Targeted Learning in Estimating Heterogeneous Effects and Transporting Direct and Indirect Effects

Levy, Jonathan Mark

2019

Targeted Learning in Estimating Heterogeneous Effects and Transporting Direct and Indirect Effects

By

Jonathan M. Levy

A dissertation completed in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alan E. Hubbard, Chair
Professor Mark J. van der Laan
Professor John M. Colford, Jr.

Spring 2019

Targeted Learning in Estimating Heterogeneous Effects and Transporting Direct and Indirect Effects

Abstract

Targeted Learning in Estimating Heterogeneous Effects and Transporting Direct and Indirect Effects

by

Jonathan M. Levy

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan E. Hubbard, Chair

Targeted learning offers a framework for applying state-of-the-art machine learning in computing estimates, while providing reliable measures of uncertainty for non-parametric and semi-parametric models. Often when applying data adaptive estimation necessary for accurate prediction to reduce bias we lose the ability to bootstrap non-parametrically for inference. This is where targeted maximum likelihood estimators succeed in providing valid inference under conditions we detail, through very inexpensive computation of the standard deviation of the efficient influence curve approximation. We apply the framework mainly to three new parameters of interest, particularly relevant to the field of causal inference and heterogeneous response to treatment. The first two are the variance and cumulative distribution function of the stratum-specific treatment effect function (VTE and TE CDF). The third is transporting from one site to another treatment effects in the presence of an intermediate confounder as well as a mediator, known as stochastic direct and indirect effects (SDE and SIE). We mainly consider SDE and SIE defined by the data in that the stochastic intervention on the mediator is defined by an estimate of the mediator and intermediate confounder mechanisms. We also consider SDE and SIE for both a restricted and unrestricted model that are relevant in practice. We prove efficiency and robustness properties for all the estimators used in this paper as well as software and provide extensive simulations to verify the properties and compare performance with other estimators.

This manuscript contains a generalized method of deriving efficient influence functions central to applying targeted learning for these parameters and others for large models, including for the fixed transported stochastic direct and indirect effects parameters for both restricted and unrestricted models, where the stochastic intervention is defined by the true mechanisms for the mediator and intermediate confounder. The method comes out of a tutorial, featuring the necessary tools of measure theory, integration, functional analysis and efficiency theory, which enables statisticians to embrace estimation for large models often realistic for practical scientific questions. Lastly, this paper implements a new way to perform the targeting in the TMLE process via the discovery of the canonical least favorable submodels (CLFM's) which, are one-dimensional submodels applicable for high dimensional parameters. CLFM's,

used in this paper for estimating many points on the TE CDF, are not only fast but also hold promise for mitigating practical positivity violations. Finally, we employ an easily implementable CV-TMLE procedure, applied on real data for estimating the VTE, that we show retains the attractive properties of Zheng and van der Laan's original CV-TMLE formulation.

Contents

TE Variance and TE CDF: Non Doubly Robust Parameters

1.1 Background and Motivation

A clinician might observe highly variable results for a treatment and want to know how much of this variation is due to confounders, thus motivating more precision in how treatment is assigned. In terms of evaluating public policies, we also might want to know if there are significant portions of the population who receive substantial benefit or harm from an intervention on average. The stratum-specific treatment effect or TE function is defined as the average treatment effect for a randomly drawn stratum of measured patient characteristics and therefore captures heterogeneous average response to treatment. To address the above questions, we first construct the targeted maximum likelihood estimator (TMLE)(van der Laan and Daniel Rubin 2006; van der Laan and Rose 2011) and its cross-validated counterpart (CV-TMLE) (Zheng and van der Laan 2010), to simultaneously estimate the average treatment effect (ATE) and the variance of the TE function (VTE), which gives a sense of the spread of the TE function. With VTE, for instance, one may apply chebyshev's inequality to tail bound the TE distribution.

We then consider the cumulative distribution of the TE function which, tells us the portion of the population whose average effect is below a threshold. The TE CDF is not pathwise differentiable and thus, we will estimate a kernel smoothed version of the parameter, which we will refer to as the smoothed TE CDF, which will depend on the bandwidth used for the kernel. We will provide precise conditions under which we are guaranteed asymptotic efficiency for estimating the smoothed parameter for fixed bandwidth as well as conditions for obtaining a normal limiting distribution when allowing the bandwidth to approach 0 with increasing sample size.

1.1.1 Previous considerations

Much consideration has been given to the distribution of $Y_1 - Y_0$, where Y_a is the counterfactual outcome under the intervention to set treatment to $a \in \{0, 1\}$, as per the Neyman-Rubin potential outcomes framework (Neyman 1923; Donald Rubin 1974). Knowing the distribution of $Y_1 - Y_0$ would give an analyst the individual response to treatment. However, such estimation hinges on recovering the joint distribution of Y_1 and Y_0, a fact Neyman, 1923, realized when computing standard errors for estimating the mean of $Y_1 - Y_0$ in small samples. Assumptions needed to estimate the joint distribution are, in turn, hard to verify. Fisher, 1951 suggests one can essentially leap frog over the issue to create the counterfactual $Y_1 - Y_0$ by careful design. Heckman and Smith, 1998 estimate the quantiles of $Y_1 - Y_0$ via the assumption of quantiles being preserved from Y_1 to Y_0 given a strata of confounders. Using tail bounds (Frechet 1951) to estimate the quantiles of $Y_1 - Y_0$ via the marginals of Y_1 and Y_0 tends to leave too big of a measure of uncertainty to be useful (Heckman and J. Smith 1997).

$var(Y_1 - Y_0)$, in the event it is not 0, is similarly hard to identify. Heckman, 1997 mentions combining the results of Cambanis, 1976 and Frechet, 1951 to tail bound $var(Y_1 - Y_0) = var(Y_1) + var(Y_0) - 2\gamma_{Y_1,Y_0} var(Y_1) var(Y_0)$ as a means to test if the lower bound of the confidence interval includes a variance of 0. For randomized trial data, Ding, Feller et al., 2016, construct a Fisher randomization test of the null hypothesis that $var(Y_1 - Y_0) = 0$ under the untestable assumption that there exists a universal τ so that $Y_1 = Y_0 + \tau$. However, testing this hypothesis is not helpful when it comes to assigning treatment based on confounders. Cox, 1958, assumes $var(Y_1 - Y_0) = 0$ for predefined homogeneous subgroups, essentially assuming the distribution of $Y_1 - Y_0$ is the same as the distribution of $E[Y_1 - Y_0 \mid W]$ for a finite set of W. The variance and CDF of $E[Y_1 - Y_0 \mid W]$ is what we aim to estimate, requiring none of the aforementioned assumptions.

A basic example

As a simple case to give intuition as to what VTE captures, consider $W =$ indicator of male or female, and binary outcome indicating survival if the outcome is 1. Suppose the men have a TE value of $\mathbb{E}[(Y_1 - Y_0)|W = male] = -0.3$ and the females, $\mathbb{E}[(Y_1 - Y_0)|W = female] = 0.7$. Assuming men and women are of equal proportion for the population at hand, then the VTE is 0.25 and ATE is 0.2. This would mean the patient gains from treatment an average 20% with a standard deviation of 50%. One should be reminded that the VTE gives a more personal measure of what to expect from treatment, but not an individual effect variance. For instance, within the male subgroup one might have a high or low varying random variable, $(Y_1 - Y_0 \mid male)$ and such does not count toward the VTE. Hence a clinician's perception of highly varying outcomes does not mean the VTE is high. Rather one would want to estimate VTE to see if the varying outcomes were due to lack of precision in applying the treatment. A similar intuition applies for the TE CDF (although our TE CDF is assumed to be continuous) in that we might have $E[\mathbb{I}(Y_1 - Y_0 \leq -0.5)] = 0.3$. However, the TE CDF is 0 at the TE value of -0.5 in that $Pr\left(E[Y_1 - Y_0 \mid W] \leq -0.5\right) = 0$. In other words, on average, neither men nor women have an effect below -0.5, even though there might be a chance some individuals will have an effect below -0.5.

Estimating VTE and the smoothed TE CDF with a plug-in estimator naturally depends on using an estimate of the outcome model, $\mathbb{E}[Y_a|W]$, from which to estimate the TE function, $\mathbb{E}[Y_1|W] - \mathbb{E}[Y_0|W]$. If one knows the TE function, for instance, one would know an optimal dynamic rule for treatment (A. Luedtke and van der Laan 2016). One could also find subgroup specific treatment effects via the TE function. Lu et al (Tian et al. 2014) offered a way to isolate interactions of treatment with confounders in a randomized trial by transforming the predictors of a parametric model. The main idea is to form a variable, $z = 2A - 1$, where A is the usual treatment indicator, and then put the interaction of this variable with the predictors in the outcome regression. This enables direct estimation of the TE function from which one could obtain a point estimate of the VTE and TE CDF. One could also employ recursive partitioning to divide the data into homogeneous subgroups as far as treatment effects (Athey and Imbens 2016) as well as employ random forests (Athey and Imbens 2015). We could use such subgroups to compute the VTE and smoothed TE CDF but as noted in (Bitler, Gelbach, and Hoynes 2014), establishing too rough subgroups can miss detecting treatment effect heterogeneity. In applying the CV-TMLE or TMLE, our estimators of choice, we also may use tree regression methods within our machine learning ensemble but we are only interested in the predictive power of these methods in eliminating second order remainder term bias, as we will discuss at length.

1.2 Commonalities Between VTE and TE CDF

We will generally follow the roadmap to estimation (van der Laan and Rose 2011), the first steps of which are very similar for VTE and TE CDF. The general TMLE conditions will also be common for all parameters we discuss in this paper.

1.2.1 Full Data Statistical Model and the link to the Observed Data

The system which generates the full data can be used, ideally, to generate counterfactuals (Donald Rubin 1974), i.e. to generate a world under treatment and a parallel world under control and observe every individual's outcome under treatment versus control. This is the ideal situation we would like to have in order to determine causal parameters of interest we precisely define shortly. In this case, our full-data is the same for VTE and the TE CDF.

Our full data, including unobserved measures, is assumed to be independent identically distributed data, generated according to the following structural equations (S. Wright 1921; Strotz and Wold 1960; Pearl 2000). We can assume a joint distribution, $U = (U_W, U_A, U_Y) \sim P_U$, an unknown distribution of unmeasured variables. $X = (W, A, Y)$ are the measured variables. In the time ordering of occurrence we have $W = f_W(U_W)$ where W is a vector of confounders, $A = f_A(U_A, W)$, where A is a binary treatment and $Y = f_Y(U_Y, W, A)$, where Y is the outcome, either binary or continuous. We thusly define a distribution $P_{U,X}$, via $(U, X) \sim P_{U,X}$.

Y_a is a random outcome under $P_{U,X}$ where we intervene on the structural equations to set treatment to $a \in \{0, 1\}$, i.e. $Y_a = f_Y(U_Y, a, W)$. The full model, $\mathcal{M}^F$, consists of all possible P_{UX}. The observed data model, $\mathcal{M}$, is linked to $\mathcal{M}^F$ in that we observe $X = (W, A, Y) \sim P$ where $X = (W, A, Y)$ is generated by P_{UX} according to the structural equations above. Our true observed data distribution is an element of the statistical model, $\mathcal{M}$, which will be non-parametric. In the case of a randomized trial or if we have some knowledge of the treatment mechanism, $\mathcal{M}$ is considered a semi-parametric model and we will incorporate such knowledge, which we will see is much more helpful for estimating ATE than for VTE or the TE CDF.

We can scale a continuous outcome to be in $[0, 1]$ via the transformation $Y_s = \frac{Y-a}{b-a}$ where a and b are minimum and maximum outcomes respectively, obtained from the data or known a priori. A given distribution P in our model defines an outcome model with conditional mean, $\bar{Q}(A, W) = E_P[Y \mid A, W]$, and loss function, $L(P)(w, a, y) = -\log p_Y(y \mid a, w)$, where p_Y is the conditional likelihood of Y given A and W. We also define P_0 as the true distribution.

$$L(P)(w, a, y)dP_0 = -\big(y \log(\bar{Q}(a, w)) + (1-y)\log(1-\bar{Q}(a, w))\big) \tag{1.1}$$

is commonly called quasibinomial loss for a continuous outcome scaled between 0 and 1. Whether the outcome is continuous or binary $\mathbb{E}_{P_0} L(P)(W, A, Y)$ is minimized at the true outcome model (Wedderburn 1974; McCullagh 1983). The targeting portion, which uses either quasibinomial or log-likelihood loss in a logistic regression fluctuation model (see section 1.3.1), will thusly remain the same for binary or scaled continuous outcome. After the CV-TMLE or TMLE algorithm is complete, one may convert the outcomes back to their original scale and they will be naturally constrained within a and b (Gruber and van der Laan 2010). We can then form confidence bands for the parameter of interest on the original scale, for which we offer instruction.

1.2.2 Identification of the Parameters of Interest

We define the stratum-specific treatment effect function or TE function as $b_{P_{UX}}(W) = \mathbb{E}_{P_{UX}}[Y_1|W] - \mathbb{E}_{P_{UX}}[Y_0|W]$. For simultaneously estimating ATE and VTE our parameter of interest is considered to be a mapping $\mathcal{M}^F$ to R^2 defined by

$$\Psi^F(P_{UX}) = (\mathbb{E}_{P_{UX}} b_{P_{UX}}(W), var_{P_{UX}} b_{P_{UX}}(W))$$

For the TE CDF, we define the parameter mapping from $\mathcal{M}^F$ to R^d defined by

$$\Psi^F(P_{U,O}) = (\mathbb{E}_{P_{U,O}} \mathbb{I}(b_{P_{UX}}(W) \leq t_1), \mathbb{E}_{P_{U,O}} \mathbb{I}(b_{P_{UX}}(W) \leq t_2), ..., \mathbb{E}_{P_{UX}} \mathbb{I}(b_{P_{U,O}}(W) \leq t_d))$$

Under the randomization assumption (Robins 1986; Greenland and Robins 1986), $Y_a \perp A|W$ as well as positivity, $0 < E_P[A = a \mid W] < 1$ for all a and W we have that $b_P(W) = \mathbb{E}_P[Y|A = 1, W] - \mathbb{E}_P[Y|A = 0, W] = b_{P_{UX}}(W)$. We can now identify the ATE and VTE as a mapping from the observed data model, $\mathcal{M}$, to $\mathbb{R}^2$ via the gcomp formula (Robins 1986) $\Psi(P) = (\mathbb{E}_P b_P(W), var_P b_P(W)) = \Psi(P^F_{UX})$.

