Research Article

Open Access

David Benkeser* and Jialu Ran

Nonparametric inference for interventional effects with multiple mediators

DOI: DOI, Received ..; revised ..; accepted ..

Abstract: Understanding the pathways whereby an intervention has an effect on an outcome is a common scientific goal. A rich body of literature provides various decompositions of the total intervention effect into pathway specific effects. Interventional direct and indirect effects provide one such decomposition. Existing estimators of these effects are based on parametric models with confidence interval estimation facilitated via the nonparametric bootstrap. We provide theory that allows for more flexible, possibly machine learning-based, estimation techniques to be considered. In particular, we establish weak convergence results that facilitate the construction of closed-form confidence intervals and hypothesis tests and prove multiple robustness properties of the proposed estimators. Simulations show that inference based on large-sample theory has adequate small-sample performance. Our work thus provides a means of leveraging modern statistical learning techniques in estimation of interventional mediation effects.

Keywords: Mediation, Causal inference, Augmented inverse probability of treatment weighted estimator, Targeted minimum loss estimator, Machine learning

MSC: 62G05,62G08,62G20

1 Introduction

Recent advances in causal inference have provided rich frameworks for posing interesting scientific questions pertaining to the mediation of effects through specific biologic pathways (among others, Imai et al. [11], Naimi et al. [13], Pearl [15], Valeri and VanderWeele [20], VanderWeele and Tchetgen Tchetgen [24], Yuan and MacKinnon [28], Zheng and van der Laan [30]). Foremost amongst these advances is the provision of model-free definitions of mediation parameters, which enables researchers to develop robust estimators of these quantities. A debate in this literature has emerged pertaining to the reliance of methodology on *cross-world* independence assumptions that are fundamentally untestable even in randomized controlled experiments [7, 14, 17]. One approach to this problem is to utilize methods that attempt to estimate bounds on effects (among others, Robins and Richardson [17], Tchetgen Tchetgen and Phiri [19]). A second approach considers seeking alternative definitions of mediation parameters that do not require such cross-world assumptions (among others, Rudolph et al. [18], VanderWeele et al. [25]). Rather than considering deterministic interventions on mediators (i.e., a hypothetical intervention that fixes every every individuals mediator to a particular value), these approaches consider *stochastic* interventions on mediators (i.e., hypothetical interventions where the mediator is drawn from a particular conditional distribution). In this class of approaches, that of Vansteelandt and Daniel [26] is particularly appealing. Building on the prior work of VanderWeele et al. [25], the authors provide a simple decomposition of the total effect into direct effects and pathway-specific effects via multiple mediators. Interestingly, their decompositions hold even when the structural dependence between mediators is unknown.

Vansteelandt and Daniel [26] described two approaches to estimation of the effects using parametric working models for relevant nuisance parameters. In both cases, the nonparametric bootstrap was rec-

***Corresponding Author: David Benkeser, Jialu Ran:** Emory University Rollins School of Public Health; E-mail: benkeser@emory.edu

ommended for inference. A potential limitation of the proposal is that correctly specifying a parametric working model may be difficult in many settings. In these instances, we may rely on flexible estimators of nuisance parameters, for example, based on machine learning. When such techniques are employed, the nonparametric bootstrap does not generally guarantee valid inference [6]. This fact motivates the present work, where we develop nonparametric efficiency theory for the interventional mediation effect parameters. This theory allows us to utilize frameworks for nonparametric efficient inference to develop estimators of the quantities of interest. We propose a one-step and a targeted minimum loss-based estimator and demonstrate that under suitable regularity conditions, both estimators are nonparametric efficient amongst the class of regular asymptotically linear estimators. The estimators also enjoy a multiple robustness property, which ensures consistency of effect estimates if at least some combination of nuisance parameters are consistently estimated. Another benefit enjoyed by our estimators is the availability of closed-form confidence intervals and hypothesis tests.

2 Interventional Effects

Adopting the notation of Vansteelandt and Daniel [26], suppose the observed data are represented as n independent copies of the random variable $O = (C, A, M_1, M_2, Y) \sim P$, where $C \in \mathcal{C}$ is a vector of confounders, $A \in \{a, a^\star\}$ is a binary intervention, $M_1 \in \mathcal{M}_1$ and $M_2 \in \mathcal{M}_2$ are mediators, and $Y \in \mathcal{Y}$ is a relevant outcome. Our developments pertain to both discrete and real-valued mediators, while without loss of generality, we assume $\mathcal{Y} = (0, 1)$. We assume $\text{pr}_P\{0 < \text{pr}_P(A = a \mid C) < 1\} = 1$; that is, any subgroup defined by covariates C that is observed with positive probability should have some chance of receiving both interventions. We also assume that for $a_0 = a, a^\star$, the probability distribution of (M_1, M_2) given $A = a_0, C$ has density $q_{a_0,M_1,M_2}(m_1, m_2 \mid c)$ with respect to some dominating measure and this density satisfies $\text{pr}_P\{\inf_{m_1,m_2} q_{a_0,M_1,M_2}(m_1, m_2 \mid C)\} > 0\} = 1$, where the infimum is taken over $\mathcal{M}_1 \times \mathcal{M}_2$. Similarly, we assume that for all $\sup_{c,m_1,m_2} q_{a_0,M_1,M_2}(m_1, m_2 \mid c) < \infty$. Beyond these conditions, $\mathcal{P}$ encodes no assumptions about P; however, the efficiency theory that we develop still holds under a model that makes assumptions about $\text{pr}_P(A \mid C)$, including the possibility that this quantity is known exactly, as in a stratified randomized trial.

To define interventional mediation effects, notation for counterfactual random variables is required. For $a_0 \in \{a, a^\star\}$, and $j = 1, 2$, let $M_j(a_0)$ denote the counterfactual value for the j-th mediator when A is set to a_0. Similarly, let $Y(a_0, m_1, m_2)$ denote the counterfactual outcome under an intervention that sets $A = a_0, M_1 = m_1$, and $M_2 = m_2$. As a point of notation, when introducing quantities whose definition depends on particular components of the random variable O, we will use lower case letters to denote the particular value and assume that the definition at hand applies for all values in the support of that random variable.

The total effect of intervening to set $A = a$ versus $A = a^\star$ is $\psi = \mathbb{E}\{Y(a, M_1(a), M_2(a))\} - \mathbb{E}\{Y(a^\star, M_1(a^\star), M_2(a^\star))\}$, where we use $\mathbb{E}$ to emphasize that we are taking an expectation with respect to a distribution of a counterfactual random variable. The total effect describes the difference in counterfactual outcome considering an intervention where we set $A = a$ and allow the mediators to naturally assume the value that they would under intervention $A = a$ versus an intervention where we set $A = a^\star$ and allow the mediators to vary accordingly. To contrast with forthcoming effects, it is useful to write the total effect in integral form. Specifically, we use $\bar{\mathbb{Q}}_{a_0}(m_1, m_2, c)$ to denote the covariate-conditional mean of the counterfactual outcome $Y(a_0, m_1, m_2)$, $\mathbb{Q}_{M_1(a_0),M_2(a_0)}(\cdot, \cdot \mid c)$ to denote the covariate-conditional bivariate cumulative distribution function of $(M_1(a_0), M_2(a_0))$, and Q_C to denote the marginal distribution of C.

The total effect can be written as

$$\psi = \int_{\mathcal{C}} \Bigg\{ \int_{\mathcal{M}_1 \times \mathcal{M}_2} \bar{\mathbb{Q}}_a(m_1, m_2, c) \; d\mathbb{Q}_{M_1(a), M_2(a)}(m_1, m_2 \mid c) - \int_{\mathcal{M}_1 \times \mathcal{M}_2} \bar{\mathbb{Q}}_{a^\star}(m_1, m_2, c) \; d\mathbb{Q}_{M_1(a^\star), M_2(a^\star)}(m_1, m_2 \mid c) \Bigg\} dQ_C(c) \ .$$

The total effect can be decomposed into interventional direct and indirect effects. The interventional direct effect is the difference in average counterfactual outcome under two population-level interventions. The first intervention sets $A = a$, and subsequently for individuals with $C = c$ draws mediators from $\mathbb{Q}_{M_1(a^\star), M_2(a^\star)}(\cdot \mid c)$. Thus, on a population level the covariate conditional distribution of mediators in this counterfactual world is the same as it would be in a population where everyone received intervention $A = a^\star$. This is an example of a stochastic intervention [12]. The second intervention sets $A = a^\star$, and subsequently allows the mediators to naturally assume the value that they would under intervention $A = a^\star$, so that the population level mediator distribution is again $\mathbb{Q}_{M_1(a^\star), M_2(a^\star)}(\cdot \mid c)$. The interventional direct effect compares the average outcome under these two interventions,

$$\psi_A = \int_{\mathcal{C}} \int_{\mathcal{M}_1 \times \mathcal{M}_2} \{\bar{\mathbb{Q}}_a(m_1, m_2, c) - \bar{\mathbb{Q}}_{a^\star}(m_1, m_2, c)\} d\mathbb{Q}_{M_1(a^\star), M_2(a^\star)}(m_1, m_2 \mid c) dQ_C(c) \ .$$

