under the same conditions as those in section 2.2, the natural indirect effect can be identified as

$$\begin{aligned} & E(Y(1,Z(1)) - E(Y(1,Z(0)) \overset{A1,A2,A3}{=} \Psi_{NIE}(P_0) \\ & \equiv P_{W,0}\left\{\sum_z \bar{Q}_{Y,0}(W,A=1,z)\left[Q_{Z,0}(z|W,A=1) - Q_{Z,0}(z|W,A=0)\right]\right\}. \end{aligned} \tag{20}$$

The results of Robins and Richardson (2010) thus have the same implications on the difficulty of identifying the natural indirect effect in real experiments, due to the conditional counterfactual independence assumption A3. In such cases, what kind of causal interpretation can the statistical parameter (20) still offer? If assumption A3 fails but randomization assumptions A1 and A2 hold, the statistical parameter in (20) equals

$$\Psi_{NIE}(P_0) \overset{A1,A2}{=} E_W\left\{\sum_z E(Y(1,z)|W)\left[p(Z(1)=z|W) - p(Z(0)=z|W)\right]\right\}.$$

The interpretation of the right hand side is not as intuitive as in the natural direct effect case. But since $p(Z(1)=z|W) - p(Z(0)=z|W)$ measures the effect of A on Z, at its face value this alternative effect parameter can be viewed as weighting the different outcomes $E(Y(1,z)|W)$ under z by these effect measures. However, we remind the reader again that this alternative causal parameter only serves to provide a causal interpretation for the statistical parameter (20) and one should be cautious about putting it into the context of the traditional total effect decomposition.

The parameter $\Psi_{NIE}(P)$ is also a function of Q alone. To extend the discussions above to the natural indirect effect parameter (20), we now consider the *mediated mean outcome* map $Q \mapsto \psi_{NIE,Z}(Q)$, where $\psi_{NIE,Z}(Q) : \mathscr{A} \times \mathscr{W} \to \mathbb{R}$ is given by

$$(w,a) \mapsto \psi_{NIE,Z}(Q)(w,a) \equiv E_{Q_Z}\left(\bar{Q}_Y(W=w,A=1,Z)|W=w,A=a\right).$$

This way, the parameter can be regarded as $\Psi_{NIE}(Q) = \Psi_{NIE}\left(Q_W, \psi_{NIE,Z}(Q)\right)$.

The efficient score for this parameter (derived inTchetgen Tchetgen and Shpitser (2011b)) is given by

$$\begin{aligned} & D^*_{NIE}(Q,g,\Psi_{NIE}(Q)) \\ & = \frac{I(A=1)}{g(1|W)}\left\{Y - \psi_{NIE,Z}(Q)(W,1) - \frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)}\left(Y - \bar{Q}_Y(W,1,Z)\right)\right\} \\ & - \frac{I(A=0)}{g(0|W)}\left(\bar{Q}_Y(W,1,Z) - \psi_{NIE,Z}(Q)(W,0)\right) \\ & + \psi_{NIE,Z}(Q)(W,1) - \psi_{NIE,Z}(Q)(W,0) - \Psi_{NIE}(Q). \end{aligned} \tag{21}$$

The general robustness conditions of Tchetgen Tchetgen and Shpitser (2011b) apply to both natural direct and indirect effects. By the same reasoning (and analogous proof) as that of lemma 1, we note again that conditions (i) and (iii) may be weakened to: (i) the conditional mean outcome $\bar{Q}_Y(W,A,Z)$ and the mediated outcome map $\psi_{NIE,Z}(Q)(W,A)$ are both correct; (iii) the exposure mechanism and mediator density, or the exposure mechanism and the conditional distribution $p(A|W,Z)$, are correct. Therefore, in situations where Z is high dimensional, similar practical implications as those discussed in remarks following lemma 1 apply. However, note that a regression-based estimation procedure for $\psi_{NIE,Z}(Q_0)$ now regresses $\bar{Q}_Y(W,1,Z)$ on W among treated observations to obtain the conditional mean $\psi_{NIE,Z}(Q)(W,1)$, and among control observations to obtain $\psi_{NIE,Z}(Q)(W,0)$.