The TE CDF is identified via

$$\Psi(P) = (\mathbb{E}_P \mathbb{I}(b_P(W) \leq t_1), \mathbb{E}_P \mathbb{I}(b_P(W) \leq t_2), ..., \mathbb{E}_P \mathbb{I}(b_P(W) \leq t_d)) \quad = (F(t_1), ..., F(t_d))$$

where F is the CDF of the TE function with respect to P. Ψ is not pathwise differentiable (van der Vaart 2000) so instead we consider the smoothed version of the parameter mapping, using kernel, k, with bandwidth, δ, which is pathwise differentiable and hence, provides a strategy for providing inference for the smoothed TE CDF as well as the TE CDF itself. Here we will suppress k in the notation for convenience and define the i^{th} component of the $d - dimensional$ parameter mapping as

$$\Psi_{\delta,t_i}(P) = \mathbb{E}_W \int_x \frac{1}{\delta} k\left(\frac{x - t_i}{\delta}\right) \mathbb{I}(b(W) \leq x) dx = \int_x \frac{1}{\delta} k\left(\frac{x - t_i}{\delta}\right) F(x) dx$$

so we can write the parameter mapping as

$$\begin{aligned}\Psi_\delta(P) &= \left(\int_x \frac{1}{\delta} k\left(\frac{x - t_1}{\delta}\right) F(x) dx, \int_x \frac{1}{\delta} k\left(\frac{x - t_2}{\delta}\right) F(x) dx, ..., \int_x \frac{1}{\delta} k\left(\frac{x - t_d}{\delta}\right) F(x) dx\right) \\ &= (F_\delta(t_1), ..., F_\delta(t_d))\end{aligned}$$

where $F_\delta(t_i)$ is a shortened notation for the smoothed CDF, $F(t_i)$.

1.2.3 Estimation Methodology

For all the parameters in this report we construct TML estimators. The CV-TMLE involves cross-validating the targeting step, which we detail in 1.3.1 and 3.5. CV-TMLE and TMLE are an integral part of the broad spectrum of targeted learning (van der Laan and Rose 2011), where we employ ensemble machine learning to break from unrealistically narrow parametric model assumptions which, cause large bias and poor coverage. We will see the CV-TMLE has an advantage over the TMLE in that it does not require a donsker condition on the initial predictions for the relevant nuisance parameters, enabling more flexibility in the ensemble learning we employ. In the case of our prediction methods being costly, TMLE and CV-TMLE save considerable time over a non-parametric bootstrap approach to inference, since our inference will be obtained by taking a sample standard deviation of the efficient influence curve approximation. In addition, without the targeting step in our estimator, as outlined in sections and , the non-parametric bootstrap might not even guarantee valid inference for our plug-in estimates (van der Vaart and Wellner 1996).

Despite clear advantages of targeted learning, we will see that estimating VTE and the smoothed TE CDF lacks the desirable robustness properties present when estimating ATE. Particularly for a randomized trial, our CV-TMLE estimate for ATE is consistent where as knowledge of the treatment mechanism does not guarantee consistent VTE or smoothed TE CDF estimates. The lack of robustness is due to stubborn second order remainder terms which, we will discuss in detail.

1.2.4 General TMLE Conditions for Asymptotic Efficiency

We refer the reader to the Targeted Learning Appendix (van der Laan and Rose 2011) as well as (van der Laan 2016; van der Laan and Gruber 2016; van der Laan and Daniel Rubin 2006) for a more detailed look at the theory of TMLE and the use of targeted learning that yields our algorithm below. We offer the reader a brief overview in service of our estimation problems at hand.

In our general discussion, we consider our case of a d-dimensional efficient influence curve, $D^\star_\Psi(P) = (D^\star_{\Psi_1}(P), ..., D^\star_{\Psi_d}(P))$, where $d = 2$ for ATE and VTE and the number of estimated points on the curve for TE CDF. The efficient influence curve is defined at a distribution, P, and is a function of the observed data, $O \sim P$. The variance of $D^\star_{\Psi_j}(P)$ gives the generalized Cramer-Rao lower bound for the variance of any regular asymptotically linear estimator of Ψ_j (van der Vaart 2000). We will shorten the notation to $D^\star_i(P)$ for the i^{th} component of the efficient influence curve.

We will employ the notation, $P_n f$, to be the empirical average of function, $f(\cdot)$, and Pf to be $\mathbb{E}_P f(O)$. We define a loss function, $L(P)(O)$, which is a function of the observed data, O, and indexed at the distribution on which it is defined, P, such that $E_{P_0} L(P)(O)$ is minimized when $P = P_0$, the true data generating distribution. The TMLE procedure maps an initial estimate, $P^0_n \in \mathcal{M}$, of the true data generating distribution to $P^\star_n \in \mathcal{M}$ such that $P_n L(P^\star_n) \leq P_n L(P^0_n)$ and such that $P_n D^\star(P^\star_n) = o_P(n^{-0.5})_{d\times 1}$. $P^\star_n$ is called the TMLE of the initial estimate P^0_n. We can then write an expansion with second order remainder term, R_2, as follows: $\Psi(P^\star_n) - \Psi(P_0) = (P_n - P_0)D^\star_\Psi(P^\star_n) + R_2(P^\star_n, P_0)$.

Theorem 1.2.1. *Define the norm $\|f\|_{L^2(P)} = \sqrt{\mathbb{E}_P f^2}$. Assume the following TMLE conditions:*

1. *$D^\star_j(P^\star_n)$ is in a P-Donsker class for all j, where $D^\star_j$. This condition can be dropped in the case of using CV-TMLE (Zheng and van der Laan 2010). We show the advantages to CV-TMLE in our simulations.*

2. *Second order remainder condition: $R_{2,j}(P^*_n, P_0)$ is $o_p(1/\sqrt{n})$ for all j, where $R_{2,j}$ is the j^{th} component of remainder term.*

3. *$D^\star_j(P^\star_n) \overset{L^2(P_0)}{\longrightarrow} D^\star_j(P_0)$ for all j. This condition will follow from the previous item so usually is not a necessary condition in and of itself for asymptotic efficiency.*

then $\sqrt{n}(\Psi(P^\star_n) - \Psi(P_0)) \overset{D}{\Longrightarrow} N[0_{2\times 1}, cov_{P_0}(D^\star_\Psi(P_0)_{d\times d}]$ where $cov_{P_0}(D^\star_\Psi(P_0)(O)$ is a $d \times d$ matrix in our case with the (i, j) entry given as $E_{P_0} D^*_i(P_0)(O) D^*_j(P_0)(O)$. The i^{th} diagonal of $cov_{P_0}(D^\star_\Psi(P_0)(O)$ is the variance of the $D^*_i(P_0)$ and the limiting variance of $\sqrt{n}(\Psi_i(P^*_n) - \Psi_i(P_0))$ under TMLE conditions. Thus, our plug-in TMLE estimates and individual CI's are given by

$$\Psi_j(P^\star_n) \pm z_\alpha * \frac{\widehat{\sigma}_n(D^\star_j(P^\star_n))}{\sqrt{n}}$$

will be as small as possible for any regular asymptotically linear estimator at significance level, $1 - \alpha$, where $Pr(|Z| \leq z_\alpha) = \alpha$ for Z standard normal and $\widehat{\sigma}_n(D^\star_j(P^\star_n))$ is the sample standard deviation of $\{D^\star_j(P^\star_n)(O_i) \mid i \in 1 : n\}$ (van der Laan and Daniel Rubin 2006).

Theorem 1.2.2. *Note, that if the TMLE conditions hold for the initial estimate, P^0_n, then they will also hold for the updated model, $P^\star_n$, thereby placing importance on our ensemble machine learning in constructing P^0_n (van der Laan 2016).*

Now that we have identified the parameters of interest and reviewed the basic requirements for our estimators to be asymptotically efficient, we will delve into the details of the estimators for each parameter.

1.3 The Variance of the Stratum-Specific Treatment Effect Function, VTE

1.3.1 One-step CV-TMLE Algorithm for ATE and VTE

The reader may consult van der Laan and Gruber, 2016 , for more detail on the one-step TML algorithm, which employs a universal least favorable submodel (ulfm). In section 3.4 we introduce a new iterative analog to the one-step TMLE procedure which also uses one-dimensional parametric submodels, called canonical least favorable submodels (clfm) (Levy 2018c), from which we also define the universal least favorable submodel. van der Laan and Rubin, 2006, constructed a TMLE based on a locally least favorable submodel (lfm) that utilizes parametric submodels of dimension the same dimension as the parameter of interest. TMLE's based the clfm or the lfm require iteration in certain cases, where as the one step TMLE does not. It has therefore been conjectured that the one-step TMLE may better preserve the properties of the initial fit, P_n^0, than the iterative versions, thereby leading to better finite-sample behavior of the second-order remainder term $R_2(P_n^*, P_0)$ (van der Laan and Gruber 2016). If this conjecture is correct, then we would expect similar gains by using the one-step TMLE in our setting, however, in our simulations we found no appreciable differences. For the ATE and VTE estimation we will employ the one-step TMLE but when speed is an issue as with the smoothed TE CDF, we will employ the clfm-based TMLE.

The efficient influence curve for $\Psi(P) = (\Psi_1(P), \Psi_2(P))$, i.e. ATE and VTE, has two components given by

$$D_1^\star(P)(W, A, Y) = \frac{2A-1}{g(A|W)}(Y - \bar{Q}(A, W)) + b_P(W) - \Psi_1(P)$$

$$D_2^\star(P)(W, A, Y) = 2(b_P(W) - \mathbb{E}_P b_P)\frac{2A-1}{g(A|W)}(Y - \bar{Q}(A, W)) + (b_P(W) - \mathbb{E}_P b_P)^2 - \Psi_2(P)$$

where W is a possibly high dimensional set of confounders, A is a binary treatment indicator and Y is a binary outcome or a continuous outcome scaled between 0 and 1. $b_P(W) = E_P[Y \mid A = 1, W] - E_P[Y \mid A = 0, W]$. The reader may visit 3.1.5 for the derivation.

The "Learning" Part of Targeted Learning: Obtaining Initial Estimates To perform a TMLE we use an ensemble learning package such as sl3 (Coyle, Malenica, et al. 2018a) or SuperLearner (Polley et al. 2017) to construct the initial fit, $\bar{Q}_n^0$, of outcome model $E_P[Y \mid A, W]$, and the initial fit, g_n, of the treatment mechanism, $E_P[A \mid W]$, thus providing the estimates $\bar{Q}_n^0(A_i, W_i)$ and $g_n(A_i \mid W_i)$, $i \in 1 : n$, i.e., for all n subjects.

Initialize Targeting Step Compute the negative log-likelihood loss for our outcome predictions. Note, Y_i is the true outcome:

$$P_n L(P_n^0) = -\frac{1}{n}\sum_{i=1}^{n}\left[Y_i \text{log}\bar{Q}_n^0(A_i,W_i) + (1-Y_i)\text{log}(1-\bar{Q}_n^0(A_i,W_i))\right] = L_0 \text{ our starting loss}$$

Compute $H_1^0(A_i,W_i) = \frac{2A_i-1}{g_n(A_i|W_i)}$ and $H_2^0(A_i,W_i) = 2\left(b_n^0(W_i) - \frac{1}{n}\sum_{i=1}^n b_n^0\right)\left(\frac{2A_i-1}{g_n(A_i|W_i)}\right)$ and note H_1 will stay fixed for the entire process for this parameter. Note $b_n^0(W) = \bar{Q}_n^0(1,W) - \bar{Q}_n^0(0,W)$. Compute $\|P_n D_\Psi^\star(P_n^0)\|_2$, where $\|\cdot\|_2$ is the euclidean norm and

$$\begin{aligned} P_n D_\Psi^\star(P_n^0) &= P_n\left(D_1^*(P_n^0), P_n D_2^*(P_n^0)\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n H_1^0(A_i,W_i)(Y_i - \bar{Q}_n^0(A_i,W_i)), \frac{1}{n}\sum_{i=1}^n H_2^0(A_i,W_i)(Y_i - \bar{Q}_n^0(A_i,W_i))\right) \end{aligned}$$

Our initial estimate of the parameter is $(\frac{1}{n}\sum_{i=1}^n b_n^0(W_i), \frac{1}{n}\sum_{i=1}^n (b_n^0(W_i) - \frac{1}{n}\sum_{i=1}^n b_n^0(W_i))^2)$, our sample mean and variance of our estimated TE function values.

The Targeting Step **step 2:** If $|P_n D_j^\star(P_n^m)| < \hat{\sigma}_n(D_j^\star(P_n^m))/n$ for $j \in \{1,2\}$ then $P_n^\star = P_n^m$ and go to step 4. $\hat{\sigma}_n(\cdot)$ denotes the sample standard deviation as in Theorem 1.2.1. This insures that we stop the process once the bias is second order as recursions after this occurs are not fruitful. If $|P_n D_{\Psi_j}^\star(P_n^m)| > \hat{\sigma}/n$, then $m = m+1$ and go to step 3.

step 3

Define the following recursion, using euclidean inner product notation, $\langle\cdot,\cdot\rangle_2$, the same as a dot product:

$$\bar{Q}_n^m(A,W) = expit\left(logit(\bar{Q}_n^{m-1}(A,W)) - d\epsilon\left\langle (H_1^{m-1}(A,W), H_2^{m-1}(A,W)), \frac{P_n(D_\Psi^*(P_n^{m-1})}{\|P_n(D^*(P_n^{m-1})\|_2}\right\rangle_2\right) \quad (1.2)$$

where dϵ is set to 0.0001 (going smaller only costs more without improving accuracy). This recursively defines an estimate, $\bar{Q}_n^m(A,W)$, of the true outcome model, $\bar{Q}_0(A,W) = E_{P_0}[Y \mid A,W]$. Compute $L_m = -\sum_{i=1}^n \left[Y_i\text{log}\bar{Q}_n^m(A_i,W_i) + (1-Y_i)\text{log}(1-\bar{Q}_n^m(A_i,W_i))\right]$. If $L_m \leq L_{m-1}$ then return to step 2. Otherwise $\bar{Q}_n^m = \bar{Q}_n^*$ and continue to step 4.

step 4

Our estimate for ATE and VTE is $\left(\frac{1}{n}\sum_{i=1}^n b_n^*(W_i), \frac{1}{n}\sum_{i=1}^n \left(b_n^*(W_i) - \frac{1}{n}\sum_{i=1}^n b_n^*(W_i)\right)^2\right)$, where $b_n^*(W_i) = \bar{Q}_n^\star(1,W_1) - \bar{Q}_n^\star(0,W_i)$. If the outcome was scaled as $\frac{Y-a}{b-a}$ (see section 2, paragraph 1), then ATE and VTE is $\left(\frac{b-a}{n}\sum_{i=1}^n b_n^*(W_i), \frac{(b-a)^2}{n}\sum_{i=1}^n \left(b_n^*(W_i) - \frac{1}{n}\sum_{i=1}^n b_n^*(W_i)\right)^2\right)$. We compute the standard error estimates by computing the sample standard deviation of the influence curve approximation for each component (see Theorem 1.2.1) i.e. the standard error estimate for the j^{th} component of the estimate is given by $\frac{\widehat{\sigma}_n(D^\star_j(P^\star_n))}{\sqrt{n}}$. If the outcomes were scaled according to $Y_s = \frac{Y-a}{b-a}$, then the standard error estimates for ATE and VTE are multiplied by $b-a$ and $(b-a)^2$, respectively. We can then use the standard normal quantiles z_α as detailed in the previous section (1.96 for 95% CI) for individual confidence bounds or follow the procedure detailed for simultaneous confidence bounds below.