For interventional indirect effects, we require definitions for the covariate-conditional distribution of each mediator, which we denote for $j = 1, 2$ by $\mathbb{Q}_{M_j(a_0)}(\cdot \mid c)$. The interventional indirect effect through M_1 is

$$\psi_{M_1} = \int_{\mathcal{C}} \left[\int_{\mathcal{M}_2} \int_{\mathcal{M}_1} \bar{\mathbb{Q}}_a(m_1, m_2, c) \{ d\mathbb{Q}_{M_1(a)}(m_1 \mid c) - d\mathbb{Q}_{M_1(a^\star)}(m_1 \mid c) \} d\mathbb{Q}_{M_2(a^\star)}(m_2 \mid c) \right] \times dQ_C(c) \ .$$

As with the direct effect, this effect considers two interventions. Both interventions set $A = a$. The first intervention draws mediator values independently from the marginal mediator distributions $\mathbb{Q}_{M_1(a)}(\cdot \mid c)$ and $\mathbb{Q}_{M_2(a^\star)}(\cdot \mid c)$, while the second intervention draws mediator values independently from the marginal mediator distributions $\mathbb{Q}_{M_1(a^\star)}(\cdot \mid c)$ and $\mathbb{Q}_{M_2(a^\star)}(\cdot \mid c)$. The effect thus describes the average impact of shifting the population level distribution of M_1, while holding the population level distribution of M_2 fixed. The interventional indirect effect on the outcome through M_2 is similarly defined as

$$\psi_{M_2} = \int_{\mathcal{C}} \left[\int_{\mathcal{M}_1} \int_{\mathcal{M}_2} \bar{\mathbb{Q}}_a(m_1, m_2, c) d\mathbb{Q}_{M_1(a)}(m_1 \mid c) \{ d\mathbb{Q}_{M_2(a)}(m_2 \mid c) - d\mathbb{Q}_{M_2(a^\star)}(m_2 \mid c) \} \right] \times dQ_C(c) \ .$$

Note that when defining interventional indirect effects, mediators are drawn *independently* from marginal mediator distributions. The final effect in the decomposition essentially describes the impact of drawing the mediators from marginal rather than joint distributions. Thus, we term this effect the *covariant mediator effect*, defined as

$$\psi_{M_1, M_2} = \int_{\mathcal{C}} \int_{\mathcal{M}_1 \times \mathcal{M}_2} \bar{\mathbb{Q}}_a(m_1, m_2, c) \Big[d\mathbb{Q}_{M_1(a), M_2(a)}(m_1, m_2 \mid c) - d\mathbb{Q}_{M_1(a) \times M_2(a)}(m_1, m_2 \mid c) - \{ d\mathbb{Q}_{M_1(a^\star), M_2(a^\star)}(m_1, m_2 \mid c) - d\mathbb{Q}_{M_1(a^\star) \times M_2(a^\star)}(m_1, m_2 \mid c) \} \Big] dQ_C(c) \ ,$$

where $d\mathbb{Q}_{M_1(a_0) \times M_2(a_0)}(m_1, m_2 \mid c) = d\mathbb{Q}_{M_1(a_0)}(m_1 \mid c) d\mathbb{Q}_{M_2(a_0)}(m_2 \mid c)$. Vansteelandt and Daniel [26] discuss situations where these effects are of primary interest.

From the above definitions, we have the following effect decomposition $\psi = \psi_A + \psi_{M_1} + \psi_{M_2} + \psi_{M_1, M_2}$. These component effects can be identified using the observed data under the following assumptions:

(i) the effect of A on Y is unconfounded given C, $Y(a, M_1(a), M_2(a)) \perp\!\!\!\perp A \mid C$;

(ii) the effect of M_1 and M_2 on Y is unconfounded given A and C, $Y(a_0, M_1(a_0), M_2(a_0)) \perp\!\!\!\perp M_1, M_2 \mid A = a_0, C$;

(iii) the effect of A on M_1, M_2 is unconfounded given C, $M_1(a_0), M_2(a_0) \perp\!\!\!\perp A \mid C$.

Under these assumptions, the counterfactual mean $\bar{\mathbb{Q}}_{a_0}(m_1, m_2, c)$ is identified by $\bar{Q}_{a_0}(m_1, m_2, c) = E_P(Y \mid A = a_0, M_1 = m_1, M_2 = m_2, C = c)$, commonly referred to as the *outcome regression* as it may generally be estimated using mean regression of the outcome Y onto treatment, mediators, and confounders. The cumulative distribution of $(M_1(a_0), M_2(a_0))$ given $C = c$ is identified by $Q_{a_0,M_1,M_2}(m_1, m_2 \mid c) = \mathrm{pr}_P(M_1 \le m_1, M_2 \le m_2 \mid A = a_0, C = c)$. We assume the existence of a density q_{a_0,M_1,M_2} for the mediators with respect to a dominating measure and define marginal mediator densities $q_{a_0,M_i}(m_i \mid c) = \int_{m_j \in \mathcal{M}_j} dQ_{a_0,M_1,M_2}(m_1, m_2 \mid c)$ for $i, j = 1, 2$ and $i \neq j$. We subsequently refer to these objects as *marginal* mediator distributions, though they are in fact conditional on $A = a_0$ and C.

The identifying formula for each effect can now be written as a statistical functional of the observed data distribution by substituting the outcome regression for $\bar{\mathbb{Q}}_{a_0}(m_1, m_2, c)$ and the observed-data mediator distributions for the respective counterfactual distributions in the integral expressions above.

We note that the above assumptions preclude the existence of treatment-induced confounding of the mediator-outcome association. In the Web Supplement, we provide relevant extensions to this setting.

3 Methods

3.1 Efficiency theory

In this section, we develop efficiency theory for nonparametric estimation of interventional effects. This theory centers around the efficient influence function of each parameter. The efficient influence function is important for several reasons. First, it allows us to utilize of two existing estimation frameworks, one-step estimation [2, 10] and targeted minimum loss-based estimation [22, 23], to generate estimators that are nonparametric efficient. That is, under suitable regularity conditions, they achieve the smallest asymptotic variance amongst all regular estimators that, when scaled by $n^{1/2}$, have an asymptotic Normal distribution. We discuss how these estimators can be implemented in section 3.2. The second important feature of the efficient influence function is that its variance equals the variance of the limit distribution of the scaled estimators. Thus, an estimate of the variance of the efficient influence function is a natural standard error estimate, which affords closed-form Wald-style confidence intervals and hypothesis tests (Section 3.3). Finally, the efficient influence function also characterizes robustness properties of our proposed estimators (Section 3.4).

To introduce the efficient influence function, several additional definitions are required. For a given distribution $P' \in \mathcal{P}$, we define $g'_{a_0}(c) = \mathrm{pr}_{P'}(A = a_0 \mid C = c)$, commonly referred to as a *propensity score*. For $i, j = 1, 2$ and $i \neq j$, we introduce the following partially marginalized outcome regressions, $\tilde{Q}_{a,M_i^\star}(m_j, c) = \int \bar{Q}_a(m_1, m_2, c) dQ_{a,M_i}(m_i \mid c)$. We also introduce notation for the indicator function $\mathbb{1}_a : \{a, a^\star\} \to \{0, 1\}$ defined by $\mathbb{1}_a(\tilde{a}) = 1$ if $\tilde{a} = a$ and zero otherwise. $\mathbb{1}_{a^\star}$ is similarly defined.

Theorem 1. *Under sampling from $P' \in \mathcal{P}$, the efficient influence function evaluated on a given observation $\tilde{o}$ for the total effect is*

$$D^*(P')(\tilde{o}) = \frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})}\{\tilde{y} - \tilde{Q}'_{a,M_1,M_2}(\tilde{c})\} - \frac{\mathbb{1}_{a^\star}(\tilde{a})}{g'_{a^\star}(\tilde{c})}\{\tilde{y} - \tilde{Q}'_{a^\star,M_1^\star,M_2^\star}(\tilde{c})\}$$
$$+ \tilde{Q}'_{a,M_1,M_2}(\tilde{c}) - \tilde{Q}'_{a^\star,M_1^\star,M_2^\star}(\tilde{c}) - \psi' \ .$$

The efficient influence function for the interventional direct effect is

$$D^*_A(P')(\tilde{o}) = \frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})} \frac{q'_{a^\star,M_1,M_2}(\tilde{m}_1, \tilde{m}_2 \mid \tilde{c})}{q'_{a,M_1,M_2}(\tilde{m}_1, \tilde{m}_2 \mid \tilde{c})}\{\tilde{y} - \bar{Q}'_a(\tilde{m}_1, \tilde{m}_2, \tilde{c})\}$$

$$
\begin{aligned}
&- \frac{\mathbb{1}_{a^\star}(\tilde{a})}{g'_{a^\star}(\tilde{c})}\{\tilde{y} - \bar{Q}'_{a^\star}(\tilde{m}_1, \tilde{m}_2, \tilde{c})\} \\
&+ \frac{\mathbb{1}_{a^\star}(\tilde{a})}{g'_{a^\star}(\tilde{c})}\left[\bar{Q}'_a(\tilde{m}_1, \tilde{m}_2, \tilde{c}) - \bar{Q}'_{a^\star}(\tilde{m}_1, \tilde{m}_2, \tilde{c}) - \{\tilde{Q}'_{a,M_1^\star,M_2^\star}(\tilde{c}) - \tilde{Q}'_{a^\star,M_1^\star,M_2^\star}(\tilde{c})\}\right] \\
&+ \tilde{Q}'_{a,M_1^\star,M_2^\star}(\tilde{c}) - \tilde{Q}'_{a^\star,M_1^\star,M_2^\star}(\tilde{c}) - \psi'_A \ .
\end{aligned}
$$