Since the parameter (20) is given by

$$\Psi_{NIE}(Q) = E_{Q_W}\Big(\psi_{NIE,Z}(Q)(W,1) - \psi_{NIE,Z}(Q)(W,0)\Big), \tag{22}$$

the targeted MLE only needs to focus on estimation of the components $Q_{W,0}$, $\bar{Q}_{Y,0}$ and $\psi_{NIE,Z}(Q_0)$ of the likelihood. We first rewrite the efficient score in (21) as

$$\begin{aligned}
&D^*_{NIE}(Q,g,\Psi_{NIE}(Q)) \\
&= \frac{I(A=1)}{g(1|W)}\left(1-\frac{Q_Z(Z|W,0)}{Q_Z(Z|W,1)}\right)\big(Y-\bar{Q}_Y(W,A,Z)\big) \\
&+ \frac{2A-1}{g(A|W)}\left\{\bar{Q}_Y(W,1,Z)) - \psi_{NIE,Z}(Q)(W,A)\right\} \\
&+ \psi_{NIE,Z}(Q)(W,1) - \psi_{NIE,Z}(Q)(W,0) - \Psi_{NIE}(Q) \\
&\equiv D^*_{NIE,Y} + D^*_{NIE,Z} + D^*_{NIE,W}.
\end{aligned}$$

The reader may have readily noted the parallel between $D^*_{NIE,Z} + D^*_{NIE,W}$ and the efficient score for the familiar additive marginal treatment effect; this is because the indirect effect can viewed as an additive marginal effect of the treatment on $\bar{Q}_Y(W,A=1,Z)$ through its effect on Z, as seen in (22). In fact, as we will see shortly, the second part of the implementation of TMLE is very similar to the well-known case of additive marginal effects.

Without loss of generality, we assume that Y is bounded in the unit interval. Under the log-likelihood loss function of (3), the least favorable submodel for $\bar{Q}_Y(W,A,Z)$ through a given initial estimator $\hat{\bar{Q}}_{Y,n}$ is now given by

$$\hat{\bar{Q}}_{Y,n}(\varepsilon_1) \equiv expit\left(logit(\hat{\bar{Q}}_{Y,n}) + \varepsilon_1 C_Y(\hat{Q}_{Z,n},\hat{g}_n)\right),$$

where $C_Y(\hat{Q}_{Z,n},\hat{g}_n)(O) = \frac{I(A=1)}{\hat{g}_n(1|W)}\left(1-\frac{\hat{Q}_{Z,n}(Z|W,0)}{\hat{Q}_{Z,n}(Z|W,1)}\right)$. Note that the dependence of $\hat{\bar{Q}}_{Y,n}(\varepsilon_1)$ on $\hat{Q}_{Z,n}$ and $\hat{g}_n$ are suppressed in the notation. The targeted MLE of $\bar{Q}_{Y,0}$ is $\hat{\bar{Q}}^*_{Y,n} \equiv \hat{\bar{Q}}_{Y,n}(\hat{\varepsilon}^*_1)$ and is similarly defined as in section 3.1.

Next, consider an estimating procedure $\hat{\psi}_{NIE,Z,n}(P_n)(\cdot)$ for $\psi_{NIE,Z}(Q_0)$, and let $\hat{\psi}_{NIE,Z,n} \equiv \hat{\psi}_{NIE,Z}(P_n)$. We are reminded that the function $\hat{\psi}_{NIE,Z,n}$ depends on the estimating procedure $\hat{\psi}_{NIE,Z}(\cdot)$ and the observed data P_n, and it can be plug-in or regressed-based. $\hat{\psi}_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})$ is an initial estimator of $\psi_{NIE,Z}(Q_0)$. We define the log-likelihood loss for $\psi_{NIE,Z}(Q)(W,A)$ as

$$\begin{aligned} &L_Z(\psi_{NIE,Z}(Q))(O) \\ &= -\log\left\{\psi_{NIE,Z}(Q)(W,A)^{\bar{Q}_Y(W,1,Z)}\left(1-\psi_{NIE,Z}(Q)(W,A)\right)^{1-\bar{Q}_Y(W,1,Z)}\right\}. \end{aligned}$$