Performing a CV-TMLE

We will adjust the original CV-TMLE procedure (Zheng and van der Laan 2010) for ease of computation without losing any theoretical properties or finite sample performance. The convenience here is that once we obtain initial estimates, there is no difference between CV-TMLE and TMLE as far as implementation is concerned. The reader may consult section 3.5 for the difference between this procedure and the originally defined CV-TMLE (Zheng and van der Laan 2010) regarding our parameter of interest and why neither require condition 1 in Theorem 1.2.1.

To perform a CV-TMLE we would define a split, B_n, which is a mapping on $1 : n$, such that $B_n(i) = 1$ means the i^{th} observation is in the training set and $B_n(i) = 0$ means the i^{th} observation is in the validation set. We usually define 10 splits for which the validation sets are disjoint and comprise all n observations, as in typical 10-fold cross-validation. A CV-TMLE is defined as an average across the splits of estimates computed on the validation sets.

On the training set of each split B_n, we would use an ensemble learning package such as sl3 (Coyle, Malenica, et al. 2018a) or SuperLearner (Polley et al. 2017) to construct the initial fit, $\bar{Q}^0_{n,B_n}$, of outcome model $E_P[Y \mid A, W]$, and the initial fit, g_{n,B_n}, of the treatment mechanism, $E_P[A \mid W]$. We would then predict the outcome and treatment probabilities on the validation set defined by B_n. For all $i \in 1 : n$, we therefore provide an estimate $\bar{Q}^0_n(A_i, W_i)$ and $g_n(A_i \mid W_i)$ for when observation i was in the validation set of one of the splits. With these predictions we can proceed with steps 2 through 4 above, yielding a CV-TMLE.

Simultaneous Estimation and Confidence bounds

We often want to provide confidence intervals that simultaneously cover all the coordinates of $\Psi(P_0)$ at a given significance level. The following is an added benefit of having the efficient influence curve at hand for we can account for correlated estimates in a tighter manner than a bonferroni correction (Dunn 1961). The reader may note we offer the general d-dimensional version here and thus, for (ATE, VTE) $d = 2$. After completing the above algorithm we have, $D^*_\Psi(P^*_n)(O_i) = (D^*_1(P^*_n)(O_i), ..., D^*_d(P^*_n)(O_i))$, for each subject indexed by $i \in 1 : n$. Consider the d-dimensional random variable $Z_n = (Z_{n,1}, ..., Z_{n,d}) \sim N(0_{d\times 1}, \Sigma_n)$, defined by two by two matrix, Σ_n, the sample correlation matrix of $D^\star_\Psi(P^\star_n)$. Let $q_{n,\alpha}$ be the α^{th} quantile of the random variable $M_n = max(|Z_{n,d}|, ..., |Z_{n,d}|)$. Let $Z = (Z_1, ..., Z_d) \sim N[0_{d\times 1}, \Sigma]$, where Σ is the correlation matrix of $D^\star_\Psi(P_0)$. Let q_α be the α^{th} quantile of the random variable $M = max(|Z_1|, ..., |Z_d|)$, i.e., the α^{th} quantile of the random variable giving the max number of standard deviations over the coordinates of Z. We monte-carlo sample 5 million draws from the random variables M_n to find $q_{n,\alpha}$. We note that 5 million is a sufficient number to guarantee very little error in finding the true $q_{n,\alpha}$. Applying the continuous mapping theorem (van der Vaart and Wellner 1996) assures us under TMLE conditions that $Z_{n,i} \pm q_\alpha$ covers Z_i for all $i \in 1 : d$, at $(1-\alpha) \times 100\%$. Then we can apply the continuous mapping theorem again to assure us $q_{n,\alpha} \longrightarrow q_\alpha$. $q_{n,\alpha}$ is therefore an estimate of the number of standard errors needed to simultaneously cover both true parameter values at $(1-\alpha) \times 100\%$. This results in the confidence bands

$$\hat{Psi}_{j,n} \pm q_{n,\alpha} * \frac{\widehat{\sigma}_n(D^\star_j(P^\star_n))}{\sqrt{n}}$$

which, will asymptotically cover all coordinates, $\Psi_j(P_0)$ of $\Psi(P_0)$, simultaneously at the significance level, $1-\alpha$. The reader may note $q_{n,\alpha}$ is very close to the bonferroni correction (Dunn 1961) if Σ_n is the identity matrix.

Remark. From here on, all theorems will apply to either TMLE or CV-TMLE so we will use the lighter TMLE notation where we need not keep track of the splits, B_n.

1.3.2 The Unforgiving Remainder Term in VTE Estimation

Computation of the remainder term is in the section 3.2.1 and is accompanied by more rigorous analysis. Here we provide the reader with the necessary results for our discussion. For convenience we define the true outcome model to be $\bar{Q}_0(A, W) = E_{P_0}[Y \mid A, W]$ and the true treatment mechanism as $g_0(A \mid W) = E_{P_0}[A \mid W]$. Let $\bar{Q}^0_n$ be the initial estimate of $\bar{Q}_0$, and g_n be the estimate for g_0. For estimating ATE and VTE, we will fluctuate an initial outcome model fit, $\bar{Q}^0_n$ to $\bar{Q}^*_n$ but g_n will not change. $b^*_n(W) = \bar{Q}^*_n(1, W) - \bar{Q}^*_n(0, W)$. We also note, that we have used the empirical distribution, $Q_{W,n}$, to estimate Q_W, the distribution of W, which also remains fixed as in van der Laan and Rubin, 2006. The second order remainder term for VTE is:

$$\begin{aligned}
&R_2(P^*_n, P_0) = \Psi(P^*_n) - \Psi(P_0) + P_0\left(D^\star_\Psi(P)\right) && (1.3)\\
&= \left(\mathbb{E}_0 b_0(W) - P_n b^*_n(W)\right)^2 && (1.4)\\
&+ \mathbb{E}_0\Bigg[2\left(b^*_n(W) - P_n b^*_n(W)\right) * \Bigg(\frac{g_0(1|W) - g(1|W)}{g(1|W)} * \\
&\left(\bar{Q}_0(1, W) - \bar{Q}^*_n(1, W)\right) - \frac{g_0(0|W) - g(0|W)}{g(0|W)}\left(\bar{Q}_0(0, W) - \bar{Q}^*_n(0, W)\right)\Bigg)\Bigg] && (1.5)\\
&- \mathbb{E}_0\left(b_0(W) - b^*_n(W)\right)^2 && (1.6)
\end{aligned}$$

Considering (1.4) and (1.5) above, we need

$$\|\bar{Q}^*_n - \bar{Q}_0\|_{L^2(P_0)} \|g_n - g_0\|_{L^2(P_0)} \tag{1.7}$$

to be $o_P(n^{-0.5})$. If the first factor is $o_P(n^{r_{\bar{Q}}})$ and the second is $o_P(n^{r_g})$, then $r_{\bar{Q}} + r_g \leq -0.5$ will satisfy the TMLE remainder term condition 2 of Theorem 1.2.1. It is notable the terms disappear in the case of a randomized trial where we incorporate the known g_0. (1.7) is also a generous upper bound for the first two terms, which depend on $\int(\bar{Q}_n^* - \bar{Q}_0)(g_n - g_0))dP_0$ because the integrand can change sign. However, (1.6) is not generous in this way because the integrand is a square. Precisely, we require $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)}$ to be $o_P(n^{-0.25})$ with no help provided by knowing the treatment mechanism. Hence, VTE estimation is not doubly robust. In an randomized clinical trial, the TMLE or CV-TMLE estimate for ATE will be consistent but such is not the case for VTE. We can apply a large data adaptive ensemble of state-of-the-art machine learning algorithms to mitigate this remainder term but we still have found it can cause bias leading to poor coverage.

1.3.3 Simulations for VTE

We performed two different kinds of simulations, the first primarily to verify the remainder conditions in the theory of TMLE (condition 2, Theorem 1.2.1). The rest were performed to get a sense of what might occur with real data. Inference for all TMLE's used the sample standard deviation of the efficient influence curve approximation to form confidence intervals as per section 2. For logistic regression plug-in estimators of ATE and VTE, confidence bands were formed by using the delta method and the influence curve for the beta coefficients for intercept, main terms and interactions (see section 3.3.3). SuperLearner initial estimates had no accompanying measure of uncertainty since there is little theory for such, even if non-parametrically bootstrapping (van der Vaart and Wellner 1996).

Simulations with Controlled Noise

Instead of drawing W then A and then Y under a data generating distribution and then trying to recover the truth with various predictors or SuperLearner as we do later, we directly add heteroskedastic noise to $\bar{Q}_0$ in such a way that the conditions of TMLE hold and then use the noisy estimate as the initial estimate in the TMLE process. This does not necessarily match what happens in practice because the noise we add is not related to the noise in the draw of Y given A and W. However, it is a valid way to directly test the conditions of TMLE in that we can control the noise so that the TMLE conditions hold and watch the asymptotics at play. We also note that we will assume g_0 is known because the other second order terms for VTE, involving bias in estimating g_0, are dependent on double robustness in the same way as for the ATE, for which the properties of TMLE are already well-known (van der Laan and Daniel Rubin 2006; van der Laan and Rose 2011).

Simulation Set-up $W_1 \sim uniform[-3,3]$, $W_2 \sim binomial(1,.5)$, $W_3 \sim N[0,1]$ and $W_4 \sim N[0,1]$. We define $g_0(A|W) = expit(.5*(-0.8*W_1 + 0.39*W_2 + 0.08*W_3 - 0.12*W_4 - 0.15))$, which is the true density of A given W. We kept our propensity scores between about 0.17 and 0.83 so as to avoid poor performance from positivity violations (Petersen et al. 2012). $\mathbb{E}_0[Y|A,W] = \bar{Q}_0(A,W) = expit(.2*(.1*A + 2*A*W_1 - 10*A*W_2 + 3*A*W_3 + W_1 + W_2 + .4*W_3 + .3*W_4))\}$ which defines the density of Y given A and W for a binary outcome. Define the TE function as $b(W) = \mathbb{E}_0[Y|A=1,W] - \mathbb{E}_0[Y|A=0,W]$ and we have $\Psi(P_0) = var_0(b(W)) = 0.0636$. This is a substantial VTE to avoid getting near the parameter boundary at 0.

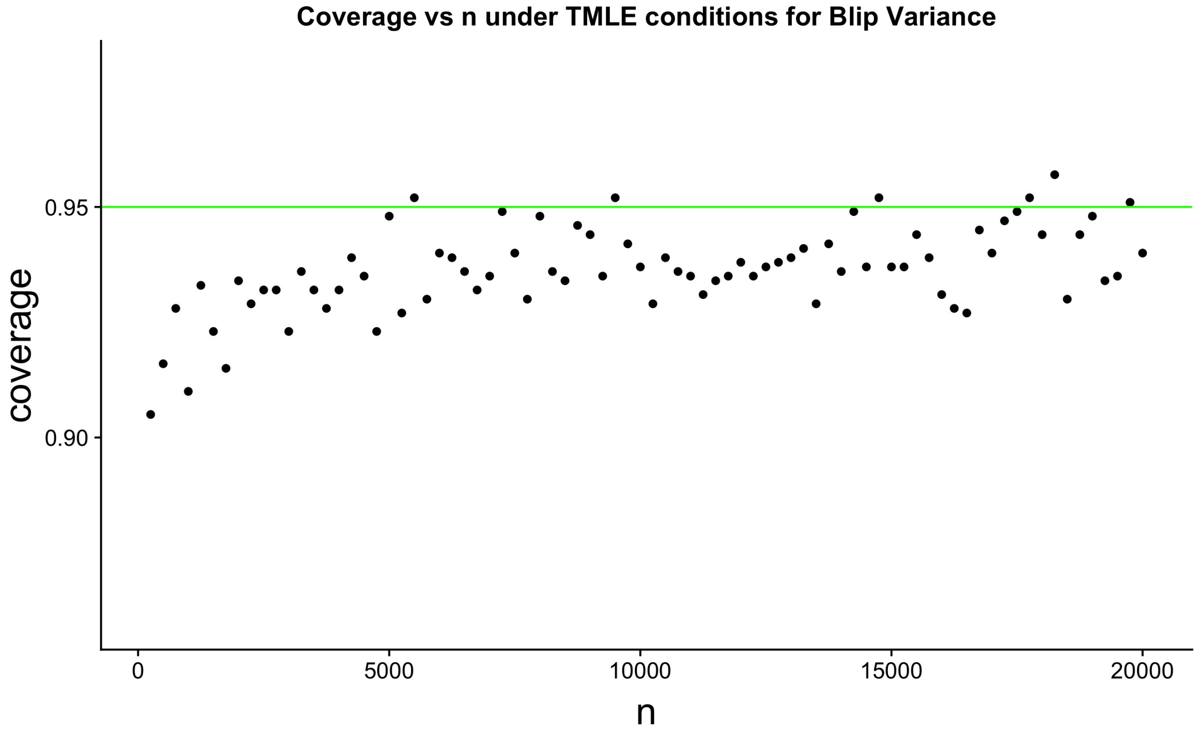
Below we illustrate the process for one simulation. For each sample size, n, we performed the simulation 1000 times. We note that *rate* is some number which we will set to less than -1/4 (-1/3 in this case) in order to satisfy TMLE conditions.

1. define $bias(A, W, n) = 1.5n^{rate}(-.2 + 1.5A + 0.2W_1 + W_2 - AW_3 + W_4)$
2. define heteroskedasticity: $\sigma(A, W, n) = 0.8n^{rate}|3.5+0.5W_1+0.15W_2+0.33W_3W_4-W_4|$
3. define $b(A, W, n, Z) = bias(A, W, n) + Z \times \sigma(A, W, n)$ where Z is standard normal
4. draw $\{Z_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ each from standard normals
5. $\bar{Q}_n^0(1, W_i) = expit\left(logit\left(\bar{Q}_0(1, W_i)\right) + b(1, W_i, n, Z_i)\right)$
6. $\bar{Q}_n^0(0, W_i) = expit\left(logit\left(\bar{Q}_0(0, W_i)\right) + 0.5b(1, W_i, n, Z_i) + \sqrt{0.75}b(0, W_i, n, X_i)\right)$
7. $\bar{Q}_n^0(A, W) = A * \bar{Q}_n^0(1, W) + (1 - A)\bar{Q}_n^0(0, W)$

We note that we placed correlated noise on the true $\bar{Q}_0(1, W)$ and $\bar{Q}_0(0, W)$ so as to make the TE function"estimates" of similar noise variance as the initial "estimates" for $\bar{Q}_0(A, W)$. By a Taylor series expansion about the truth, it is easy to see the above procedure will satisify the remainder term conditions of Theorem 1.2.1. We have that $\bar{Q}_n^0(1, W) = \bar{Q}_0(1, W)+\bar{Q}_0(1, W)(1-\bar{Q}_0(1, W))b(1, W, n, Z)+O(b^2(1, W, n, Z))$ and likewise for $\bar{Q}_n^0(0, W)$ and thus trivially, $\sqrt{\mathbb{E}_0\left(b_n^0(W) - b_0(W)\right)^2}$ is of order n^{rate} with $rate < -1/4$. As previously mentioned, we need not worry about any second order terms but $\mathbb{E}_0\left(b_n^0(W) - b_0(W)\right)^2$ because we are using the true g_0. Condition 1 of Theorem 1.2.1 is easily satisfied and Condition 3, the donsker condition, is satisfied since our "estimated" influence curve, $D^*(\bar{Q}_n^0, g_0)$, depends on a fixed function of A and W with the addition of independently added random normal noise.