The efficient influence function for the interventional indirect effect through M_1 is

$$
\begin{aligned}
D^*_{M_1}(P')(\tilde{o}) = &\frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})}\frac{\{q'_{a,M_1}(\tilde{m}_1 \mid \tilde{c}) - q'_{a^\star,M_1}(\tilde{m}_1 \mid \tilde{c})\}q'_{a^\star,M_2}(\tilde{m}_2 \mid \tilde{c})}{q'_{a,M_1,M_2}(\tilde{m}_1, \tilde{m}_2 \mid \tilde{c})}\{\tilde{y} - \bar{Q}'_a(\tilde{m}_1, \tilde{m}_2, \tilde{c})\} \\
&+ \frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})}\{\tilde{Q}'_{a,M_2^\star}(\tilde{m}_1, \tilde{c}) - \tilde{Q}'_{a,M_1 \times M_2^\star}(\tilde{c})\} \\
&- \frac{\mathbb{1}_{a^\star}(\tilde{a})}{g'_{a^\star}(\tilde{c})}\{\tilde{Q}'_{a,M_2^\star}(\tilde{m}_1, \tilde{c}) - \tilde{Q}'_{a,M_1^\star \times M_2^\star}(\tilde{c})\} \\
&+ \frac{\mathbb{1}_{a^\star}(\tilde{a})}{g'_{a^\star}(\tilde{c})}\left[\tilde{Q}'_{a,M_1}(\tilde{m}_2, \tilde{c}) - \tilde{Q}'_{a,M_1^\star}(\tilde{m}_2, \tilde{c}) - \{\tilde{Q}'_{a,M_1 \times M_2^\star}(\tilde{c}) - \tilde{Q}'_{a,M_1^\star \times M_2^\star}(\tilde{c})\}\right] \\
&+ \tilde{Q}'_{a,M_1 \times M_2^\star}(\tilde{c}) - \tilde{Q}'_{a,M_1^\star \times M_2^\star}(\tilde{c}) - \psi'_{M_1} \ .
\end{aligned}
$$

The efficient influence function for the interventional indirect effect through M_2 is

$$
\begin{aligned}
D^*_{M_2}(P')(\tilde{o}) = &\frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})}\frac{\{q'_{a,M_2}(\tilde{m}_2 \mid \tilde{c}) - q'_{a^\star,M_2}(\tilde{m}_2 \mid \tilde{c})\}q'_{a,M_1}(\tilde{m}_1 \mid \tilde{c})}{q'_{a,M_1,M_2}(\tilde{m}_1, \tilde{m}_2 \mid \tilde{c})}\{\tilde{y} - \bar{Q}'_a(\tilde{m}_1, \tilde{m}_2, \tilde{c})\} \\
&+ \frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})}\{\tilde{Q}'_{a,M_1}(\tilde{m}_2, \tilde{c}) - \tilde{Q}'_{a,M_1 \times M_2}(\tilde{c})\} \\
&- \frac{\mathbb{1}_{a^\star}(\tilde{a})}{g'_{a^\star}(\tilde{c})}\{\tilde{Q}'_{a,M_1}(\tilde{m}_2, \tilde{c}) - \tilde{Q}'_{a,M_1 \times M_2^\star}(\tilde{c})\} \\
&+ \frac{\mathbb{1}_a(\tilde{a})}{g'_a(\tilde{c})}\left[\tilde{Q}'_{a,M_2}(\tilde{m}_1, \tilde{c}) - \tilde{Q}'_{a,M_2^\star}(\tilde{m}_1, \tilde{c}) - \{\tilde{Q}'_{a,M_1 \times M_2}(\tilde{c}) - \tilde{Q}'_{a,M_1 \times M_2^\star}(\tilde{c})\}\right] \\
&+ \tilde{Q}'_{a,M_1 \times M_2}(\tilde{c}) - \tilde{Q}'_{a,M_1 \times M_2^\star}(\tilde{c}) - \psi'_{M_2} \ .
\end{aligned}
$$

The efficient influence function for the covariant interventional effect is $D^*_{M_1,M_2} = D^* - D^*_A - D^*_{M_1} - D^*_{M_2}$.

A proof of Theorem 1 is provided in the web supplement.

3.2 Estimators

We propose estimators of each interventional effect using one-step and targeted minimum loss-based estimation. Both techniques develop along a similar path. We first obtain estimates of the propensity score, outcome regression, and joint mediator distribution; we collectively refer to these quantities as *nuisance parameters*. With estimated nuisance parameters in hand, we subsequently apply a correction based on the efficient influence function to the nuisance estimates.

To estimate the propensity score, we can use any suitable technique for mean regression of the binary outcome A onto confounders C. Working logistic regression models are commonly used for this purpose, though semi- and nonparametric alternatives would be more in line with our choice of model. We denote by $g_{n,a_0}(c)$ the chosen estimate of $g_{a_0}(c)$. Similarly, the outcome regression can be estimated using mean regression of the outcome Y onto A, M_1, M_2, and C. For example, if the study outcome is binary, logistic regression could again be used, though more flexible regression estimators may be preferred. As above, we denote by $\bar{Q}_{n,a_0}$ the estimated outcome regression evaluated under $A = a_0$, with $\bar{Q}_{n,a_0}(m_1, m_2, c)$ providing an estimate of $E_P(Y \mid A = a_0, M_1 = m_1, M_2 = m_2, C = c)$. To estimate the marginal cumulative distribution of C, we will use the empirical cumulative distribution function, which we denote by $Q_{n,C}$.

Estimation of the conditional joint distribution of the mediators is a more challenging proposition, as fewer tools are available for flexible estimation of conditional multivariate distribution functions. We hence focus our developments on the development of approaches for discrete-valued mediators. The approach we adopt could be extended to continuous-valued mediators by considering a fine partitioning of the mediator values. We examine this approach via simulation in Section 4. To develop our density estimators, we use the approach of Dìaz Muñoz and van der Laan [9], which considers estimation of a conditional density via estimation of discrete conditional hazards. Briefly, consider estimation of the distribution of M_2 given A and C, and, for simplicity, suppose that the support of M_2 is $\{1, 2, 3\}$. We create a long-form data set, where the number of rows contributed by each individual contribute is equal to their observed value of M_2. An example is illustrated in Table 1. We see that the long-form data set includes an integer-valued column named "bin" that indicates to which value of M_2 each row corresponds, as well as a binary column $\mathbb{1}_{\text{bin}}(M_2)$ indicating whether the observed value of M_2 corresponds to each bin. These long-form data can be used to fit a regression of the binary outcome $\mathbb{1}_{\text{bin}}(M_2)$ onto C, A, and bin. This naturally estimates $\lambda_b(a_0, c) = P(M_2 = b \mid M_2 > b - 1, A = a_0, C = c)$, the conditional discrete hazard of M_2 given A and C. Let $\lambda_{n,\cdot}$ denote the estimated hazard obtained from fitting this regression. An estimate of the density at $m_2 \in \mathcal{M}_2$ is

$$q_{n,a_0,M_2}(m_2 \mid c) = \frac{\lambda_{n,m_2}(a_0,c) \prod\limits_{b=1}^{m_2-1} \{1 - \lambda_{n,b}(a_0,c)\}}{\sum\limits_{m \in \mathcal{M}_2} \left[\lambda_{n,m}(a_0,c) \prod\limits_{b=1}^{m-1} \{1 - \lambda_{n,b}(a_0,c)\} \right]} \ .$$

Similarly, an estimate $q_{n,a_0,M_1}(\cdot \mid m_2, c)$ of the conditional distribution of M_1 given $A = a_0, M_2 = m_2, C = c$ can be obtained. An estimate of the joint conditional density is implied by these estimates, $q_{n,a_0,M_1,M_2}(m_1, m_2 \mid c) = q_{n,a_0,M_1}(m_1 \mid m_2, c) q_{n,a_0,M_2}(m_2 \mid c)$, while an estimate of the marginal distribution of M_1 is $q_{n,a_0,M_1}(m_1, \mid c) = \sum_{m_2 \in \mathcal{M}_2} q_{n,a_0,M_1,M_2}(m_1, m_2 \mid c)$.

In principle, one could reverse the roles of M_1 and M_2 in the above procedure. That is, we could instead estimate the distribution of M_1 given $A = a_0, C$ and of M_2 given $A = a_0, C, M_1$. Cross-validation could be used to pick between the two potential estimators of the joint distribution. Other approaches to conditional density estimation are permitted by our procedure as well. For example, approaches based on working copula models may be particularly appealing in this context, as they allow separate specification of marginal vs. joint distributions of the mediators.