The least favorable submodel through the initial estimator $\hat{\psi}_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})$ is given by

$$\hat{\psi}_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})(\varepsilon_2) = expit\left(logit\left(\hat{\psi}_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) + \varepsilon_2 C_Z(\hat{g}_n)\right),$$

where $C_Z(\hat{g}_n) = \frac{2A-1}{\hat{g}_n(A|W)}$. The dependence of the submodel on $\hat{g}_n$ is also suppressed in the notation. In a similar fashion as section 3.1, we obtain the targeted MLE $\hat{\psi}^*_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n}) \equiv \hat{\psi}_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})(\hat{\varepsilon}^*_2)$. Finally, the targeted MLE of the parameter $\Psi_{NIE}(Q_0)$ is given by

$$\hat{\psi}^*_{NIE,n} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}^*_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})(W_i,1) - \hat{\psi}^*_{NIE,Z,n}(\hat{\bar{Q}}^*_{Y,n})(W_i,0)\right).$$

We remind the reader again that the role of the ratio of Q_Z in C_Y may be replaced by ratios of $g(A|W)$ and $p(A|W,Z)$.

The resulting estimator satisfies the efficient score equation, and therefore is asymptotically unbiased if (i) the conditional mean outcome $\bar{Q}_Y$ and the mediated outcome map $\psi_{NIE,Z}(Q)$ are both correct; (ii) the conditional mean outcome and the exposure mechanism $g(A|W)$ are correct; (iii) the exposure mechanism and mediator density $Q_Z(Z|W,A)$, or the exposure mechanism and the conditional distribution $p(A|W,Z)$, are correct. An estimating equation estimator $\hat{\psi}^{ee}_{NIE}$ is also discussed in Tchetgen Tchetgen and Shpitser (2011b). As mentioned in section 4, $\hat{\psi}^*_{NIE}$ and $\hat{\psi}^{ee}_{NIE}$ will inherit the same robustness properties from the efficient score, since both satisfy the efficient score equation. Conditions for asymptotic linearity are analogous to those of theorem 1, we omit their derivations here.

7 Summary and Concluding Remarks

In this article, we applied the targeted maximum likelihood framework of van der Laan and Rubin (2006) and van der Laan and Rose (2011) to construct a semiparametric efficient, multiply robust, plug-in estimator for the natural direct effect of a binary treatment. This estimator has the property that it satisfies the efficient

score equation (derived in Tchetgen Tchetgen and Shpitser (2011b)), and hence also inherits its robustness properties. We noted that the robustness conditions in Tchetgen Tchetgen and Shpitser (2011b) may be weakened (lemma 1), thereby placing less reliance on the estimation of the mediator density. More precisely, the proposed estimator is asymptotically unbiased if either one of the following holds: i) the conditional mean outcome given exposure, mediator, and confounders, and the mediated mean outcome difference are consistently estimated; (ii) the exposure mechanism given confounders, and the conditional mean outcome are consistently estimated; or (iii) the exposure mechanism and the mediator density, or the exposure mechanism and the conditional distribution of the exposure given confounders and mediator, are consistently estimated. If all three conditions hold, then the effect estimate is asymptotically efficient. We also extended our results analogously to the case of natural indirect effect.

In applications, the components that are difficult to estimate are often times the conditional mean outcome and/or the mediator density. For a high-dimensional Z, few tools are available to estimate the conditional mediator density Q_Z. On the other hand, there is abundant literature addressing estimation of conditional means. This can be used to estimate the mediated mean outcome difference $\psi_Z(Q) \equiv E_{Q_Z}(\bar{Q}_Y(W,1,Z) - \bar{Q}_Y(W,0,Z)|W,A=0)$, and the conditional distributions of a categorical A. Lemma 1 implies that estimation of the mediator density may be replaced by estimations of $g(A|W)$, $p(A|W,Z)$, and the conditional expectation $\psi_Z(Q)$.