The simulation result, displayed in 1.1, is in alignment with the theory established for the TMLE estimator of VTE but how fast the asymptotics come into play is an important issue as to the relevance of the asymptotic theory.

Figure 1.1



Coverage vs n under TMLE conditions for Blip Variance

coverage slowly becomes nominal as expected
blip bias is O_P(n^{-1/3} i.e. o_P{n^{-1/4}
n increases by 250 for each point moving to the right, starting at 250

Simulations That Are More Realistic

We will stick with binary outcome and treatment, though the results will be comparable for continuous outcome. Unless otherwise noted, sample size n = 1000 and the number of simulations = 1000. Throughout the simulations we generated the covariates as follows: $W_1 \sim uniform[-3, 3]$, $W_2 \sim$ standard normal, $W_3 \sim$ standard normal and $W_4 \sim$ standard normal. These simulations are more realistic in that we try to recover via machine learning, an ”unknown” treatment mechanism and outcome model. When we specify the models correctly we are considering a ”best case” scenario where our regressions achieve parametric rates of convergence to the truth. When we misspecify a model in the data generating system, we try to recover its non-linear functional form with ensemble machine learning, in the event that a linear model including interactions (to pick up heterogeneity) is catastrophic for estimating VTE. If we are going to estimate VTE, a main terms linear model will assume VTE is essentially 0, so comparing ensemble learning methods with such is not very informative.

Well-specified TMLE Initial Estimates, Skewing

Here we apply logistic regression to the correct functional form for both outcome model, $\mathbb{E}[Y|A, W] = \bar{Q}(A, W)$ and treatment mechanism, $\mathbb{E}[A|W] = g_0(A, W))$, thus achieving parametric rates of convergence. The only point of these simulations is to show that TMLE preserves well-specified initial estimates and also to show approximately what size sample will lead to skewing (and therefore bias) of the sampling distribution for TE variance when the truth is near the lower parameter bound of 0. We can say as a rule of thumb, a sample size of 500 or more is probably needed to even hope to get reliable estimates for TE variances in the neighborhood of 0.025 (15.8% standard deviation), a rule confirmed by figures 1.2, 1.3 and 1.4. $\bar{Q}(A,W) = expit(0.14(2A + W1 + aAW_1 - bAW_2 + W_2 - W_3 + W_4))$ for the outcome regression, varying a and b to adjust the size of the TE variance. $\mathbb{E}[A|W] = g0 = expit(-0.4 * W1 + 0.195 * W2 + 0.04 * W3 - 0.06 * W4 - 0.075)$ was the true treatment mechanism and we avoid positivity violations here by keeping our propensity scores mostly between 0.10 and 0.90.

Figure 1.2

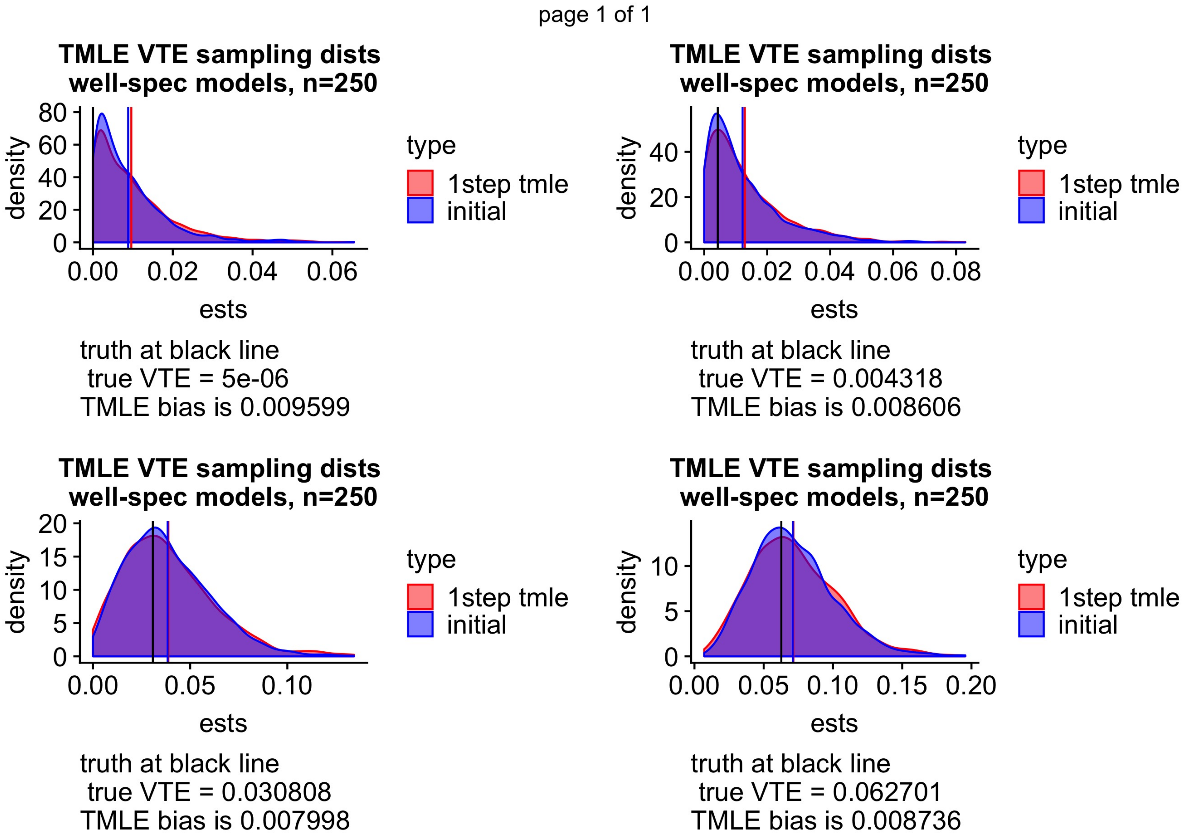
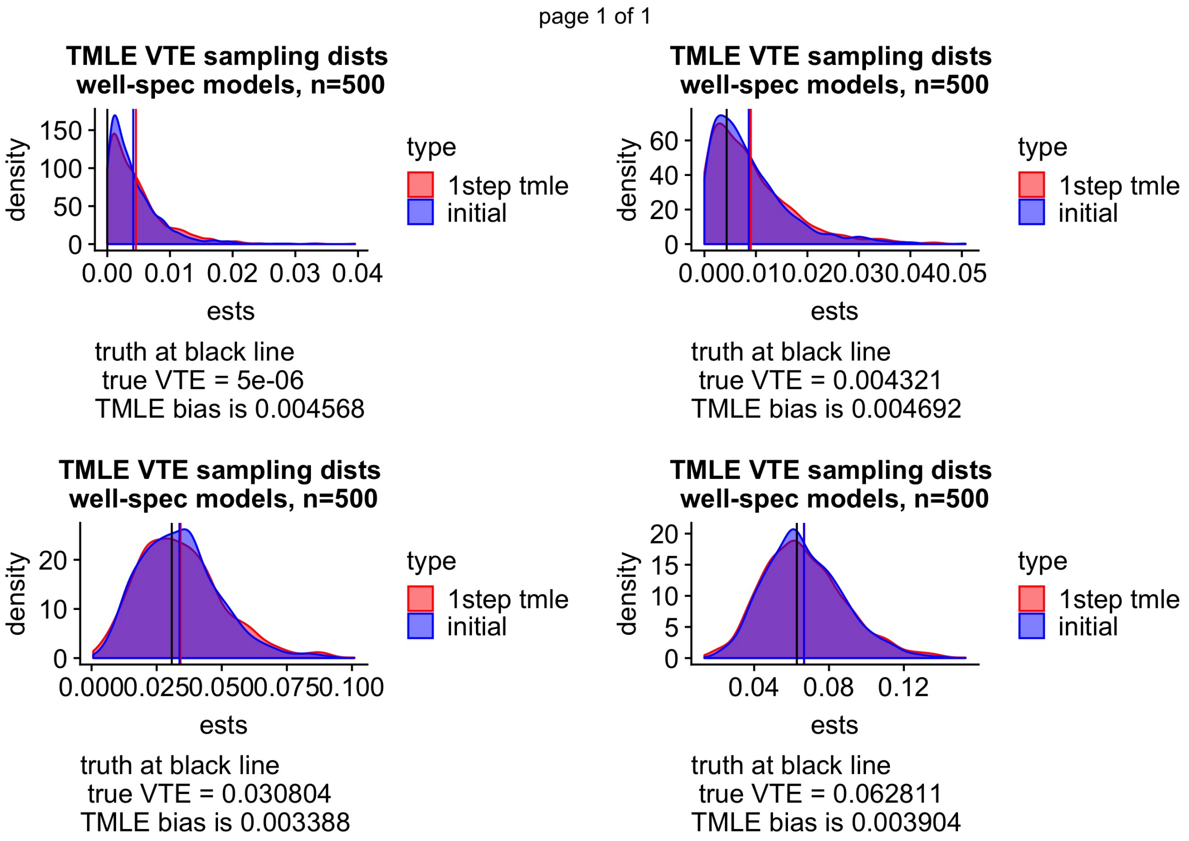


Figure 1.3



TMLE VTE sampling dists well-spec models, n=500

truth at black line
true VTE = 5e-06
TMLE bias is 0.004568

TMLE VTE sampling dists well-spec models, n=500

truth at black line
true VTE = 0.004321
TMLE bias is 0.004692

TMLE VTE sampling dists well-spec models, n=500

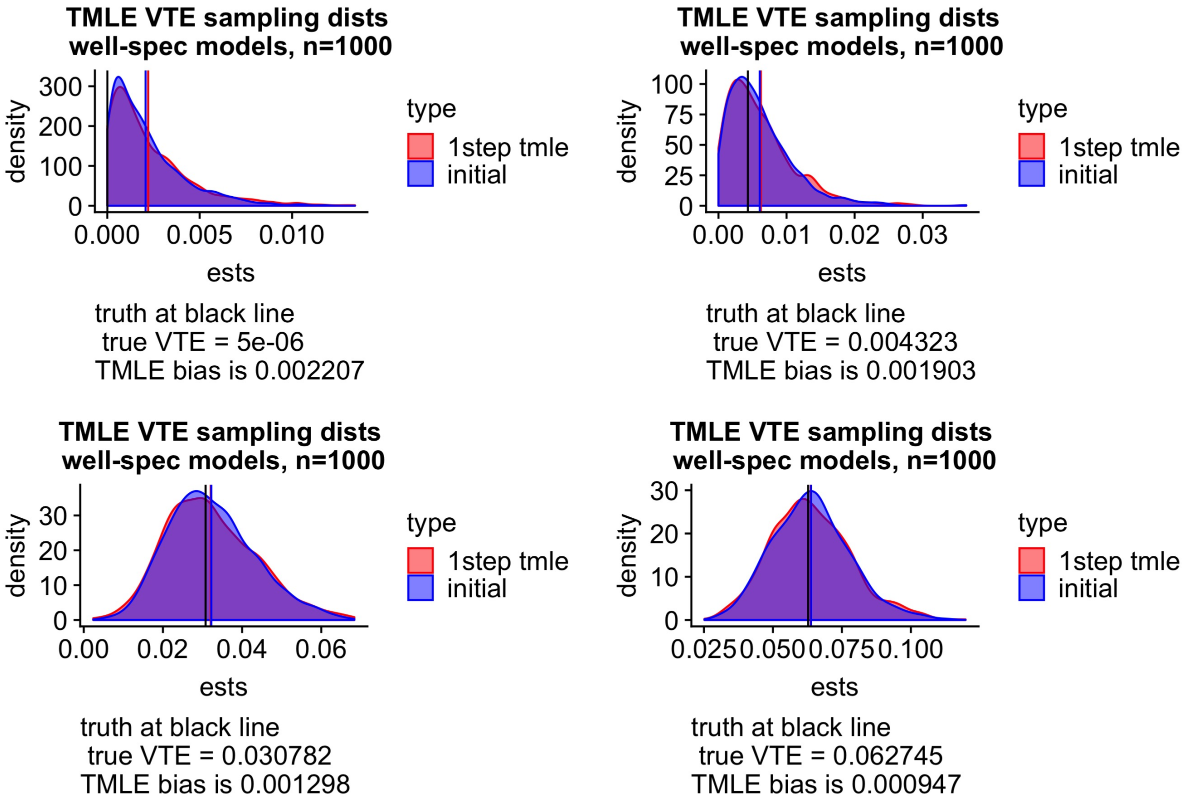
truth at black line
true VTE = 0.030804
TMLE bias is 0.003388

TMLE VTE sampling dists well-spec models, n=500

truth at black line
true VTE = 0.062811
TMLE bias is 0.003904

Figure 1.4

page 1 of 1



SuperLearner Details For Remaining Simulations

Targeted learning (van der Laan and Rose 2011) features the use of data adaptive prediction methods optimized by the ensemble learning R packages, such as SuperLearner (Polley et al. 2017), H2O (LeDell 2017) or the most recent sl3 (Coyle, Malenica, et al. 2018a). Superlearner, which picks the best single algorithm in the library, as decided by the cross-validation of a valid loss function, has risk that converges to the oracle selector at rate $O\left(log\left(k(n)\right)/n\right)$ where $k(n)$ is the number of candidate algorithms, under very mild assumptions on the library of estimators (van der Laan, Polley, and Hubbard 2007). Generally the best or nearly best learner in the library is the optimal convex combination of algorithms that forms the SuperLearner predictor which, might be more familiar to the reader as a form of model stacking (Wolpert 1992).

Use of Highly Adaptive Lasso: Making Initial Predictions $\bar{Q}_n^0$ and g_n

It is notable that if we use the highly adaptive lasso (HAL) (van der Laan 2016; Benkeser and van der Laan 2016) for our nuisance parameter fits (outcome model and treatment mechanism, if treatment mechanism is unknown), we will yield asymptotically efficient TMLE's, assuming the true models are right-hand continuous with left-hand limits and have variation norm smaller than a constant M (van der Laan 2016). In finite samples, however, some machine learning algorithms might be better suited for prediction and so we rely on ensemble learning with HAL as one of the candidate estimators, thus still retaining the above guarantee.

SuperLearner Library 1, termed SL1, Avoiding Overfitting

This library will be indicated by ”SL1” in the simulation results.

1. SL.gam3, a gam (Hastie 2017) using degree 3 smoothing splines, screening main terms, top 10 correlated variables with the outcome and top 6.