ID	C	A	M_2
1	**1**	**0**	**1**
2	**0**	**1**	**3**
⋮	⋮	⋮	⋮
n	**0**	**1**	**2**

ID	C	A	**bin**	$\mathbb{1}_{\textbf{bin}}(M_2)$
1	**1**	**0**	**1**	**1**
2	**0**	**1**	**1**	**0**
2	**0**	**1**	**2**	**0**
2	**0**	**1**	**3**	**1**
⋮	⋮	⋮	⋮	⋮
n	**0**	**1**	**1**	**0**
n	**0**	**1**	**2**	**1**

Table 1. An illustration of how to make a long form data set suitable for estimating mediator distributions. An ID is uniquely assigned to each independent data unit and a single confounder C is included in the mock data set.

Given estimates of nuisance parameters, we now illustrate one-step estimation for the interventional direct effect. One-step estimators of other effects can be generated similarly. A plug-in estimate of the conditional interventional direct effect given $C = c$ is the difference between

$$\begin{aligned} \tilde{Q}_{n,a,M_1^\star,M_2^\star}(c) &= \int\limits_{\mathcal{M}_1 \times \mathcal{M}_2} \bar{Q}_{n,a}(m_1, m_2, c) dQ_{n,a^\star,M_1,M_2}(m_1, m_2 \mid c) \text{ and} \\ \tilde{Q}_{n,a^\star,M_1^\star,M_2^\star}(c) &= \int\limits_{\mathcal{M}_1 \times \mathcal{M}_2} \bar{Q}_{n,a^\star}(m_1, m_2, c) dQ_{n,a^\star,M_1,M_2}(m_1, m_2 \mid c) \ . \end{aligned} \tag{1}$$

To obtain a plug-in estimate $\psi_{n,A}$ of ψ_A, we standardize the conditional effect estimate with respect to $Q_{n,C}$, the empirical distribution of C. Thus, the plug-in estimator of ψ_A is $\psi_{n,A} = \int_{\mathcal{C}} \{\tilde{Q}_{n,a,M_1^\star,M_2^\star}(c) - \tilde{Q}_{n,a^\star,M_1^\star,M_2^\star}(c)\} dQ_{n,C}(c)$.

The one-step estimator is constructed by adding an efficient influence function-based correction to an initial plug-in estimate. Suppose we are given estimates of all relevant nuisance quantities and let P_n' denote any probability distribution in $\mathcal{P}$ that is compatible with these estimates. The efficient influence function for ψ_A under sampling from P_n' is $D_A^*(P_n')$, and the one-step estimator is $\psi_{n,A,+} = \psi_{n,A} + n^{-1} \sum_{i=1}^n D_A^*(P_n')(O_i)$. All other effect estimates are generated in this vein: estimated nuisance parameters are plugged in to the efficient influence function, the resultant function is evaluated on each observation, and the empirical average of this quantity is added to the plug-in estimator.

While one-step estimators are appealing in their simplicity, the estimators may not obey bounds on the parameter space in finite samples. For example, if the study outcome is binary, then the interventional effects each represent a difference in two probabilities and thus are bounded between -1 and 1. However, one-step estimators may fall outside of this range. This motivates estimation of these quantities using targeted minimum loss-based estimation, a framework for generating plug-in estimators. The implementation of such estimators is generally more involved than that of one-step estimators. In this approach, a second-stage model fitting is used to ensure that nuisance parameter estimates satisfy efficient influence function estimating equations. The approach for this second-stage fitting is dependent on the specific effect parameter considered and the procedure differs subtly for the various effect measures presented here. The web supplement includes a detailed exposition of how such estimators can be implemented.

3.3 Large sample inference

We now present a theorem establishing the joint weak convergence of the proposed estimators to a random variable with a multivariate normal distribution. Because the asymptotic behavior of the one-step and targeted minimum loss estimators are equivalent, we present a single theorem. A discussion of the differences in regularity conditions required to prove the theorem for one-step versus targeted minimum loss estimation is provided in the web supplement. Let $\psi_{n,\cdot}$ denote the vector of (one-step or targeted minimum loss) estimates of $\psi_{\cdot} = (\psi_A, \psi_{M_1}, \psi_{M_2}, \psi_{M_1,M_2})^\top$ and let $D_{\cdot}^*(P')$ denote the vector of efficient influence functions defined by

$$\tilde{o} \mapsto \left(D_A^*(P')(\tilde{o}), D_{M_1}^*(P')(\tilde{o}), D_{M_2}^*(P')(\tilde{o}), D_{M_1,M_2}^*(P')(\tilde{o})\right)^\top .$$

In the theorem, we use $|| \cdot ||$ to denote the $L_2(P)$-norm, define for any P-measurable f as $||f||^2 = \int f(o) dP(o)$.

Theorem 2. *Under sampling from $P \in \mathcal{P}$, if for $a_0 = a, a^\star$,*

(i) $sup_c |g_{n,a_0}(c) - g_{a_0}(c)| \rightarrow 0$ *in probability as* $n \rightarrow \infty$*,*
(ii) $sup_{m_1,m_2,c} |q_{n,a_0,M_1,M_2}(m_1, m_2 \mid c) - q_{a_0,M_1,M_2}(m_1, m_2 \mid c)| \rightarrow 0$ *in probability as* $n \rightarrow \infty$*,*
(iii) $||g_{n,a_0} - g_{n,a_0}|| = o_p(n^{-1/4})$*,*
(iv) $||\bar{Q}_{n,a_0} - \bar{Q}_{a_0}|| = o_p(n^{-1/4})$*,*
(v) $||q_{n,a_0,M_1,M_2} - q_{a_0,M_1,M_2}|| = o_p(n^{-1/4})$*,*
(vi) $||q_{n,a_0,M_1} q_{n,a_0,M_2} - q_{a_0,M_1} q_{a_0,M_2}|| = o_p(n^{-1/4})$*,*
(vii) $||q_{n,a_0,M_1} - q_{a_0,M_1}|| = o_p(n^{-1/4})$*,* $||q_{n,a_0,M_2} - q_{a_0,M_2}|| = o_p(n^{-1/4})$*, and*
(viii) $\int \{D_{\cdot}^*(P_n')(o) - D_{\cdot}^*(P)(o)\}^2 dP(o) \rightarrow 0$ *in probability as* $n \rightarrow \infty$ *and* $D_{\cdot}^*(P_n')$ *falls in a P-Donsker class with probability tending to 1,*

then $n^{1/2}(\psi_{n,\cdot} - \psi_{\cdot}) \rightarrow_d$ *Normal*$(0, \Sigma)$*, where* $\Sigma = \int D_{\cdot}^*(P)(o) D_{\cdot}^*(P)(o)^\top dP(o)$*.*

The regularity conditions required for Theorem 2 are typical of many problems in semiparametric efficiency theory. We provide conditions in terms of $L_2(P)$-norm convergence, as this is typical of this literature; however, alternative and potentially weaker conditions are possible to derive. For further discussion, see the supplementary material. As with any nonparametric procedure, there is a concern relating to the dimensionality C, particularly in situations with real-valued mediators. Minimum loss estimators (MLE) in certain function classes can attain the requisite convergence rates. For example, an MLE in the class of functions that are right-continuous with left limits (i.e., càdlàg) with variation norm bounded by a constant achieves an $L_2(P)$ convergence rate faster than $n^{-1/4}$ irrespective of the dimension of the conditioning set [1]. However, this may not allay all concerns pertaining to the curse of dimensionality due to the fact that in moderately high dimensions, these function classes can be restrictive and thus the true function may fall outside this class. Nevertheless, we suggest (and our simulations show) that in spite of concerns pertaining to the curse of dimensionality our procedure will enjoy reasonable finite-sample performance in many settings.

The covariance matrix Σ may be estimated by the empirical covariance matrix of the vector $D^*(P'_n)$ applied to the observed data, where P'_n is any distribution in the model that is compatible with the estimated nuisance parameters. With the estimated covariance matrix, it is straightforward to construct Wald confidence intervals and hypothesis tests about the individual interventional effects or comparisons between them. For example, a straightforward application of the delta method would allow for a test of the null hypothesis that $\psi_{M_1} = \psi_{M_2}$.

3.4 Robustness properties

As with many problems in causal inference, consistent estimation of interventional effects requires consistent estimation only of *certain combinations* of nuisance parameters. To determine these combinations, we may study the stochastic properties of the efficient influence function. In particular, consider a parameter whose value under P is $\tilde{\psi}$ and whose efficient influence function under sample from P' can be written $\tilde{D}^*(P', \tilde{\psi}')$, where $\tilde{\psi}'$ is the value of the parameter of interest under P'. Then we may study the circumstances under which $\int \tilde{D}^*(P', \tilde{\psi}) dP(o) = 0$. This generally entails understanding which parameters of P' must align with those parameters of P to ensure that the influence function $\tilde{D}^*(P', \tilde{\psi})$ has mean zero under sampling from P. We present the results of this analysis in a theorem below and refer readers to the web supplement for the proof.