We have also described general conditions for the estimator to be asymptotically linear. More specifically, 1) estimators of each component must converge to their respective limits at a reasonable speed, and 2) if there is a component that is not consistently estimated, the consistent estimators of the remaining components must meet stricter asymptotic linearity conditions. These conditions provide a guideline for situations where influence curve based variance estimates are realistic.

Estimators that use of the efficient score are robust, but are generally sensitive to practical positivity violations. We refer to Petersen, Porter, S.Gruber, Wang, and van der Laan (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 aim to provide more stable estimates in such situations. However, identification of the parameter depends ultimately on the information available in a given finite sample. A way to improve finite sample robustness is the Collaborative TMLE (C-TMLE) 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 that explain the residual bias of the estimator of Q. We aim to investigate applications of C-TMLE to the effect mediation problem.

DOI: 10.2202/1557-4679.1361

Appendix A

App1. Proof of lemma 1

Let $\tilde{\psi}_Z$ be a map $Q \mapsto \tilde{\psi}_Z(Q)$, where $\tilde{\psi}_Z(Q)$ is a function from $\mathscr{W}$ to $\mathbb{R}$. Note that $\tilde{\psi}_Z(Q)$ may or may not make use of the density Q_Z, but it surely uses $\bar{Q}_Y$. Then

$$
\begin{aligned}
& P_0 D^*(Q, g, \tilde{\psi}_Z(Q), \psi_0) \\
& = P_{W,0}\left\{ \frac{g_0(1|W)}{g(1|W)} \sum_z Q_{Z,0}(z|W,1) \frac{Q_Z(z|W,0)}{Q_Z(z|W,1)} \left(\bar{Q}_{Y,0}(W,1,z) - \bar{Q}_Y(W,1,z)\right) \right\} && (23) \\
& - P_{W,0}\left\{ \frac{g_0(0|W)}{g(0|W)} \sum_z Q_{Z,0}(z|W,0) \left(\bar{Q}_{Y,0}(W,0,z) - \bar{Q}_Y(W,0,z)\right) \right\} && (24) \\
& + P_{W,0}\left\{ \frac{g_0(0|W)}{g(0|W)} \sum_z Q_{Z,0}(z|W,0) \left(\bar{Q}_Y(W,1,z) - \bar{Q}_Y(W,0,z)\right) \right\} && (25) \\
& - P_{W,0}\left\{ \frac{g_0(0|W)}{g(0|W)} \tilde{\psi}_Z(Q)(W) \right\} && (26) \\
& + P_{W,0}\left\{ \tilde{\psi}_Z(Q)(W) \right\} - \psi_0 && (27)
\end{aligned}
$$

Suppose (i) holds, i.e. $\bar{Q}_Y = \bar{Q}_{Y,0}$ and $\tilde{\psi}_Z(Q)(W) = \psi_Z(Q_0)(W)$. Then (23) and (24) are each exactly 0; the expectation in (25) and (26) are the same; and $P_{W,0}\tilde{\psi}_Z(Q)(W) = P_{W,0}\psi_Z(Q_0)(W) = \psi_0$. Notice that in this case, it was not necessary that $Q_Z = Q_{Z,0}$. But rather, any function $\tilde{\psi}_Z(Q)$ that equals the true mediated mean difference $\psi_Z(Q_0)$ will yield the desired result.

Suppose now that (ii) holds. Then (23) and (24) are each exactly 0. The expression in (26) equals $P_{W,0}\tilde{\psi}_Z(Q)(W)$, and the expression in (25) equals ψ_0. Therefore the mean is zero.

Suppose that (iii) holds. Then, rearranging (23) and (24) we rewrite the above expectation as