2. SL.glmnet_1, SL.glmnet_2 and SL.glmnet_3 (Friedman, Hastie, and Tibshirani 2010) performed a lasso, equal mix between lasso and ridge penalty and ridge regressions.

3. nnetMain_screen.Main (Venbles and B. D. Ripley 2002) is a neural network with decay = 0.1 and size = 5 using main terms.

4. earthMain (Milborrow 2017) is data adaptive penalized regression spline fitting method. They allow for capturing the subtlety of the true functional form. We allowed degree = 2, which is interaction terms with the default penalty = 3 and a minspan = 10 (minimum observations between knots).

5. SL.glm (R Core Team 2017) logistic regression and we used main terms, top 6 correlated variables with outcome and top 10 as well as a standard glm with main terms and interactions (glm_mainint_screen.Main)

6. SL.stepAIC (R Core Team 2017) uses Akaike criterion in forward and backward step regression

7. SL.hal is the highly adaptive lasso (Benkeser, Kennedy, and Sofrygin 2016)

8. SL.mean returns the mean outcome for assurance against overfitting

9. rpartPrune (Therneau, Atkinson, and B. Ripley 2017) is recursive partitioning with cp = 0.001 (must decrease the loss by this factor) minsplit = 5 (min observations to make a split), minbucket = 5 (min elements in a terminal node)

SuperLearner Library 2 termed SL2, More Aggressive, overfits a little

This library will be indicated by ”SL2” in the simulation results. This library is identical to Library 1, except we added the following learners, which were tuned to maximize cross validated loss on a few draws from case 2a data generating distribution. Thus these additions do not severely overfit, in general.

1. SL.ranger (M. N. Wright and Ziegler 2017): A random forest which picked 3 features at a time formed 2500 trees and had a minimum leaf size set to 10.

2. SL.xgboost (Chen et al. 2017): One xgboost fit on all main terms and interactions with stumps (depth 1 trees), allowing a minimum of 3 observations per node, a learning rate of 0.001 and summing 10000 trees. We also included an xgboost using depth of 4 trees on main terms only with same shrinkage and minimum observations per node but only 2500 trees.

Case 1: Well-Specified Treatment Mechanism, Misspecified Outcome

The following example, encapsulated in figure 5, demonstrates three things

1. **Enormous gains possible with flexible estimation**. Using a logistic regression with main terms and interactions plug-in estimator and the delta method for inference, yielded a bias of -0.065 (the truth is 0.079), missing almost the entire TE variance and covering at 0%. The TMLE could not help the initial estimates using the same logistic regression so reliance on a parametric model can be a disaster as opposed to ensemble learning.

2. **Difference in robustness** between estimating causal risk difference and TE variance. The severely misspecified logistic regression with main terms and interactions initial estimate for the outcome model and well-specified treatment mechanism yielded a TMLE for causal risk difference (which is doubly robust) that covered at 95.6%, where as for TE variance it never covers the truth.

3. **The advantage of CV-TMLE over TMLE**. The same SuperLearner used for initial estimates yields some skewing and bad outliers as well as bigger bias and variance for TMLE as opposed to a normally distributed CV-TMLE sampling distribution. Just some overfitting by random forest about 20% of the time out of the library of 18 learners managed to cause outliers for TMLE, ruining normality of the sampling distribution and causing higher bias and variance, where as CV-TMLE appeared unaffected by the overfitting. Overfitting means essentially that the metric entropy of the class of functions considered by random forest was too big. To give some intuition behind the donsker TMLE condition, an example of a large donkser class is the set of functions of bounded variation (van der Vaart and Wellner 1996) meaning the function class is smooth in some sense, not allowing unlimited ups and downs between predictions, such as overfitting allows. Since the influence curve approximation is defined partially in terms of the mean outcome model, overfitting causes the class of functions for the influence curve approximation to be non-donsker as well. When trying to do a good job estimating the mean outcome model as in this simulation, CV-TMLE allows highly adaptive machine learning we need to minimize the second order remainder bias without paying a big price for overfitting.

Simulation Set-up, Case 1

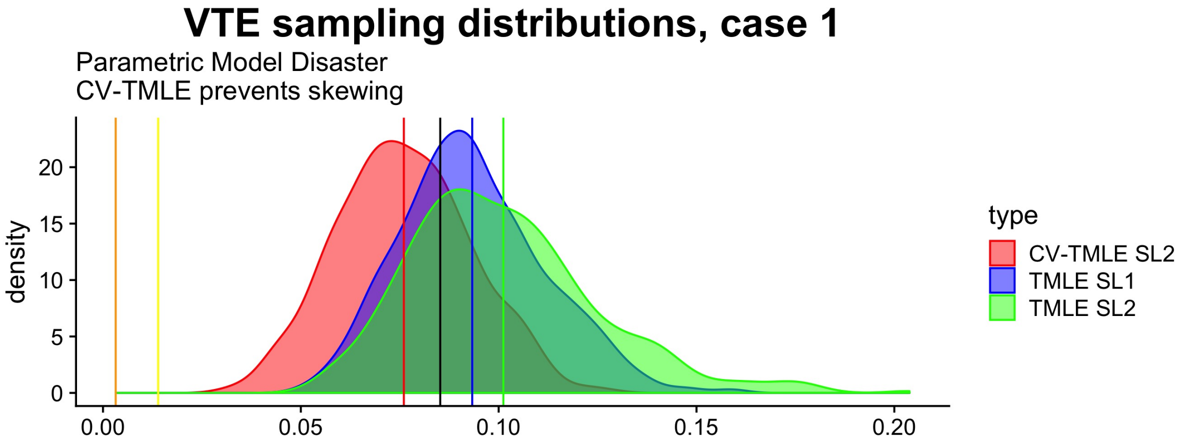
$\mathbb{E}[Y|A,W] = Q0 = expit(0.28 * A + 2.8 * cos(W1) * A + cos(W1) - 0.56 * A * (W2^2) + 0.42 * cos(W4) * A + 0.14 * A * W1^2)$. $\mathbb{E}[A|W] = g0 = expit(-0.4 * W1 + 0.195 * W2 + 0.04 * W3 - 0.06 * W4 - 0.075)$, which we will specify model correctly in all cases with a linear logisitic fit. True Causal Risk Difference = 0.078. True CATE Variance = 0.085.

Table 1.1: Performance of the Estimators, Case 1

	var	bias	mse	coverage
TMLE LR	0.00001	-0.08207	0.00675	0
LR plug-in	0.00005	-0.07134	0.00514	0
CV-TMLE SL2	0.00028	-0.00930	0.00037	0.87375
CV-TMLE SL2∗	0.00028	-0.00924	0.00037	0.88577
TMLE SL2	0.00057	0.01584	0.00082	0.83100
TMLE SL2∗	0.00057	0.01591	0.00082	0.86000
TMLE SL1	0.00033	0.00802	0.00040	0.93193
TMLE SL1∗	0.00033	0.00804	0.00040	0.93994

* indicates causal risk difference and TE variance estimated simultaneously with 1step tmle covering both parameters for 95% simultaneous confidence intervals. LR indicates logistic regression with main terms and interactions. SL1 Library did not overfit out 20% of the time with one out of 18 algorithms, still causing outliers, necessitating CV-TMLE as a precaution.

Figure 1.5



VTE sampling distributions, case 1

Parametric Model Disaster
CV-TMLE prevents skewing

Truth is at black vline. Orange and yellow lines mark means of TMLE using logistic regression with main terms and interactions for the initial outcome predictions and the like regression plug-in estimator respectively.
Both of these never cover the truth and are disastrously biased.
TMLE SL2, which used Superlearner Library 2 for initial ests
is skewed, has many outliers and covers at 83.1% is both more biased and more variant.
TMLE SL1 uses a non-overfitting SuperLearner and covers near nominally at 93.2%
CV-TMLE SL2 does not require the donsker condition lowest MSE, no skewing and covers at 87.4%. despite use of overfitting SL2.

1.3.4 Mixed Results Pointing to a Need for Future Refinements

We again demonstrate how employing targeted learning with CV-TMLE can recover a misspecified treatment mechanism as well as outcome models when parametric models are terrible, but coverage is below nominal and at times very poor, depending on the situation. This is not a problem solely for the case of an observational study as the authors have found the main culprit in poor coverage to be the second order remainder term consisting of the integral of true TE function minus estimated TE function squared (section 1.3.2). Standard parametric models are again disastrous for both cases 2 and 3, never covering the truth and missing almost all of the VTE as in case 1. Only in case 2 does CV-TMLE, using a pretty small superlearner library of 8 algorithms including, xgboost, neural networks, glm with main terms and interactions, earth, sample mean and the highly adaptive lasso (van der Laan 2016), achieve decent coverage of 83% and reduces bias of the initial estimate from -0.015 to -0.009. In case 3, CV-TMLE with the same SuperLearner library only covers at 32%.

For both case 2 and case 3 we used the following model for the true treatment mechanism:
$\mathbb{E}_0[A \mid W] = expit(.4 * (-0.4 * W1 * W2 + 0.63 * W2^2 - .66 * cos(W1) - 0.25))$
Case 2 true outcome model:

$$\begin{aligned}\mathbb{E}_0[Y \mid A, W] = expit(&0.1 * W1 * W2 + 1.5 * A * cos(W1) + \\ &0.15 * W1 - .4 * W2 * (abs(W2) > 1) - 1 * W2 * (abs(W2 <= 1)))\end{aligned}$$

Case 3 true outcome model:
$\mathbb{E}_0[Y \mid A, W] = expit(0.2*W1*W2+0.1*W2^2-.8*A*(cos(W1)+.5*A*W1*W2^2)-0.35)$

Supplementary Results

The reader may visit Jonathan Levy's github for instructions and software on how to reproduce the results obtained in this paper, as well as more detailed results that were not included in this manuscript.

1.3.5 Demonstration on Real Data

The CV-TMLE estimator, as detailed in this paper, was applied to a real dataset. GER-INF is a placebo-controlled randomized trial published in 2002 that evaluated the impact on mortality of low dose steroid administration in patients hospitalized in the intensive care unit (ICU) for septic shock (Annane et al. 2002). This study was performed in 19 ICUs in France and enrolled a total of 299 patients. Because steroid supplementation in this context was expected to be beneficial in patients with relative adrenal insufficiency, a corticotropin stimulation test was performed in all patients. A pre-specified subgroup analysis was planned in patients who were nonresponders to the corticotropin stimulation test (relative adrenal insufficiency) and in those responders to the corticotropin stimulation test (no relative adrenal insufficiency). In this sample, 7 days of low dose hydrocortisone associated with fludrocortisone were associated with a reduced risk of death in patients with septic shock. As expected, this reduction was significant in patients with relative adrenal insufficiency as reflected by a lack of response following the corticotropin stimulation test. Because of the apparent heterogeneity in treatment effect, at least partially explained by the presence of a relative adrenal insufficiency, we used the proposed estimator to quantify treatment effect variability (VTE) across the strata of patients.

We controlled for the following confounders in fitting the outcome model: IGS2 or SAPS2 severity score (Simplified Acute Physiology Score to assess mortality), Sequential Organ Failure Assessment (SOFA) severity score at baseline, lactate level (LACTA), cortisol level before corticotropin (CORT0), an indicator of responding to the corticotropin stimulation test (RESPONDER), site of infection (site), mechanical ventilation at baseline (VM0), patient origin (hospital acquired infection or not) (origine), indicator of medical, elective surgery or urgent surgery (typeadmission), maximum difference in cortisol concentration before and after stimulations (DELTA_CORTmax), indicator of use of etomidate for anesthesia (drug known to alter the adrenal function), blood sugar and the pathogen responsible for the infection (GLYC) and the responsible pathogen (GERME). In this case, the treatment assignment was random and thus we can identify VTE from the data as a measure of how much of the heterogeneity in treatment effect is due to confounders normally used to assign treatment. We note, in the case of variables missing data, we create an indicator of missingness and use the median or, in the case of categorical variables, the most popular category as an imputed value. The table below summarizes our data.

GER-INF Summary, N=299

Variable	
Outcome: Renal Failure	
Yes	173
No	126
Treatment: Rec. Steroid	
Yes	150
No	149
Age in yrs	60.8 (16.1)
IGS2	62.6 (23.6)
SOFA0	11 (3.2)
missing	24
LACTA	4.4 (3.3)
CORT0	23.3 (30.9)
DELTA_CORTmax	6.1 (22.3)
RESPONDER	
GLYC	175.8 (106.2)
missing	6
Yes	70
No	229
VMO	
Yes	298
No	1
origine	
Yes	112
No	187
etomidate	
Yes	76
No	213
site	
Multiple	144
Lung	89
GI	31
Soft Tissue	16
Bacteremia	6
Other	12
missing info	1
typeadmission	
1:	179
2:	10
3:	110
GERME	
type 1	72
2	38
3	4
4	2
5	24
6	9
7	150

We provide estimates below for ATE and VTE simultaneously and used the delta method to also give a confidence interval for the $\sqrt{VTE}$, because such is on the scale of measurement of ATE. We can see the left bounds of the confidence interval for VTE and $\sqrt{VTE}$ strayed into the negative numbers, which are not possible estimates for such parameters. Log-scaling the confidence intervals only make them excessively large and therefore not useful due to the unscaled confidence bands centering close to 0. If we reference our discussion about skewing in section 3.2.1, we see that if the true VTE were as small as our estimate, then our sampling distribution under the best case scenario of well-specified outcome model is skewed. We would need a true VTE of around 0.06 to have any hope of having normal sampling distributions for estimating VTE. There is also the issue of second order remainder term bias as we discussed, which could account for missing a truly larger VTE in our estimates. The second order remainder term in Theorem 1.2.1 is the square of the L^2 norm bias in estimating the TE function so, prioritizing the TE function in our outcome model estimation is the subject for future work to improve TE function estimation in finite samples.

Table 1.2: CV-TMLE Results for Simultaneous Estimation of ATE, VTE and sqrt(VTE)

	est	se	lower	upper
ATE	-0.088	0.050	-0.199	0.023
VTE	0.002	0.004	-0.008	0.012
sqrt(VTE)	0.045	0.048	-0.061	0.152

SuperLearner

We used a SuperLearner library consisting of 40 algorithms, including main terms and interactions when applying any regression methods, such as logistic regression, bayes generalized linear models, lasso and ridge regressions (combinations of L^1 and L^2 penalty) and earth, which uses data adaptive regression splines. For boosting trees (xgboost), we used depth 2 trees to allow interaction as well as depth 1 trees with main terms and interactions as the covariates. Also for boosting, we used different hyperparameters for the number of trees in combination with different learning rates. We used recursive partitioning and random forest, which are tree methods and therefore account for interactions, as well as neural networks which are more non-parametric approaches. We also applied k nearest neighbors for prediction. In addition we applied screening of the top 25 correlated variables with the outcome in conjunction with then running the machine learning algorithm and did the same for the top 5 variables when running bayes generalized linear models as well as glm. The convex combination of learners, or SuperLearner, had the lowest cross-validated risk (negative log-likelihood) of 0.536 with random forest algorithms and the lasso with all interactions included as variables performing equally well. The discrete SuperLearner (Polley et al. 2017), or the one that chooses the lowest risk algorithm to use for each fold's validation set predictions, had a cross-validated risk of 0.57. Without knowing which algorithm would perform the best a priori, we can see SuperLearner did the job it was supposed to do in combining our ensemble in an optimal way, according to cross-validated risk.