Theorem 3. *Locally efficient estimators of the total effect and the intervention direct, indirect, and covariant effects are consistent for their respective target parameters if the following combinations of nuisance parameters are consistently estimated:*

Total effect: $(\bar{Q}_a, \bar{Q}_{a^\star}, Q_{a,M_1,M_2}, Q_{a^\star,M_1,M_2})$ *or* $(g_a, g_{a^\star})$

Interventional direct effect: $(\bar{Q}_a, \bar{Q}_{a^\star}, g_{a^\star})$ *or* $(\bar{Q}_a, \bar{Q}_{a^\star}, Q_{a,M_1,M_2}, Q_{a^\star,M_1,M_2})$ *or* $(Q_{a,M_1,M_2}, Q_{a^\star,M_1,M_2}, g_{a^\star}, g_a)$*;*

Inverventional indirect effect through M_1*:* $(\bar{Q}_a, Q_{a,M_1}, Q_{a^\star,M_1}, Q_{a^\star,M_2})$ *or* $(g_a, Q_{a,M_1,M_2}, Q_{a^\star,M_1}, Q_{a^\star,M_2})$ *or* $(\bar{Q}_a, g_a, g_{a^\star}, Q_{a^\star,M_2})$ *or* $(\bar{Q}_a, g_a, g_{a^\star}, Q_{a,M_1})$*;*

Inverventional indirect effect through M_2*:* $(\bar{Q}_a, Q_{a,M_2}, Q_{a^\star,M_2}, Q_{a,M_1})$ *or* $(g_a, Q_{a,M_1,M_2}, Q_{a^\star,M_2})$ *or* $(\bar{Q}_a, g_a, g_{a^\star}, Q_{a,M_1})$ *or* $(\bar{Q}_a, g_a, g_{a^\star}, Q_{a,M_2})$*;*

Interventional covariant effect: $(\bar{Q}_a, \bar{Q}_{a^\star}, Q_{a,M_1,M_2}, Q_{a^\star,M_1,M_2})$ *or* $(g_a, g_a^\star, \bar{Q}_a, \bar{Q}_{a^\star}, Q_{a,M_1}, Q_{a^\star,M_2})$ *or* $(g_a, g_{a^\star}, Q_{a,M_1,M_2}, Q_{a^\star,M_1}, Q_{a^\star,M_2})$.

The most interesting robustness result is perhaps that pertaining to the indirect effects. The first condition for consistent estimation is expected, as the propensity score plays no role in the definition of the indirect effect. The second condition shows that the joint mediator distribution and propensity score together can compensate for inconsistent estimation of the outcome regression, while the relevant marginal mediator distributions are required to properly marginalize the resultant quantity. The third and fourth conditions show that inconsistent estimation of the marginal distribution one, but not both, of the mediators can be corrected for via the propensity score.

We note that Theorem 3 provides sufficient, but not necessary, conditions for consistent estimation of each effect. For example, a consistent estimate of the total effect is implied by a consistent estimate of $\tilde{Q}_{a,M_1,M_2}$ and $\tilde{Q}_{a^\star,M_1^\star,M_2^\star}$, a condition that is generally weaker than requiring consistent estimation of the outcome regression and joint mediator distribution. Because our estimation strategy relies on estimation of the joint mediator distribution, we have described robustness properties in terms of the large sample behavior of estimators of those quantities.

3.5 Extensions

In the Web Supplement, we provide relevant extensions to the setting where the mediator-outcome relationship is confounded by measured covariates whose distributions are affected by the treatment. In this case, both the effects of interest and their efficient influence functions involve the conditional distribution of the confounding covariates. We discuss the relevant modifications to the estimation procedures to accommodate this setting in the supplement.

Generalization to other effect scales requires only minor modifications. First, we determine the portions of the efficient influence function that pertain to each component of the additive effect. For example, considering ψ_{M_1}, we identify the portions of the efficient influence function that pertain to the mean counterfactual under draws of M_1 from $q_{a,M_1}(\cdot \mid C)$ and of M_2 from $q_{a^\star,M_2}(\cdot \mid C)$ versus those portions that pertain to the mean counterfactual under draws of M_1 from $q_{a^\star,M_1}(\cdot \mid C)$ and of M_2 from $q_{a^\star,M_2}(\cdot \mid C)$. We then develop a one-step or targeted minimum loss estimator for each of these components separately. Finally, we use the delta method to derive the resulting influence function. In the web supplement, we illustrate an extension to a multiplicative scale.

Our results can also be extended to estimation of interventional effects for more than two mediators. As discussed in Vansteelandt and Daniel [26], when there are more than two mediators, say $M_1, \ldots, M_t$, there are many possible path-specific effects. However, our scientific interest is usually restricted to learning effects that are mediated through each of the mediators, rather than all possible path-specific effects. Moreover, strong untestable assumptions are required to infer all path-specific effects, including assumptions about the direction of the causal effects between mediators. Therefore, it may be of greatest interest to evaluate direct effects such as

$$\psi_{t,A} = \int_{\mathcal{C}} \int_{\mathcal{M}_1\times\cdots\times\mathcal{M}_t} \{\bar{\mathbb{Q}}_a(m_1,\ldots,m_t,c) - \bar{\mathbb{Q}}_{a^\star}(m_1,\ldots,m_t,c)\} \\ \times d\mathbb{Q}_{M_1(a^\star),\ldots,M_t(a^\star)}(m_1,\ldots,m_t \mid c)dQ_C(c) \ ,$$

which describes the effect of setting $A = a$ versus $A = a^\star$, while drawing all mediators from the joint conditional distribution given $A = a^\star, C$, and for $s = 1, \ldots, t$, indirect effects such as

$$\psi_{t,M_s} = \int_{\mathcal{C}} \Bigg[\int_{\mathcal{M}_1\times\cdots\times\mathcal{M}_t} \bar{\mathbb{Q}}_a(m_1,m_2,c)\{d\mathbb{Q}_{M_s(a)}(m_s \mid c) - d\mathbb{Q}_{M_s(a^\star)}(m_s \mid c)\} \\ \times \prod_{u=1}^{s-1} d\mathbb{Q}_{M_u(a)}(m_u \mid c) \prod_{v=s+1}^{t} d\mathbb{Q}_{M_v(a^\star)}(m_v \mid c) \Bigg] dQ_C(c) \ ,$$

which describes the effect of setting M_s to the value it would assume under $A = a$ versus $A = a^\star$ while drawing $M_1, \ldots, M_{s-1}$ from their respective marginal distributions given $A = a, C$ and drawing $M_{s+1}, \ldots, M_t$ from their marginal distribution given $A = a^\star, C$. We provide relevant efficiency theory for these parameters in the web supplement.

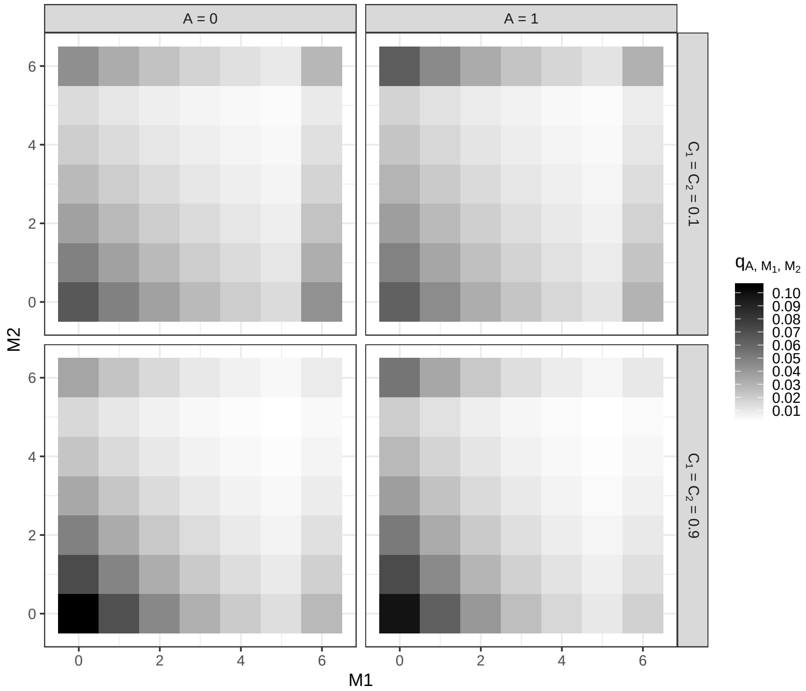

Figure 1. Joint distribution of mediators used in the simulation.

3.6

4 Simulations

4.1 Discrete mediators

We evaluated the small sample performance of our estimators via Monte Carlo simulation. Data were generated as follows. We simulated $C = (C_1, \ldots, C_5)$ by drawing C_1, C_2, C_3 independently from Uniform(0,1) distributions, and C_4, C_5 independently from Bernoulli distributions with success probability of 0.25 and 0.5, respectively. The treatment variable A was, given $C = c$, was drawn from a Bernoulli distribution with $g_a(c) = \text{logit}^{-1}(-1 + 0.125c_1 + 0.25c_2)$ and $g_{a^\star}(c) = 1 - g_a(c)$. Here, we consider $a = 1$ and $a^\star = 0$. Given $C = c, A = a_0$, the first mediator M_1 was generated by taking draws from a geometric distribution with success probability $\text{logit}^{-1}(-1.1 + 0.45c_1 + 0.125a_0)$. Any draw of six or greater was set equal to six. The second mediator was generated from a similarly truncated geometric distribution with success probability $\text{logit}^{-1}(-1.1 + 0.15c_1 + 0.2c_2 - 0.2a_0)$. Given $C = c, A = a_0, M_1 = m_1, M_2 = m_2$, the outcome Y was drawn from a Bernoulli distribution with success probability $\text{logit}^{-1}(-1 + c_1 - c_2 + 0.25m_1 + 0.25m_2 + 0.25a_0)$. The mediator distribution is visualized for combinations of c and a_0 in Figure 1. The true total effect is approximately 0.06, which decomposes into a direct effect of 0.05, an indirect effect through M_1 of -0.01, an indirect effect through M_2 of 0.02 and a covariant effect of 0.