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

App2. Proof of theorem 1

To see (9) we note firstly that for any Q and ψ

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

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

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

where the last summand can be rewritten as

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

Result (9) thus follows. Moreover, the Donsker class condition in (10) yields

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

The conditions for asymptotic linearity can be ascertained from the second order terms by a straightforward expansion:

$$
\begin{aligned}
&P_{W,0}\sum_z \left(\bar{Q}_{Y,0}(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right) Q_{Z,0}(z|W,1)\left(\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-\frac{Q_{Z,0}(z|W,0)}{Q_{Z,0}(z|W,1)}\right)\\
&+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\hat{\bar{Q}}^*_{Y,n}\right)\\
&+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(\psi_Z(Q_{Z,0},\hat{\bar{Q}}^*_{Y,n})-\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right)\\
&=P_{W,0}\sum_z \left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}^*_Y(W,1,z)\right) Q_{Z,0}(z|W,1)\left(\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)}-\frac{Q_{Z,0}(z|W,0)}{Q_{Z,0}(z|W,1)}\right)\\
&+P_0\left(C_Y(g,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\bar{Q}^*_Y\right)\\
&+P_0\left(\frac{I(A=0)}{g(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(\psi_Z(Q_{Z,0},\hat{\bar{Q}}^*_{Y,n})-\psi^*_Z(\hat{\bar{Q}}^*_{Y,n})\right)\\
&+P_{W,0}\sum_z \left(\bar{Q}^*_Y(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right) Q_{Z,0}(z|W,1)\left(\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)}\right) \qquad (28)\\
&+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g,\hat{Q}_{Z,n})\right)\left(\bar{Q}^*_Y-\hat{\bar{Q}}^*_{Y,n}\right) \qquad (29)\\
&+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g(0|W)}\right)\left(\psi^*_Z(\hat{\bar{Q}}^*_{Y,n})-\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right) \qquad (30)
\end{aligned}
$$

$$
\begin{aligned}
&+P_{W,0}\sum_z \left(\bar{Q}^*_Y(W,1,z)-\hat{\bar{Q}}^*_{Y,n}(W,1,z)\right) Q_{Z,0}(z|W,1)\left(\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)}-\frac{Q_{Z,0}(z|W,0)}{Q_{Z,0}(z|W,1)}\right) \qquad (31)\\
&+P_{W,0}\sum_z \left(\bar{Q}_{Y,0}(W,1,z)-\bar{Q}^*_Y(W,1,z)\right) Q_{Z,0}(z|W,1)\left(\frac{\hat{Q}_{Z,n}(z|W,0)}{\hat{Q}_{Z,n}(z|W,1)}-\frac{Q_Z(z|W,0)}{Q_Z(z|W,1)}\right) \qquad (32)\\
&+P_0\left(C_Y(g,\hat{Q}_{Z,n})-C_Y(g_0,\hat{Q}_{Z,n})\right)\left(\bar{Q}^*_Y-\hat{\bar{Q}}^*_{Y,n}\right) \qquad (33)\\
&+P_0\left(C_Y(\hat{g}_n,\hat{Q}_{Z,n})-C_Y(g,\hat{Q}_{Z,n})\right)\left(\bar{Q}_{Y,0}-\bar{Q}^*_Y\right) \qquad (34)\\
&+P_0\left(\frac{I(A=0)}{\hat{g}_n(0|W)}-\frac{I(A=0)}{g(0|W)}\right)\left(\psi_Z(Q_{Z,0},\hat{\bar{Q}}^*_{Y,n})-\psi^*_Z(\hat{\bar{Q}}^*_{Y,n})\right) \qquad (35)\\
&+P_0\left(\frac{I(A=0)}{g(0|W)}-\frac{I(A=0)}{g_0(0|W)}\right)\left(\psi^*_Z(\hat{\bar{Q}}^*_{Y,n})-\hat{\psi}^*_{Z,n}(\hat{\bar{Q}}^*_{Y,n})\right). \qquad (36)
\end{aligned}
$$

In this theorem we study situations pertaining to the cases (i) $\bar{Q}^*_Y=\bar{Q}_{Y,0}$, and $\psi^*_Z(\hat{\bar{Q}}^*_{Y,n})=\psi_Z(Q_{Z,0},\hat{\bar{Q}}^*_{Y,n})$; (ii) $g=g_0$ and $\bar{Q}^*_Y=\bar{Q}_{Y,0}$; or (iii) $g=g_0$, $Q_Z=Q_{Z,0}$. Under either of these cases, the first three unlabeled summands after the equal sign are exactly zero. Therefore, we only need to focus on the first order ((31), (32),(33), (34), (35), (36)) and second order ((28), (29), (30)) remainders. The rate conditions (11), (12) and (13) ensure that the second order terms (28), (29) and (30) are all

$o_P(1/\sqrt{n})$. The remaining case by case asymptotic linearity conditions ensure that the first order remainders are asymptotically linear.