1.3.6 Discussion

We can see there are two great challenges in estimating VTE, one being the fact the parameter is bounded below at 0, skewing and biasing the estimates when the true variance is too small for the sample size. In the future we might develop an improvement over log-scaling to form adjusted confidence bounds if they stray into the negative zone. To our eyes, this problem is not as crucial as obtaining reliable inference when the true variance is large enough to be estimated for the sample size. On this front the second order remainder term,

$-\mathbb{E}_0\left(b_0(W)-b(W)\right)^2$, has proven difficult to contain in finite samples. We are certain for larger samples we can show that a Superlearner library including the highly adaptive lasso (Benkeser and van der Laan 2016) will achieve the necessary rates of convergence for this term to be truly second order, however, such is not trivial to show conclusively via very time-consuming and computer intensive simulations and besides, a sample size of 1000 for 4 covariates and treatment is certainly of practical importance. To this end, the authors plan to propose in a follow-up paper, a novel way to estimate the TE function that will also yield estimates of the outcome predictions that are necessarily between 0 and 1, all the while performing asymptotically as good as just fitting the outcome model. We need TE function estimates, $b(W)$, that are compatible with the outcome predictions in order to perform the crucial targeting step of the TMLE procedure. i.e., we need $b(W)=\bar{Q}(1,W)-\bar{Q}(0,W)$.

Another approach to yielding better coverage would be to account for the second order remainder term via the use of the new highly adaptive lasso (HAL) non-parametric bootstrap (van der Laan 2017), to form our confidence intervals. HAL, as mentioned previously in this paper, guarantees the necessary rates of convergence of the second order remainder terms under very weak conditions and is thus guaranteed to yield asymptotically efficient TMLE estimates (van der Laan and Gruber 2016), using the empirical variance of the efficient influence curve approximation for the standard error. However, in finite samples such a procedure yields lower than nominal coverage due to the unforgiving second order remainder term of non-doubly robust estimators as we have for VTE. We feel the HAL CV-TMLE, using a non-parametric bootstrap addresses this issue and accounts for the second order bias, the subject of more future work. Perhaps best will be performing the novel TE fitting procedure with the HAL bootstrap.

1.4 TE CDF

$$\Psi(P)=\mathbb{E}_P\mathbb{I}(b(W)\leq t) \text{ for P}\in\mathcal{M}$$

Ψ is not pathwise differentiable (van der Vaart 2000) so instead we consider the smoothed version of the parameter mapping, using kernel, k, with bandwidth, δ, which is pathwise differentiable, as a strategy to obtain inference for both the smoothed parameter of the TE CDF itself. Here we will suppress k in the notation for convenience:

$$\Psi_{\delta,t}(P)=\mathbb{E}_w\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\mathbb{I}(b(W)\leq x)dx=\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)F(x)dx$$

NOTE: We assume throughout this paper, $Pr(b(W)=x)=0$ for all values, x. In other words, our TE distribution function is continuous.

A Brief Note on Pathwise Differentiability

Taken from van der Vat, 2000: A parameter, Ψ, is pathwise differentiable relative to the tangent space of P, if for every submodel P_t with score function, g, in the tangent space, there exists a continuous linear map from $\dot{\Psi}_P : L^2(P) \rightarrow \mathbb{R}^k$ such that as t vanishes

$$\frac{\Psi(P_t) - \Psi(P)}{t} \longrightarrow \dot{\Psi}_P(g)$$

A classic case of a non pathwise differentiable parameter is the density for a continuous distribution (in the absence of any parametric assumptions) at a point which, depends on a set of measure 0. In our case, the pathwise derivative for the TE CDF at a given value, t, does not exist because the indicator function is not differentiable where it jumps at the value of t. Many of the TE values far away from t will not be very helpful in estimating the CDF at t with much precision, so we focus on a bandwidth of TE values around t in a similar manner to a kernel density estimator. As n becomes larger we want to decrease the bandwidth so as to minimize the mean squared error.

The pathwise derivative defined above has a representation as $\int gD^*(P)dP$, where $D^*(P)$ is a unique element of the tangent space called the efficient influence curve or canonical gradient, whose variance is the cramer-rao lower bound (minimum variance possible) for any regular asymptotically linear estimator of the parameter. Knowing $D^*(P)$ enables the construction of estimators for non-parametric and semi-parametric models, that are asymptotically efficient in that asymptotically their variance attains the cramer-rao lower bound. Examples of such estimators are the one-step estimator and targeted maximum likelihood estimator (TMLE) or its cross-validated counterpart, CV-TMLE (van der Laan and Daniel Rubin 2006; Zheng and van der Laan 2010; van der Laan and Rose 2011). We prefer the CV-TMLE and TMLE, because they have the advantage of being substitution estimators and, therefore, obey natural parameter bounds which, has been shown to improve stability in finite samples (van der Laan and Rose 2011). For our case, if we plug in a model for many points on the TE CDF, we will be guaranteed that the estimates with be both monotonic and bounded within [0,1], where a non-substitution estimator holds no such guarantees. As we will see, CV-TMLE only requires one condition for asymptotic linearity as opposed to two for the TMLE and thus, it is our preferred estimator here.

1.4.1 The Cross-Validated Targeted Maximum Likelihood Estimator, CV-TMLE

Scaled continuous outcomes

Referring to section 1.2.1, the scaling of continuous outcomes changes nothing of importance because when we evaluate our parameter on the original scale we are smoothing the TE CDF $\mathbb{E}(b(W) \leq t) = \mathbb{E}(b(W)/(M-m) \leq t/(M-m))$, the parameter mapping for scaled outcomes.

Here, we will construct a clfm-based TMLE (see section 3.4). As discussed in section 1.3.1, we noticed no appreciable difference in performance for the three basic options in constructing a TMLE, however, the clfm-based TMLE here is considerably faster than the one-step TMLE. Both clfm TMLE and one-step TMLE both employ a one-dimensional submodel, which might prove useful in dimension reduction for high dimensional parameter. To construct our TMLE we need to know the efficient influence curve of our parameter of interest (van der

Vaart 2000) which, is a d-dimensional curve (for each of the d components of the parameter mapping), where the i^{th} component is given by

$$\mathbf{D}^\star_{\boldsymbol{\Psi}_{\delta,\mathbf{t_i}}}(\mathbf{P_0})(\mathbf{O})= \quad \frac{-1}{\delta}\mathbf{k}\left(\frac{\mathbf{b_0(W)-t_i}}{\delta}\right)*\frac{\mathbf{2A-1}}{\mathbf{g_0(A|W)}}(\mathbf{Y}-\bar{\mathbf{Q}}_\mathbf{0}(\mathbf{A},\mathbf{W}))+\int\frac{\mathbf{1}}{\delta}\mathbf{k}\left(\frac{\mathbf{x-t_i}}{\delta}\right)\mathbb{I}(\mathbf{b_0(W)\leq x})\mathbf{dx}-\boldsymbol{\Psi}_{\delta,\mathbf{t_i}}(\mathbf{P_0})$$

where t_i is a given TE value (average treatment effect level), k is the kernel and bandwidth is δ. The reader may find the proof in theorem 3.1.8, section 3. From here we will shorten the notation and refer to D^* as the $d-$dimensional efficient influence curve with components $D^\star_i = D_{\Psi_{\delta,t_i}}$. We define a clfm as follows from Levy, 2018c:

Definition 1.4.1. A canonical 1-dimensional locally least favorable submodel (clfm) of an estimate, P_n^0, of the true distribution, P_0 is

$$\{P^0_{n,\epsilon} \text{ s.t } \frac{d}{d\epsilon}P_nL(P^0_{n,\epsilon})\bigg|_{\epsilon=0} = \|P_nD^\star(P_n^0)\|_2, \epsilon\in[-\delta,\delta]\} \tag{1.8}$$

where $P^0_{n,\epsilon} = P^0_n$ and $\|\cdot\|_2$ is the euclidean norm.

We remind the reader, the initial estimate, P_n^0, of P_0 is defined by $\bar{Q}_n^0(A,W)$ an estimate of the outcome regression, $\bar{Q}_0$, g_n, an estimate of the treatment mechanism, g_0, and $Q_{W,n}$, the empirical distribution, which estimates $Q_{W,0}$, the distribution of W. We denote the empirical density as $q_{W,n}$, which esimtates the true density, $q_{W,0}$, of W. We can then define the d-dimensional curve

$$\begin{aligned} H^0(A,W) &= (H^0_1(A,W), H^0_2(A,W), ..., H^0_d(A,W)) \\ &= \frac{1-2A}{\delta g_n(A|W)}\left(k\left(\frac{b^0_n(w)-t_1}{\delta}\right), k\left(\frac{b^0_n(w)-t_2}{\delta}\right), ..., k\left(\frac{b^0_n(w)-t_d}{\delta}\right)\right)\end{aligned}$$

where $b^0_n(W_i) = \bar{Q}^0_n(1,W) - \bar{Q}^0_n(0,W)$. The initial empirical risk for the outcome model is given by

$$P_nL(\bar{Q}^0_n) = -\frac{1}{n}\sum_{i=1}^n\left[Y_i\text{log}\bar{Q}^0_n(A_i,W_i) + (1-Y_i)\text{log}(1-\bar{Q}^0_n(A_i,W_i))\right]$$

Our efficient influence curve approximation at the initial estimate is given by $D^*(P_n^0)$. Now define the elements of the clfm of initial estimate, P_n^0, by keeping g_n and $q_{W,n}$ fixed and defining

$$\bar{Q}^0_{n,\epsilon}(A,W) = expit\left(logit(\bar{Q}^0_n(A,W)) + \epsilon\left\langle H^0(A,W), \frac{P_nD^*(P_n^0)}{\|P_nD^*(P_n^0)\|_2}\right\rangle_2\right)$$

We can then verify this satisfies the definition of clfm above when we use quasibinomial loss for a scaled continuous outcome or negative log-likelihood loss for a binary outcome. This then gives rise to the iterative procedure detailed below in steps 1 through 4 below, which will cover both CV-TML and TML estimators. The CV-TMLE will be the same algorithm as that by Zheng and van der Laan, 2010 when using a pooled regression to fit the fluctuation parameter (the ϵ_k's below).

1.4.2 TML Algorithm

step 1: Obtaining Initial Estimates We obtain initial estimates of the data generating distribution identically to section **??**.
step 2:
Starting with $m = 0$:
If $|P_n D_j^*(P_n^m)(O)| < log(n)\hat{\sigma}(D_j^*(P_n^m)(O))/n^{1/2}$ for all j then $P_n^\star = P_n^m$ and go to step 4. Otherwise go to step 3. $\hat{\sigma}(\cdot)$ refers to the sample standard deviation of values taken over the data. To provide some clarity: If $n = 1000$ then $log(n) \approx 7$, so the above stopping criterion ensures any bias is second order at that point. More iterations after this are only time-consuming and do not help with coverage to any appreciable extent.

step 3:
Y as the outcome, offset $= logit(\bar{Q}_n^m)(A, W)$ and so-called clever covariate is computed as

$$\left\langle (H^{m-1}(A, W), \frac{P_n D^*(P_n^m)}{\|P_n D^*(P_n^m)\|_2} \right\rangle_2.$$

where $\langle \cdot, \cdot \rangle_2$ is the dot-product or euclidean inner product. Assume ϵ_m is the coefficient computed from the logistic regression defined by

$$\bar{Q}_n^{m+1}(A, W) = expit\left(logit\left(\bar{Q}_n^m(A, W)\right) + \epsilon_n^m H^m(A, W)\right)$$

We then update the models by the following:

$$\bar{Q}_n^{m+1}(A, W) = expit\left(logit(\bar{Q}_n^m(A, W)) - \epsilon_n^m \left\langle (H_1(P_n^m)(A, W), \frac{P_n D^*(P_n^m)}{\|P_n D^*(P_n^m)\|_2} \right\rangle_2 \right) \quad (1.9)$$

set $m = m + 1$ return to step 2.

step 4:
The TMLE procedure yields $\bar{Q}_n^*(A, W)$ and our estimator is then a plug-in estimator, with j^{th} component:

$$\Psi_{\delta, t_j}(P_n^\star) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\delta}\int k\left(\frac{x - t_j}{\delta}\right)\mathbb{I}(b(W_i) \leq x)dx$$

and standard errors are given by

$$\frac{\widehat{\sigma}_n(D_j^*(P_n^*))}{\sqrt{n}}$$

where $\widehat{\sigma}_n(D_j^*(P_n^*))$ is the sample standard deviation of $\{D_j^*(P_n^*)(O_i) \mid i \in 1 : n\}$ and $b(W_i) = \bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)$, the TE function estimate.

Performing a CV-TMLE

To perform a CV-TMLE, we apply the same procedure as in section 1.3.1 and continue with steps 2 through 4.

Simultaneous Estimation and Confidence Bounds

Simultaneous confidence intervals are obtained in the identical way as for (ATE, VTE), discussed in section 1.3.1

1.5 TMLE conditions for Estimating $\Psi_\delta(P_0)$

importance of the TMLE mapping is we then have

Theorem 1.5.1.

$$\Psi_{\delta,t_j}(P_n^\star) - \Psi_{\delta,t_j}(P_0) = (P_n^0 - P_0)D_j^*(P_n^\star) + R_{j,2}(P_n^\star, P_0)$$

where $R_2(P_n^*, P_0)$ *is given by*

$$\frac{-1}{\delta}\int\left[k\left(\frac{b_n^*(w)-t}{\delta}\right)\left(\left(\frac{g_0(1|w)}{g_n(1|W)}-1\right)(\bar{Q}_0(1,w)-\bar{Q}_n^*(1,w))-\left(\frac{g_0(0|w)}{g_n(0|w)}-1\right)(\bar{Q}_0(0,w)-\bar{Q}_n^*(0,w))\right)\right]dQ_{W,0}(w)$$
$$+\frac{1}{\delta}\int\left[\int_{b(w)}^{b_0(w)}k\left(\frac{x-t}{\delta}\right)dx+k\left(\frac{b(w)-t}{\delta}\right)(b(w)-b_0(w))\right]dQ_{W,0}(w)$$

Our remainder term can be bounded is as follows:

$$\begin{aligned} R_{2,i}(P_n^0, P_0) &= \frac{1}{\delta}O\left(\|g_n - g_0\|_{L^2_{P_0}}\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2_{P_0}}\right) + \frac{1}{\delta}O\left(\|b_n^0 - b_0\|_\infty^2\right) \\ \text{or } &\frac{1}{\delta}O\left(\|g_n^0 - g_0\|_{L^2_{P_0}}\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2_{P_0}}\right) + \frac{1}{\delta^2}O\left(\|b_n^0 - b_0\|_{L^2_{P_0}}^2\right) \end{aligned}$$

The reader may see section 3.2.3 for the proof.