The nuisance parameters were estimated using regression stacking [3, 27], also known as super learning [21] using the `SuperLearner` package for the `R` language [16]. We used this package to generate an ensemble of a main-terms logistic regression (as implemented in the `SL.glm` function in `SuperLearner`), polynomial multivariate adaptive regression splines (`SL.earth`), and a random forest (`SL.ranger`). The ensemble was built by selecting the convex combination of these three estimators that minimized ten-fold cross-validated deviance.

We evaluated our proposed estimators under this data generating process at sample sizes of 250, 500, 1000, and 2000. At each sample size, we simulated 1,000 data sets. Point estimates were compared in terms of their Monte Carlo bias, standard deviation, mean squared error. We evaluated weak convergence by visualizing the sampling distribution of the estimators after centering at the true parameter value and scaling by an oracle standard error, computed as the Monte Carlo standard deviation of the estimates, as well as scaling by an estimated standard error based on the estimated variance of the efficient influence function. Similarly, we evaluated the coverage probability of a nominal 95% Wald-style confidence interval based on the oracle and estimated standard errors.

In terms of estimation, one-step and targeted minimum loss estimators behave as expected in large samples (Figure 2). The estimators are approximately unbiased in large samples and have mean squared error appropriately decreasing with sample size. Comparing the two estimation strategies, we see that one-step and targeted minimum loss estimators had comparable performance for the interventional direct effect, while the targeted minimum loss esitmator had better performance for the indirect effects. However, the one-step was uniformly better for estimating the covariant effect owing to large variability of the targeted minimum loss estimator of this quantity. Further examination of the results revealed that the second-stage model fitting required by the targeted minimum loss approach was could be unstable in small samples, leading to extreme results in several data sets.

The sampling distribution of the centered and scaled estimators were approximately a standard normal distribution (Figures 3 and 4), excepting the targeted minimum loss estimator scaled by an estimated standard error. Confidence intervals based on an oracle standard error came close to nominal coverage in all sample sizes, while those based on an estimated standard error tended to have under-coverage in small samples.

4.2 Continuous mediators

We examined the impact of discretization of the mediator distributions when in fact the mediators are continuous valued. To that end, we simulated data as follows. Covariates were simulated as above. The treatment variable A given $C = c$ was drawn from a Bernoulli distribution with $g_a(c) = \text{logit}^{-1}(-1 + 0.25c_1 - 0.5c_1c_2)$. Given $C = c, A = a_0$, M_1 and M_2 were respectively drawn from Normal distributions with unit variance and means $c_1 - 0.5a_0c_1$ and $c_2 - c_2a_0$. As above, Super Learner was used to estimate all nuisance parameters. To accommodate appropriate modeling of the interactions, we replaced the main terms GLM (`SL.glm`) with a forward stepwise GLM algorithm that included all two way interactions (`SL.step.interaction`). The true effect sizes were approximately the same as in the first simulation. We evaluated discretization of each continuous mediator distribution into 5 and 10 evenly spaced bins. For the sake of space, we focus results on the one-step estimator; results for targeted minimum loss estimator are included in the supplement.

Overall, discretization of the continuous mediator distribution had a greater impact on the performance of indirect effect estimators compared to direct effects (Figures 5 and 6). For the latter effects, oracle confidence intervals for both levels of discretization achieved nominal coverage for all sample sizes considered. For the indirect effects, we found that there was non-negligible bias in the estimates due to the discretization. The impacts in terms of confidence interval coverage were minimal in small sample sizes, but lead to under-coverage in larger sample sizes. Including more bins generally lead to better performance, but these estimates still exhibited bias in the largest sample sizes that impacted coverage. Nevertheless the performance of the indirect effect estimators was reasonable with oracle coverage $> 90\%$ for all sample sizes.

4.3 Additional simulations

In the web supplement we include several additional simulations studying the impact of the number of levels of the discrete mediator, as well as the impact of inconsistent estimation of the various nuisance parameters. For the former, we found that the results of the simulation were robust to number of mediator

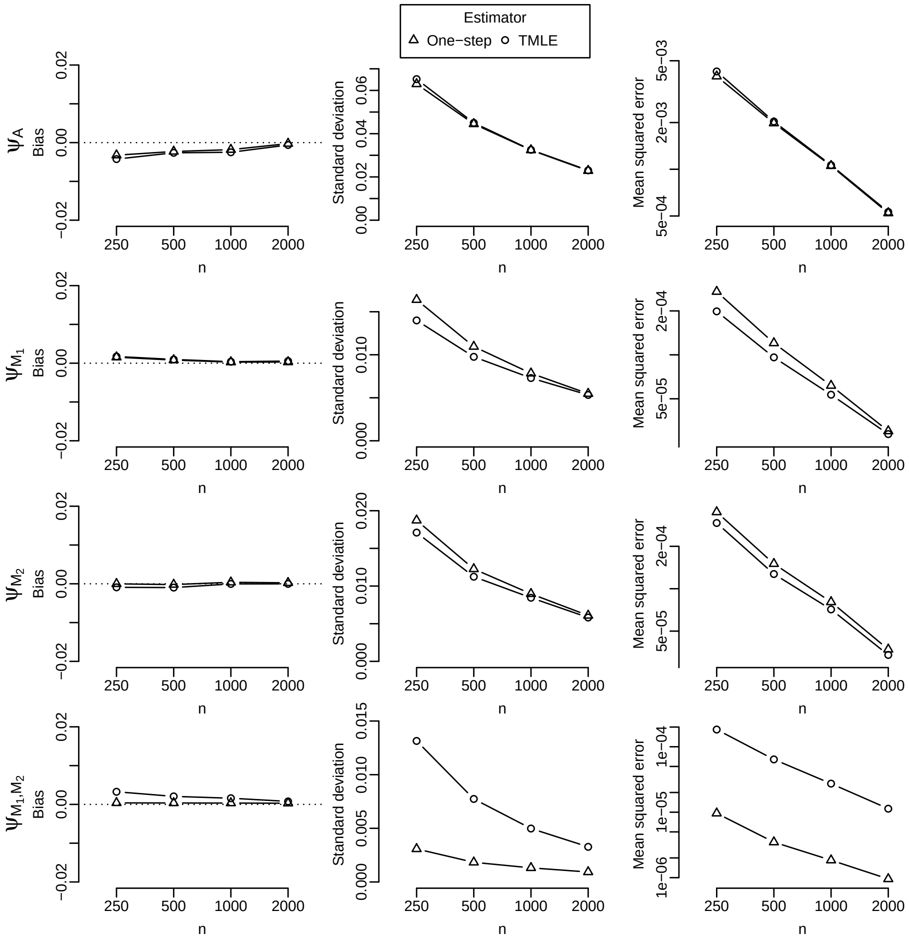

Figure 2. Comparison of one-step and targeted minimum loss estimators (TMLE) in terms of their Monte Carlo-estimated bias, standard deviation, and mean squared-error for the interventional direct (ψ_A), indirect (ψ_{M_1}, ψ_{M_2}), and covariant (ψ_{M_1,M_2}) effects.