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 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 $\bar{Q}_Y$ 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 $\bar{Q}_Y$ 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 parametric working submodel $Q_Z(\varepsilon)$ satisfying

$$\frac{d}{d\varepsilon} L_Z\left(Q_Z(\varepsilon)\mid_{\varepsilon=0}\right) = D_Z(Q_Z,g,\bar{Q}_Y). \tag{37}$$

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

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

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

$$logit Q_{Z_k}(g,\bar{Q}_Y)(\varepsilon)(1|W,A) = logit Q_{Z_k}(1|W,A) + \varepsilon C_{Z_k}(g,\bar{Q}_Y)(W,A),$$

where we define

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

if $\mathbf{Z_{k-1}}=\mathbf{0}$, and $C_{Z_k}(g,\bar{Q}_Y)(W,A)\equiv 0$ if $\mathbf{Z_{k-1}}\neq\mathbf{0}$. This way, the parametric working submodel $Q_Z(g,\bar{Q}_Y)(\varepsilon)=\prod_{k=0}^{K-1}Q_{Z_k}(g,\bar{Q}_Y)(\varepsilon)$ satisfies (37).

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 (6). Using this updated $\hat{\bar{Q}}_Y^*$, the optimal ε for the submodel of Q_Z is given by

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

and the targeted estimator of the mediator density is given by $\hat{Q}_Z(\hat{g},\hat{\bar{Q}}_Y^*)(\hat{\varepsilon}^*)$, 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:

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

It follows from (4) that $P_nD_Y^*(\hat{\bar{Q}}_Y^*,\hat{Q}_Z,\hat{g})=0$, and it follows from (37) that $P_nD_Z^*\left(\hat{\bar{Q}}_Y^*,\hat{Q}_Z^*,\hat{g}\right)=0$. Moreover, the empirical distribution $\hat{Q}_{W,n}$ of W solves the score equation $P_nD_W^*(\hat{\bar{Q}}_Y^*,\hat{Q}_Z^*,\hat{Q}_{W,n})=0$. Therefore the resulting targeted estimator also solves the efficient score equation.

References

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

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1997): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag.

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

Gruber, S. and M. van der Laan (2011): "Bounded continuous outcomes," in M. van der Laan and S. Rose, eds., *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer.

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

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

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

Jo, B., E. Stuart, D. MacKinnon, and A. Vinokur (2011): "The use of propensity scores in mediation analysis," *Multivariate Behavioral Research*, 46, 425–452.

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

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

Kosorok, M. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer-Verlag.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, Citeseer, 411–420.

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

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

Petersen, M., K. Porter, S.Gruber, Y. Wang, and M. van der Laan (2010): "Diagnosing and responding to violations in the positivity assumption," Technical report 269, Division of Biostatistics, University of California, Berkeley, URL

Petersen, M., S. Sinisi, and M. van der Laan (2006): "Estimation of direct causal effects." *Epidemiology*, 17, 276–284.

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

Robins, J. (2003): "Semantics of causal dag models and the identification of direct and indirect effects," in N. H. P. Green and S. Richardson, eds., *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, 70–81.

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

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

Robins, J. M. and T. S. Richardson (2010): "Alternative graphical causal models and the identification of direct effects." in P. Shrout, ed., *In Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, Oxford University Press.

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

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

Tchetgen Tchetgen, E. and I. Shpitser (2011a): "Semiparametric estimation of models for natural direct and indirect effects," Technical report 129, Biostatistics, Harvard University, URL

Tchetgen Tchetgen, E. and I. Shpitser (2011b): "Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis," Technical report 130, Biostatistics, Harvard University, URL

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

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

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

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

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

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

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

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

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

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

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

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

Zheng, W. and M. van der Laan (2011): "Cross-validated targeted minimum-loss-based estimation," in M. van der Laan and S. Rose, eds., *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer.

DOI: 10.2202/1557-4679.1361