The Use of Highly Adaptive Lasso

When using the highly adaptive lasso (HAL) (van der Laan 2016; van der Laan and Gruber 2016) to perform the initial estimates, we are guaranteed $\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2(P_0)}$ and $\|g_n - g_0\|_{L^2(P_0)}$ are $o_P(n^{-0.25})$ under the conditions that $\bar{Q}_0$ and g_0 are of bounded sectional variation norm and continuous from the right with left-hand limits. The use of HAL along with the previous theorem and Theorems 1.2.1 and 1.2.2 yield the following corollary:

Corollary. *When using HAL to form initial estimates of* $\bar{Q}_0$ *and* g_0*, the TML estimator of* $\Psi_{\delta,t_i}(P_0)$ *(fixed bandwidth,* δ*) will be asymptotically efficient.*

The reader may note that this parameter does not allow for doubly robust estimation.

1.5.1 Allowing the Bandwidth, δ, to Vanish for n Large

The reader may notice that below we bound the remainder term in two different ways in Theorem 1.5.1, one of which has δ in the denominator and the other which has δ^2 in the denominator. If we let δ approach 0 as a function of n, then we would prefer to only have δ in the denominator so as to allow δ to approach 0 faster and hence, a lower mean squared error. However, that condition is more difficult to guarantee as we will point out.

We will refer to the following facts, where P_n^0 is an initial fit of P_0 and P_n^* is a TMLE update of P_n^0.

1. The asymptotic variance of $\sqrt{n}(\Psi_{\delta,t_i}(P_n^*) - \Psi_{\delta,t_i}(P_0))$ is of order $1/\delta$. See the proof of this fact in theorem 3.2.4.
2. The bias between unsmoothed TE CDF value at t_i and the smoothed parameter, $\Psi_{t_i}(P_0) - \Psi_{\delta,t_i}(P_0)$, is of order δ^J, where J is the order of the kernel (power of the kernel's first non-zero moment) and we assume the TE CDF to have J continuous derivatives. The reader may see the proof of this fact in theorem 3.2.5 in the Appendix.

Theorem 1.5.2. *Assume the TE CDF has J continuous derivatives. Assume we allow our bandwidth* $= \delta_n = O(n^{-1/(2J+1)})$. *Referring to corollary 1.5, then if* $r_{g_n} + r_{\bar{Q}_n} \leq \frac{J+1}{2(2J+1)}$ *and either of*

- *A1:* $\|\bar{Q}_n^0 - \bar{Q}_0\|_\infty = o_P\left(-\frac{J+1}{2(2J+1)}\right)$
- *A2:* $\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2(P_0)} = o_P\left(-\frac{2J+3}{4(2J+1)}\right)$

$\sqrt{\delta_n n} R_2(P_n^0, P_0) \xrightarrow{p} 0$

This statement follows immediately from Theorem 1.5.1 at the beginning of this section.

Theorem 1.5.3. *If using bandwidth of order* $\delta_n = O(n^{-1/(2J+1)})$ *and HAL to form initial predictions then if we use a kernel of order* $J > \frac{4r+3}{2}$ *and the TE CDF has J continuous derivatives,* $\sqrt{\delta_n n} R_2(P_n^0, P_0) \xrightarrow{p} 0$.

The statement follows from the fact HAL guarantees $\|f_0 - f_n^0\|_{L^2(P_0)} = O_P(n^{-1/4-1/8(r+1)})$, when fitting a function, f_0 of finite sectional variation norm that is continuous from the right with left-hand limits (van der Laan 2016).

Remark. The motive for this theoerm is that if we wanted to minimize the MSE based on items 1 and 2 above (theorems 3.2.5 and 3.2.4 in the appendix) as for a kernel density estimator, we would want $\delta_n = O(n^{-1/(2J+1)})$. However, we also want the remainder term to become truly second order when blown up by $\sqrt{\delta_n n}$ in order for $\sqrt{\delta_n n}(\Psi_{\delta_n,t_i}(P_n^*) - \Psi_{\delta_n,t_i}(P_0))$ to have a limiting distribution. Thus, perhaps higher order kernels can be useful in relaxing the requirements of theorem 1.5.2 in fitting the treatment mechanism and, especially, the outcome model. For fitting nuisance parameters that are functions of variables of dimension 5, we would need a kernel of order 12 or greater to guarantee theorem 1.5.1. If HAL were to guarantee $\|f_0 - f_n^0\|_{L^\infty} = O_P(n^{-1/4-1/8(r+1)})$ then using a kernel of order $J > \frac{2r+1}{2}$ and assuming necessary smoothness on the TE CDF, would guarantee $\sqrt{\delta_n n} R_2(P_n^0, P_0) \xrightarrow{p} 0$. Thus, if $r = 5$, we only require a kernel of order 6 and hence, only 6 continuous derivatives for the TE CDF.

The reader may consult Appendix 3.2.2 to see how the polynomial kernels used in this paper were constructed. These kernels enabled very fast estimation because they could be integrated exactly in closed form and one would need to compute as many integrals as the sample size for each iteration of the CV-TMLE algorithm.

1.5.2 Simulations for Fixed Bandwidth and When Using Bandwidth Selection

Well-specified Models

For well-specified logistic models where the data generating system is given by the following: W is a random normal, $Pr(A = 1 \mid W) = g(A \mid W) = expit(.2 + .2 * W)$ and $E[Y \mid A, W] = expit(A + 2.5 * A * W + W)$. The TMLE's using the MLE as an initial estimate performed very well, with normal sampling distributions, nominal coverage (93% or higher) of the smoothed parameter, as expected, and did so for all kernels if we used bandwidth $n^{-1/(2J+1)}$ where J is the order of the kernel we and let n attain values of 1000, 2500, 5000, 10000, 25000 and 50000. The MSE was lowest for the well-specified MLE plug-in, also as expected, but not appreciably. In the highly unlikely scenario that we correctly specify the outcome model with a parametric form, TMLE performance appears very reliable for covering the smoothed parameter and yields vanishing standard errors as sample size grows.

A Method for Choosing Bandwidth for a Given Kernel

We would like to form confidence bounds for the non-pathwise differentiable parameter or unsmoothed "true" parameter, $\Psi(P) = \mathbb{E}_P \mathbb{I}(b(W) \leq t)$ for $P \in \mathcal{M}$, and propose using some of the concepts in Chapter 25 of Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies by van der Laan and Rose, 2018. We start with a largest bandwidth of size $n^{-1/(2J+1)}$ where J is the order of the kernel. Then we divide the bandwidth into 20 equal increments from $n^{-1/(2J+1)}/20, 2n^{-1/(2J+1)}/20, ..., n^{-1/(2J+1)}$. We then find the smallest set of 5 or more consecutive bandwidths that are monotonic estimates with respect to the bandwidth. If no such 5 or more consecutive bandwidths are found then we choose the bandwidth $n^{-1/(2J+1)}$. Let us call the consecutive bandwidth sequence, $B_c = \{h_1, ..., h_c\}$, where h_1 is the smallest. We also monotonize the variance so as to force it to be increasing as the bandwidth gets smaller. We then form confidence intervals using the monotonized variance for each bandwidth in B_c. If the sequence of estimates is decreasing (increasing) as bandwidth decreases (for bandwidths in B_c), then we choose the confidence interval with the minimum (maximum) right (left) bound. The idea is that we are minimizing the MSE while maintaining nominal coverage, assuming that the smoothed parameters are monotonic for the bandwidths in B_c and that this monotonicity represents the monotonicity as the bandwidth approaches 0. We still need to refine the theory as our increments for the bandwidth (20 in this case) and definition of being monotonic (5 consecutive or more as described above) are somewhat arbitrary. On the positive side, we noticed coverage of the smoothed parameter maintained nominal levels as n grew to 50,000 when applying our bandwidth selector. Similar results held for lesser order kernels as well. Figure 1.6 below displays the heuristic behind our bandwidth selector.

Table 1.3: coverage of smoothed parameter, kernel is order 10

	n = 1000		n = 2500		n = 5000		n = 10000		n = 25000		n = 50000	
TE	meth	fixed	meth	fixed	meth	fixed	meth	fixed	meth	fixed	meth	fixed
−0.145	0.907	0.947	0.920	0.949	0.916	0.941	0.935	0.948	0.944	0.949	0.944	0.948
−0.085	0.911	0.953	0.950	0.946	0.939	0.934	0.958	0.962	0.942	0.947	0.955	0.950
−0.025	0.925	0.944	0.950	0.960	0.958	0.948	0.949	0.948	0.947	0.941	0.948	0.945
0.035	0.916	0.940	0.929	0.949	0.942	0.966	0.949	0.959	0.952	0.954	0.939	0.937
0.095	0.934	0.951	0.934	0.949	0.946	0.942	0.943	0.943	0.944	0.948	0.944	0.947
0.155	0.933	0.952	0.942	0.946	0.936	0.948	0.944	0.952	0.942	0.946	0.948	0.942
0.215	0.927	0.958	0.927	0.951	0.932	0.941	0.934	0.942	0.953	0.954	0.951	0.939
0.275	0.893	0.955	0.913	0.955	0.905	0.951	0.914	0.935	0.926	0.942	0.938	0.949

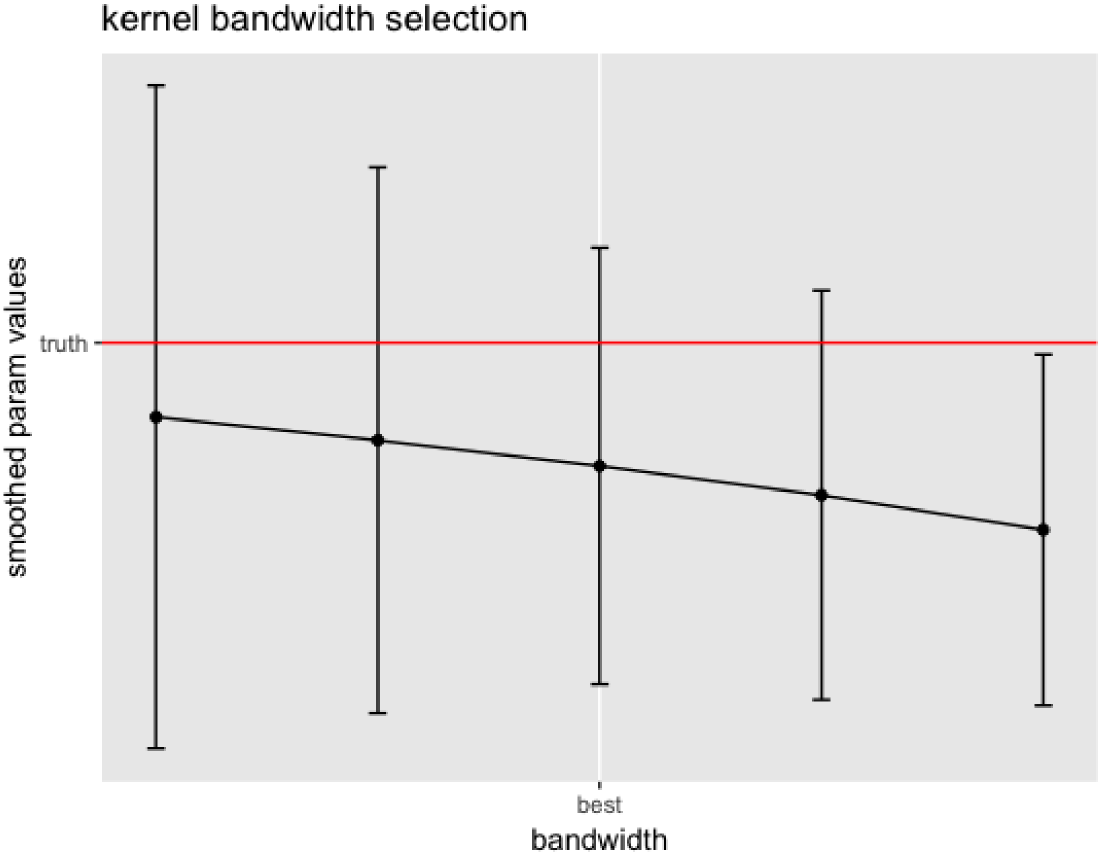
meth means we applied the bandwidth selection method, fixed means we used bandwidth $n^{-1/(2J+1)}$ where J is the kernel order.

Table 1.4: coverage for true parameter, kernel is order 10

	n = 1000		n = 2500		n = 5000		n = 10000		n = 25000		n = 50000	
TE	meth	fixed	meth	fixed	meth	fixed	meth	fixed	meth	fixed	meth	fixed
−0.145	0.671	0.001	0.436	0	0.298	0	0.294	0	0.338	0	0.325	0
−0.085	0.612	0.136	0.593	0.024	0.743	0.001	0.836	0	0.870	0	0.878	0
−0.025	0.770	0.166	0.615	0.019	0.405	0.001	0.207	0	0.076	0	0.032	0
0.035	0.850	0.071	0.924	0.001	0.927	0	0.938	0	0.903	0	0.830	0
0.095	0.747	0.070	0.895	0	0.912	0	0.924	0	0.906	0	0.801	0
0.155	0.750	0.251	0.859	0.020	0.903	0	0.907	0	0.911	0	0.945	0
0.215	0.695	0.947	0.717	0.861	0.793	0.692	0.855	0.370	0.867	0.009	0.877	0
0.275	0.858	0.008	0.817	0	0.707	0	0.493	0	0.147	0	0.010	0

meth means we applied the bandwidth selection method, fixed means we used bandwidth $n^{-1/(2J+1)}$ where J is the kernel order.

Figure 1.6



1.5.3 Simulations for Misspecified Models

We call these simulations ”misspecified” because we use the highly adaptive lasso or HAL (van der Laan and Gruber 2016) to recover the model without any specification on functional forms. The data generating system consisted of the following functions in the order listed. W is a random normal, $Pr(A = 1 \mid W) = g(A \mid W) = expit(-.1 - .5 * sin(W) - .4 * (|W| > 1) * W^2)$ and $E[Y \mid A, W] = expit(.3 * A + 5 * A * sin(W)^2 - A * cos(W))$. We simulated 1100 draws from the above data generating system and computed simultaneous TMLE's for the TE values -0.098, -0.018 0.062, 0.142, 0.222, 0.302, 0.382 and 0.462 using bandwidth $2500^{-0.2}$ and an order 1 polynomial kernel. Similar results held for the uniform kernel.

Here we show the huge advantage of data adaptive estimation in obtaining the initial estimates for CV-TMLE, using the highly adaptive lasso. TMLE_glm used used logistic regression with main terms and interactions for the initial estimates in CV-TMLE, while TMLE_HAL used HAL for the initial estimates. We can see it is catastrophic to use logistic regression here while using HAL with TMLE procedure achieved very close to nominal coverage with essentially no bias (see table 1.5). Targeting helped remove bias from the HAL initial estimates as well, shown in Figure 1.7, for one of eight points on the TE CDF simultaneously estimated by TMLE_HAL. The other seven points had very similar sampling distributions and bias.