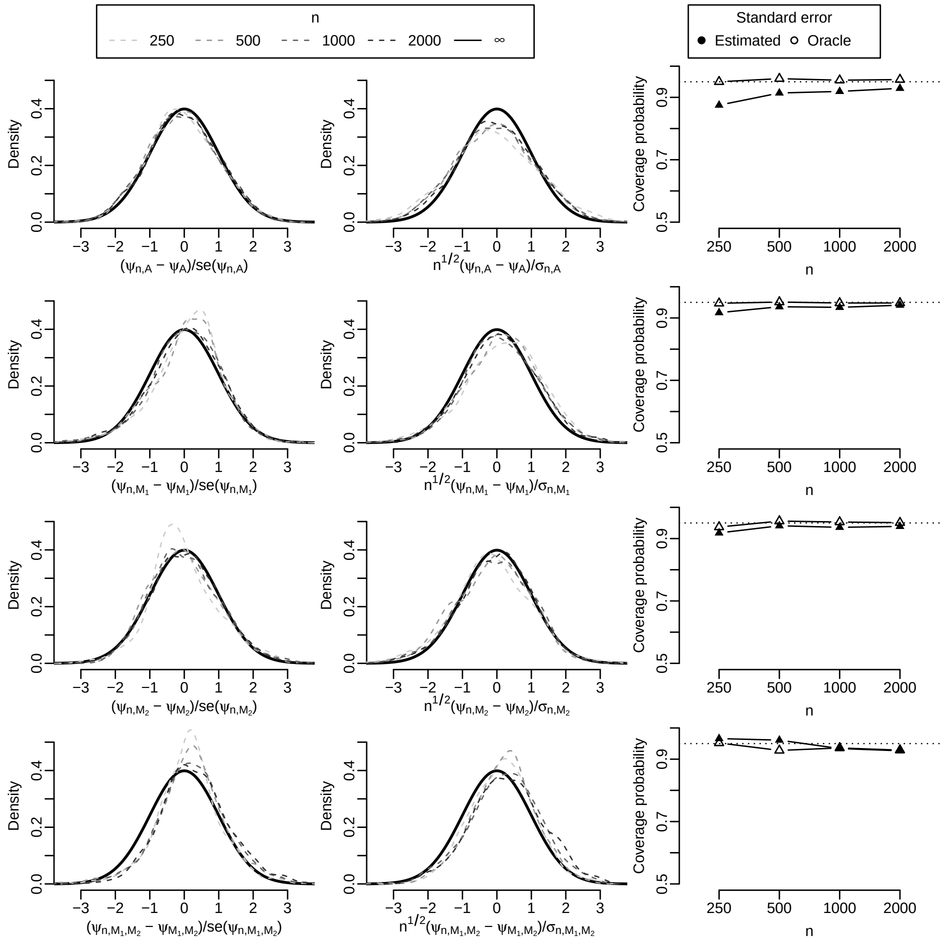

Figure 3. Illustration of weak convergence and Wald-style confidence intervals based on the one-step estimator. The left two columns show the kernel density estimate of the sampling distribution of the centered estimates of interventional effects scaled by the oracle standard error (left) and by their estimated standard error (middle). In each case, the asymptotic distribution is shown in black. The right panel shows coverage probability of a nominal 95% Wald-style confidence interval based on an oracle standard error (solid triangle) and an estimated standard error (open triangle).

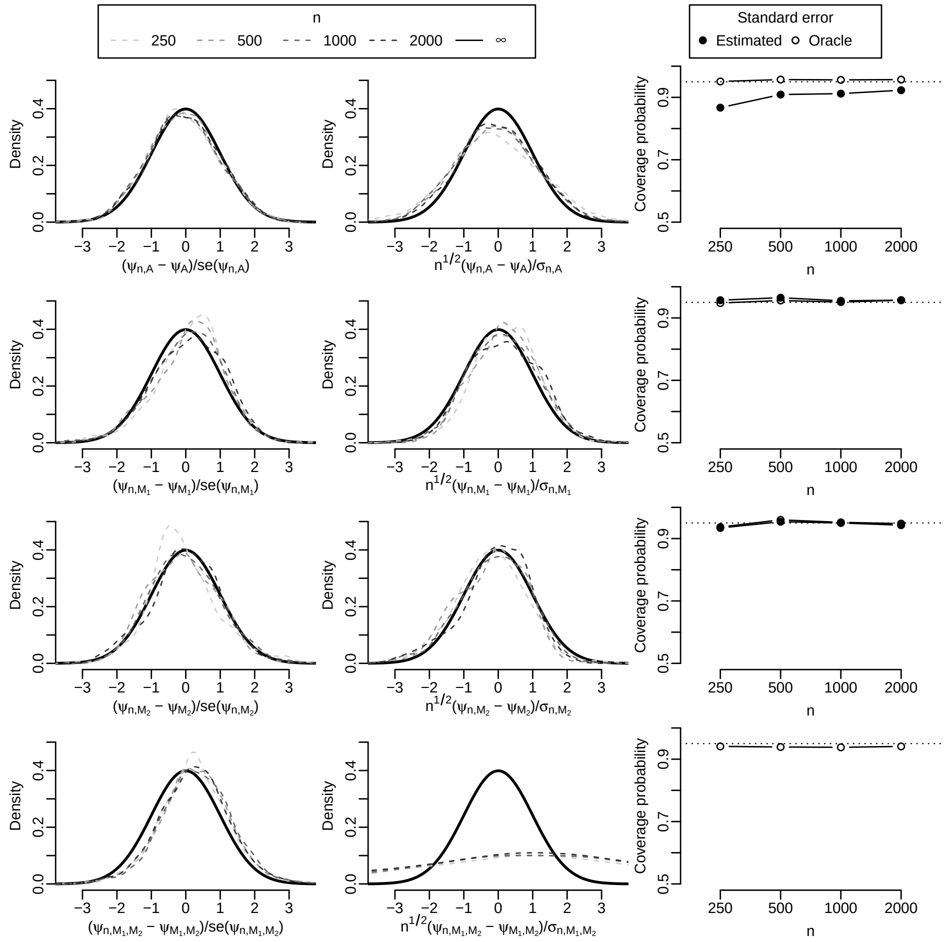

Figure 4. Illustration of weak convergence and Wald-style confidence intervals based on the targeted minimum loss estimator. The left two columns show the kernel density estimate of the sampling distribution of the centered estimates of interventional effects scaled by the oracle standard error (left) and by their estimated standard error (middle). In each case, the asymptotic distribution is shown in black. The right panel shows coverage probability of a nominal 95% Wald-style confidence interval based on an oracle standard error (solid circle) and an estimated standard error (open circle).

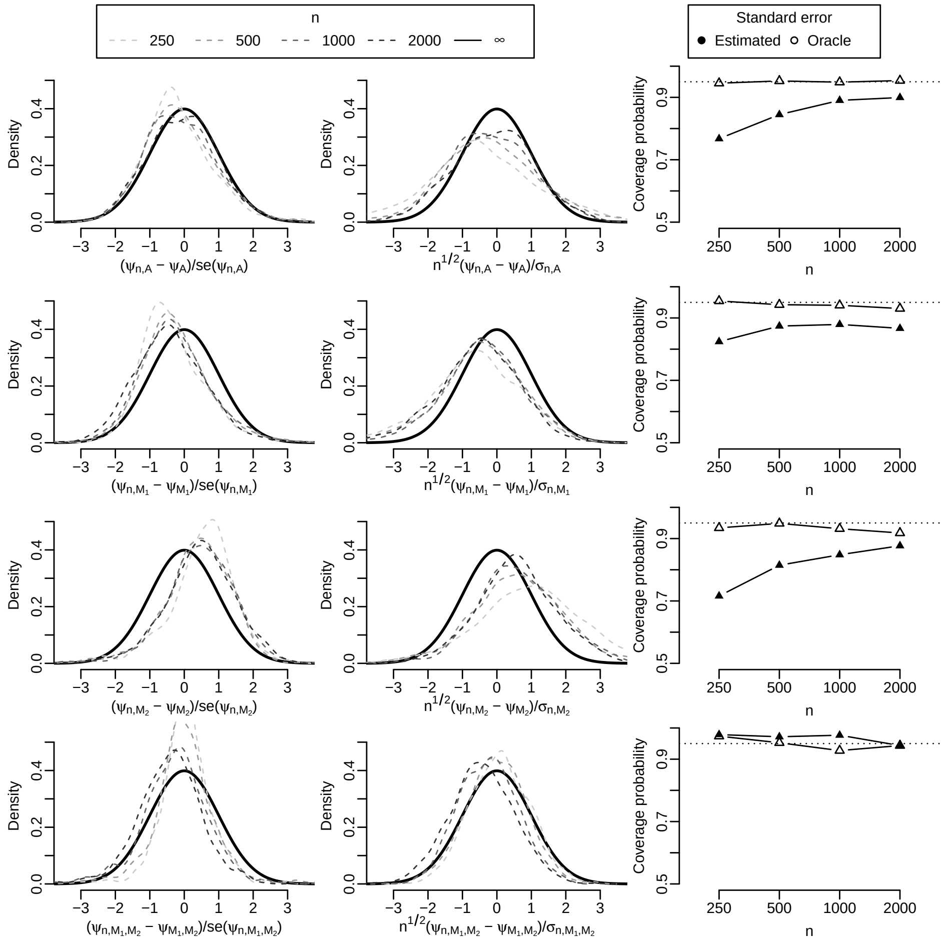

Figure 5. Illustration of weak convergence and Wald-style confidence intervals based on the one-step estimator that discretized mediator distributions into five bins. The left two columns show the kernel density estimate of the sampling distribution of the centered estimates of interventional effects scaled by the oracle standard error (left) and by their estimated standard error (middle). In each case, the asymptotic distribution is shown in black. The right panel shows coverage probability of a nominal 95% Wald-style confidence interval based on an oracle standard error (solid triangle) and an estimated standard error (open triangle).

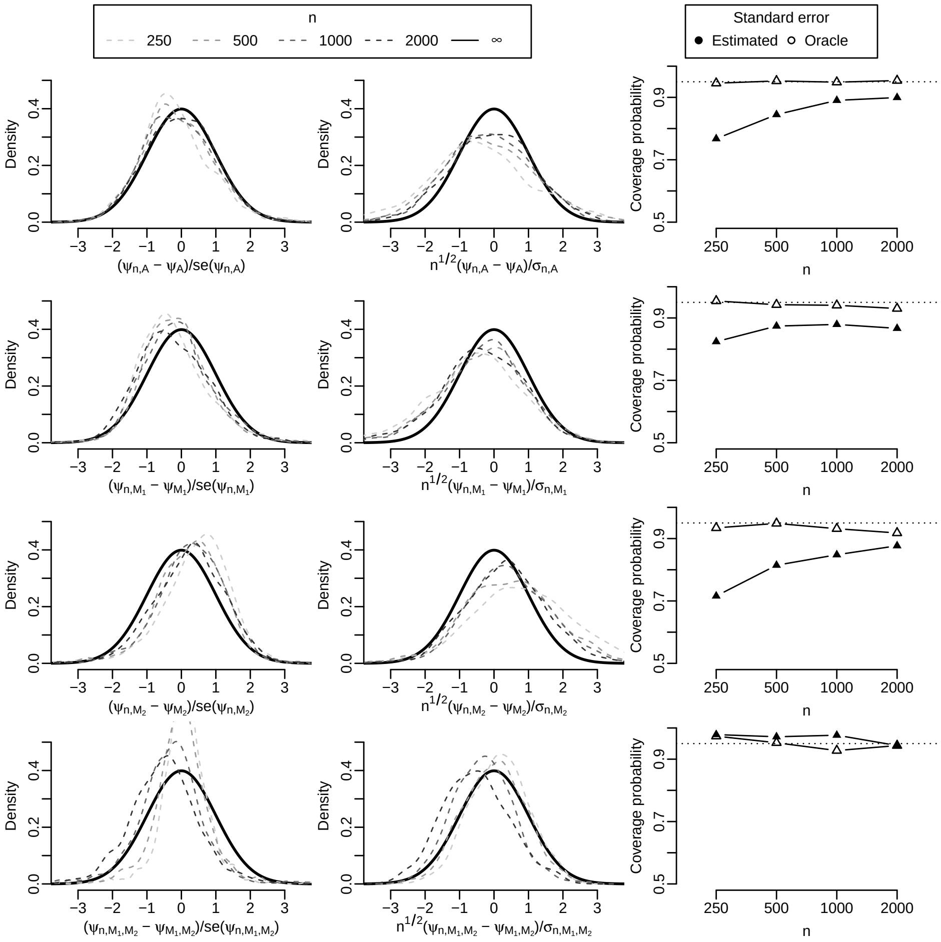

Figure 6. Illustration of weak convergence and Wald-style confidence intervals based on the one-step estimator that discretized mediator distributions into ten bins. The left two columns show the kernel density estimate of the sampling distribution of the centered estimates of interventional effects scaled by the oracle standard error (left) and by their estimated standard error (middle). In each case, the asymptotic distribution is shown in black. The right panel shows coverage probability of a nominal 95% Wald-style confidence interval based on an oracle standard error (solid triangle) and an estimated standard error (open triangle).

levels in the setting considered. For the latter, we confirmed the multiple robustness properties of the indirect effect estimators by studying the bias and standard deviation of the estimators in large sample sizes under the various patterns of misspecification given in our theorem.

5 Discussion

Our simulations demonstrate adequate performance of the proposed nonparametric estimators of interventional mediation effects in settings with relatively low-dimensional covariates (five, in our simulation). In certain settings, it may only be necessary to adjust for a limited number of covariates to adequately control confounding. For example, in the study of the mediating mechanisms of preventive vaccines using data from randomized trials, we need only adjust for confounders of the mediator/outcome relationship, since other forms of confounding are addressed by the randomized design. Generally, there are few known factors that are likely to impact vaccine-induced immune responses and so nonparametric analyses may be quite feasible in this case. For example, Cowling et al. [5] studied mediating effects of influenza vaccines, adjusting only for age. Thus, we suggest that interventional mediation estimands and nonparametric estimators thereof may be of interest for studying mediating pathways of vaccines. However, in other scenarios, it may be necessary to adjust for a high-dimensional set of confounders. For example, in observational studies of treatments (e.g., through an electronic health records system), we may require control for a high-dimensional set of putative confounders of treatment and outcome. This may raise concerns related to the curse-of-dimensionality when utilizing nonparametric estimators. Studying tradeoffs between the selection of various estimation strategies in this context will be an important area for future research.

We have developed an R package `intermed` with implementations of the proposed methods that is included in the web supplementary material. The package focuses on implementations for discrete mediators. However, our simulations demonstrate a clear need to extend the software to accommodate adaptive selection of the number of bins in the mediator density estimation procedure for continuous mediators. In small sample sizes, we found that course binning leads to adequate results, but as sample size increased, unsurprisingly there was a need for finer partitioning to reduce bias. In future versions of the software, we will include such adaptive binning strategies, as well as other methods for estimating continuous mediator densities.

The behavior of the targeted minimum loss estimator of the covariant effect in the simulation is surprising as generally we see comparable or better performance of such estimators relative to one-step estimators. This can likely be attributed to the fact that the targeted minimum loss procedure does not yield a compatible plug-in estimator of the vector $\psi_\cdot$, in the sense that there is likely no distribution P_n' that is compatible with all of the various nuisance estimators after the second-stage model fitting. A more parsimonious approach could consider either an iterative targeting procedure or a uniformly least favorable submodel that simultaneously targets the joint mediator density and outcome regression. The former is implemented in a concurrent proposal [8], where one-step and targeted minimum loss estimators of interventional effects are developed for a single mediator when the mediator-outcome relationship is subject to treatment-induced confounding. In their set up, if one can treat the treatment-induced confounder as a second mediator, then their proposal results in an estimate of one component of our indirect effect. In their simulations, they find superior finite-sample performance of the targeted minimum loss estimator relative to the one step, suggesting that targeting the mediator densities may be a more robust approach. However, their simulation involved only binary-valued mediators, so further comparison of these approaches is warranted in settings similar to our simulation, where mediators can take many values. We leave these developments to future work.

The Donsker class assumptions of our theorem could be removed by considering cross-validated nuisance parameter estimates (also known as cross-fitting) [4, 29]. This technique is implemented in our R package, but we leave to future research the examination of its impact of estimation and inference. We hypothesize that this approach will generally improve the anti-conservative confidence intervals in small samples, but will have little impact on performance of point estimates in terms of bias and variance.

6 Author's statements

D Benkeser was funded by National Institutes of Health award R01AHL137808 and National Science Foundation award 2015540. Code to reproduce simulation results is available at https://github.com/benkeser/intermed/tree/master/simulations.

References

[1] Benkeser, D. and van der Laan, M. J. (2016). The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE.

[2] Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1997). *Efficient and adaptive estimation for semiparametric models*. Springer, Berlin Heidelberg New York.

[3] Breiman, L. (1996). Stacked regressions. *Mach Learn*, 24:49–64.

[4] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1).

[5] Cowling, B. J., Lim, W. W., Perera, R. A., Fang, V. J., Leung, G. M., Peiris, J. M., and Tchetgen Tchetgen, E. J. (2019). Influenza hemagglutination-inhibition antibody titer as a mediator of vaccine-induced protection for influenza B. *Clinical Infectious Diseases*, 68(10):1713–1717.

[6] Coyle, J. and van der Laan, M. J. (2018). Targeted bootstrap. In MJ, v. and S, R., editors, *Targeted Learning for Data Science*, chapter 28, pages 523–539. Springer International Publishing.

[7] Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

[8] Dìaz, I., Hejazi, N. S., Rudolph, K. E., and van der Laan, M. J. (2020). Nonparametric efficient causal mediation with intermediate confounders.

[9] Dìaz Muñoz, I. and van der Laan, M. J. (2011). Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics*, 7(1):1–20.

[10] Ibragimov, I. and Khasminskii, R. (1981). *Statistical estimation*. Springer.

[11] Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309.

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

[13] Naimi, A. I., Schnitzer, M. E., Moodie, E. E., and Bodnar, L. M. (2016). Mediation analysis for health disparities research. *American Journal of Epidemiology*, 184(4):315–324.

[14] Pearl, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the 17th Conference*, San Francisco. Morgan Kaufmann.

[15] Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459.

[16] Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. J. (2013). *SuperLearner: Super Learner Prediction*. R package version 2.0-28.

[17] Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158.

[18] Rudolph, K. E., Sofrygin, O., Zheng, W., and van der Laan, M. J. (2017). Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidemiologic Methods*, 7(1).

[19] Tchetgen Tchetgen, E. J. and Phiri, K. (2014). Bounds for pure direct effect. *Epidemiology (Cambridge, Mass.)*, 25(5):775.

[20] Valeri, L. and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18(2):137.

[21] van der Laan, M., Polley, E., and Hubbard, A. (2007). Super learner. *Stat Appl Genet Mol*, 6(1):Article 25.

[22] van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin Heidelberg New York.

[23] van der Laan, M. and Rubin, D. B. (2006). Targeted maximum likelihood learning. *Int J Biostat*, 2(1):Article 11.

[24] VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):917–938.

[25] VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300.

[26] Vansteelandt, S. and Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258.

[27] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

[28] Yuan, Y. and MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, 14(4):301.

[29] Zheng, W. and van der Laan, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation. Technical Report 273, Division of Biostatistics, University of California, Berkeley.

[30] Zheng, W. and van der Laan, M. J. (2017). Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of Causal Inference*, 5(2).