1.5.4 Software

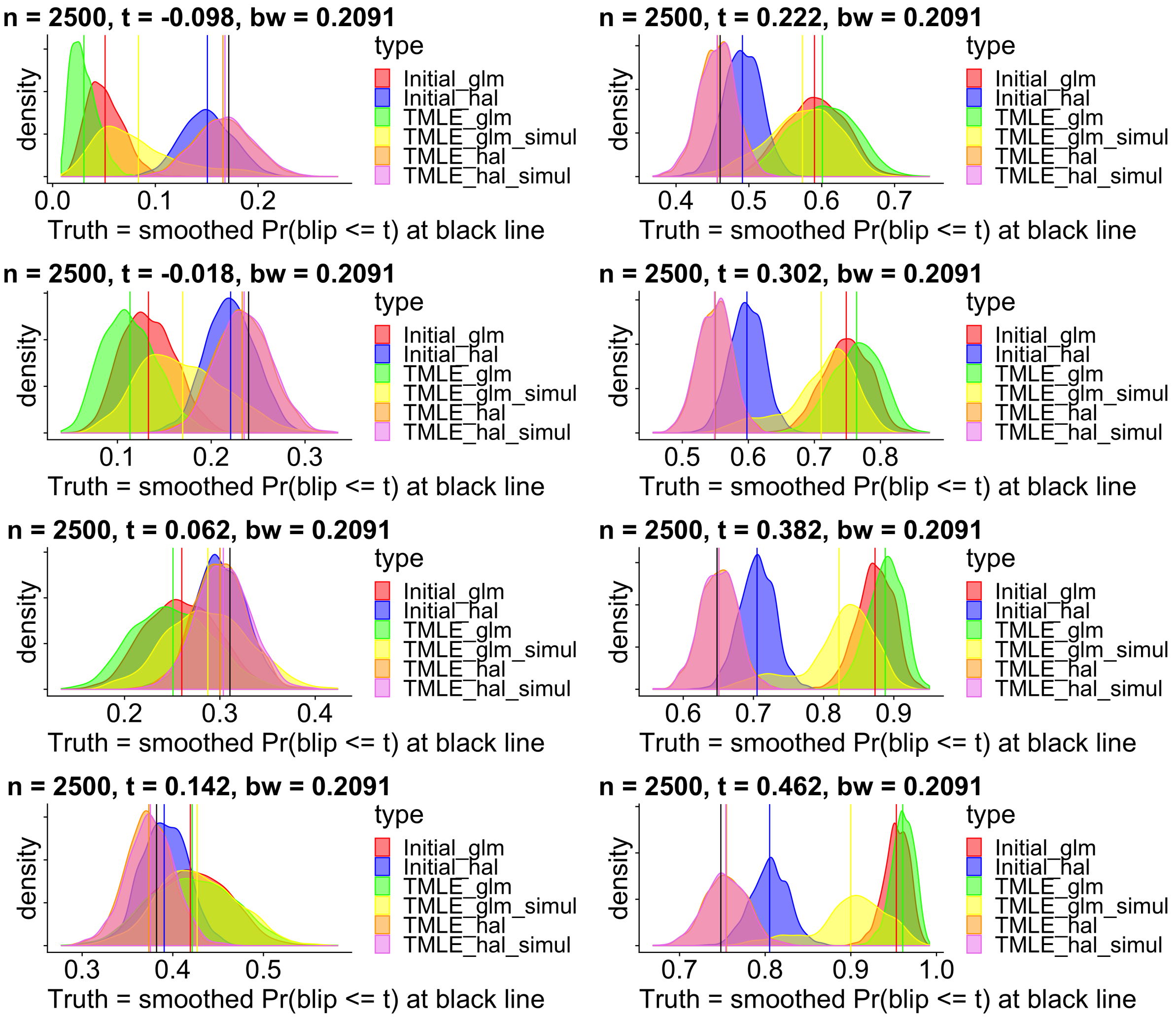
The reader may visit https://github.com/jlstiles/TECDFsim (Levy 2018d) for procedures on how to reproduce the results here-in and also visit https://github.com/jlstiles/TECDF (Levy 2018b) for software on performing the targeting step after obtaining initial estimates. This estimator is also available in the package https://github.com/tlverse (Coyle, Malenica, et al. 2018b), where the reader can also perform ensemble learning.

Table 1.5: TMLE with HAL initial estimates vs glm, coverage of smoothed parameter

	MSE TMLE_hal	MSE TMLE_glm	coverage TMLE_hal	coverage TMLE_glm
TE = -0.098	0.00083	0.02003	0.91727	0
TE = -0.018	0.00089	0.01683	0.92545	0.01818
TE = 0.062	0.00087	0.00528	0.93727	0.58455
TE = 0.142	0.00071	0.00373	0.94909	0.81000
TE = 0.222	0.00061	0.02173	0.96182	0.10455
TE = 0.302	0.00065	0.04723	0.95182	0
TE = 0.382	0.00069	0.05803	0.94455	0
TE = 0.462	0.00067	0.04528	0.94091	0

simultaneous TMLE_hal coverage was 90%, TMLE_glm coverage was 3%

Figure 1.7: Smoothed TE CDF Sampling Distributions



1.6 Discussion

We have developed an estimator to efficiently estimate, under conditions, the kernel smoothed version of the TE CDF and also allow the bandwidth to approach zero and guarantee a normal limiting distribution for the TE CDF itself. Furthermore, our estimator does not rely on any parametric assumptions on the data generating distribution. We have shown our estimator hinges on data adaptive estimation, particularly the use of the highly adaptive lasso, to make our initial estimates in the targeted learning (van der Laan and Rose 2011) framework. The TML update helps eliminate bias and provides us with immediate inference for the smoothed parameter via the sample standard deviation of the efficient influence curve approximation. Our simulations have shown that for well-specified models, choosing the bandwidth of optimal order $n^{-\frac{1}{2J+1}}$ (and hence a vanishing bandwidth in n), assuming J continuous derivatives for the TE CDF, provides normal and unbiased sampling distributions for the smoothed parameter.

The next step is to develop a way to optimally (smallest MSE possible) select the bandwidth, δ_n, and kernel so that the estimator minus the truth blown up by $\sqrt{n\delta_n}$ is normally distributed and covers the TE CDF nominally. Our bandwidth selector in this paper still gives nominal or near-nominal coverage of the smoothed parameter as the bandwidth vanishes for large n, but is not yet reliable for covering the TE CDF itself, though we show it is a big improvement over setting the bandwidth to $n^{-\frac{1}{2J+1}}$. Our bandwidth selector relies on the assumption that the smoothed parameter is monotonically increasing or decreasing toward the unsmoothed parameter as the bandwidth vanishes. Our method of determining this monotonicity is somewhat arbitrary and it also remains to be seen how this monotonicity generally holds for small bandwidths. For instance, if the monotonicity changes direction for a small bandwidth, our proposed bandwidth selector might be problematic.

Transporting Stochastic Direct and Indirect Effects

2.1 Motivation

Often, an intervention, program, or policy that works in one place or population fails to replicate in another place or population (Rudolph, Schmidt, et al. 2018) or can even have unintended harmful effects (Kling, Liebman, and Katz 2007). This is problematic from a public policy or public health perspective in that the goals of such interventions are to help—not harm, and problematic from a financial perspective in that limited resources may be not be spent optimally.

When such initiatives fail to replicate or have unintended effects in new populations, transportability theory and methods offer a chance to understand why. Transportability is the ability (based on identifying assumptions) to transport a causal effect from a source population to a new, target population, accounting for differences between the two populations (e.g., differences in compositional factors, treatment adherence, etc.) (Pearl and Bareinboim 2014). Previous work developed estimators to transport total effects from a source to target population (Rudolph and van der Laan 2017) or, similarly, to generalize effects from a sample to the population (Miettinen 1972; Stuart et al. 2011; Cole and Stuart 2010; Frangakis 2009).

In some cases, examining transportability of the total effect may shed light on reasons for lack of replication. However, in other cases, transporting the total effect may not identify the relevant differences and it may be beneficial to go further and examine transportability of the underlying mediation mechanisms. Although there has been work on the identification on transported indirect effects (Bareinboim and Pearl n.d.; Pearl and Bareinboim 2014), we are not aware of any previous work developing estimators for transporting mediation effects (direct and indirect effects) from a source to target population. Thus, we address this research gap by proposing several different estimators of stochastic direct and indirect effects: a simple inverse-probability of treatment weighted estimator, a doubly robust estimator that solves the estimating equation, and a doubly robust, efficient substitution estimator in the targeted minimum loss-based framework.

2.2 Data and Model

The full data comes from a structural causal model (SCM) (Pearl 2000) to which we would like to have access in order to find causal parameters of interest. We can consider a draw from unknown measurements, $U = (U_S, U_W, U_A, U_Z, U_M, U_Y) \sim P_U$. and then the generation of variables in the following time ordering.

$$\begin{aligned}
S &= f_S(U_S) \\
W &= f_W(U_W, S) \\
A &= f_A(U_A, W, S) \text{ in this case, just } f_A(U_A) \\
Z &= f_Z(U_Z, A, W, S) \\
M &= f_M(U_M, Z, W, S) \\
Y &= f_Y(U_Y, Z, W, M)
\end{aligned}$$

Setting $O = (Y, M, Z, A, W, S)$, we may write $(U, O) \sim P_{UO} \in \mathcal{M}^F$, the full-data model. As in the previous paper by Rudolph et. al. on stochastic direct and indirect effects (Rudolph, Sofrygin, Schmidt, et al. 2017) we consider A to be a randomly assigned instrument which has no arrow in the directed acyclic graph (Pearl 1995) to Y or M. S indicates the site, Y is an outcome (either continuous or binary), M is a mediator, Z is a confounder on the causal pathway from treatment, A, to M and W is a vector of confounders. The observed data is a random variable $(S \times YS, M, Z, A, W, S) \in \mathcal{M}$, the observed data model. We can see the observed data model is a subset of the full-data model in that we observe (S, W, A, Z, M) directly out of the full-data but Y is only observed out of the full-data when $S = 1$. Our full data model and observed data model are both semi-parametric due to the aforementioned restrictions but we can also perform the same analysis here-in under non-parametric models, i.e., as in an observational study where we can allow Y and M to be functions of A as well as their preceding variables 3.1.2. We will refer to the semi-parametric model as $\mathcal{M}_I$ and the non-parametric model as $\mathcal{M}_{II}$.

2.3 Parameter of Interest

In (Rudolf et al.) the authors defined a stochastic intervention parameter for a model not including S, where the intervention assigns M according to probability defined by

$$\hat{g}_{M|a^*,W}(M \mid W) = \sum_z Pr(M \mid Z = z, W) Pr(Z = z \mid A = a^*, W)$$

The parameter of interest was defined as $\Psi(P_{UX}) = \mathbb{E}\left[Y_{a,\hat{g}_{M|a^*,W}}\right]$ where the expectation is taken over the full data model and $Y_{a,\hat{g}_{M|a^*,W}}$ is the outcome under the stochastic intervention after having assigned treatment, a. We wish to transport this parameter to a new site where the outcome was not observed ($S = 0$):

$$\Psi^F(P_{UX}) = \mathbb{E}\left[Y_{a,\hat{g}_{M|a^*,W,s}} \mid S = 0\right]$$

where

$$\hat{g}_{M|a^*,W,s}(M \mid W) = \sum_z Pr(M \mid Z = z, W, S = s) Pr(Z = z \mid A = a^*, W, S = s)$$

where s is either 1 or 0, depending on how one wants to define the stochastic intervention based on the observed data. The site need not be involved at all in the definition of $\hat{g}$ but here we will assume we have defined $\hat{g}$ based on the data for one of the sites and hence we will place an s in the subscript of the name of the function to indicate such.
Our parameters of interest are the stochastic direct effect:

$$\Psi^F_{SDE}(P_{UX}) = \mathbb{E}\left[Y_{a=1,\hat{g}_{M|a^*=0,W,s}} \mid S = 0\right] - \mathbb{E}\left[Y_{a=0,\hat{g}_{M|a^*=0,W,s}} \mid S = 0\right]$$

Our parameters of interest are the stochastic indirect effect:

$$\Psi^F_{SIE}(P_{UX}) = \mathbb{E}\left[Y_{a=1,\hat{g}_{M|a^*=1,W,s}} \mid S = 0\right] - \mathbb{E}\left[Y_{a=1,\hat{g}_{M|a^*=0,W,s}} \mid S = 0\right]$$

2.4 Identifiability

In order to identify the parameter of interest we will need to impose additional non-testable assumptions on $\mathcal{M}^F$ and $\mathcal{M}$, listed below.

1. Positivity: For all S and W we need a positive probability of assigning any level of treatment, A=a. For all combos of S, W and $A = a$ we need to have a positive probability for any level of Z. For $S = 1$ and all combos of Z and W we need a positive probability of any level of the mediator, M.

2. Common model assumption: $E[Y \mid M, Z, W, S = 1] = E[Y \mid M, Z, W, S = 0]$.

3. Sequential Randomization: $Y_{am} \perp A \mid W, S$ and $Y_{am} \perp M \mid W, A = a, Z, S$. We treat the time ordering as per the structural equations of section 1 and therefore consider our situation at hand similar to a two-time point longitudinal intervention where at the first time point, we intervene to set the treatment, $A = a$. Then we impose the stochastic intervention for the mediator, M, which plays role of treatment for what could be considered the second time point. Y_{am} is therefore the potential outcome under intervening on the structural equations, setting treatment to a and then downstream, setting the mediator to m.

Note on notation: Any subscript used is only for descriptive purposes and is not to be considered a variable. For instance, we use a capital letter in p_Y, the conditional density of Y, because it is a density of the random variable Y given the past variables. We use W in the subscript for $\hat{g}_{M|a^*,W,s}$ because it is a conditional density of random variable M given random variable W, and values a^* and s, for which a lower case letter indicates they are fixed and the same for all participants. As arguments, sometimes we wish to use lowercase variables as when using the integral notation and at other times we wish to use uppercase letters when thinking of random variables, as in the expectation notation. We have the following identifiability result.

Theorem 2.4.1.

$$\Psi(P) = \Psi^F(P_{UX}) = \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}_{\hat{g}_{M|a^*,W,s}}\left[\mathbb{E}\left[Y|M,Z,W,S=1\right]|W,Z\right]|W,a,S=0\right]|S=0\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_m \left[\mathbb{E}Y\hat{g}_{M|a^*,W,s}(m \mid W) \mid M=m,Z,A=a,W,S=1\right] \mid A=a,W,S\right] \mid S=0\right]$$

Proof.

$$\Psi^F(P_{UX}) = \mathbb{E}\left[Y_{a,\hat{g}_{M|a^*,W,s}} \mid S=0\right]$$

$$= \sum_m \left[\mathbb{E}Y_{am}\hat{g}_{M|a^*,W,s}(m \mid W)\right] \mid S=0\right]$$

assumption 1, 2 and the tower law $\Longrightarrow$

$$= \mathbb{E}\sum_m \mathbb{E}\left[\mathbb{E}\left[Y_{am}\hat{g}_{M|a^*,W,s}(m \mid W) \mid Z,A,W,S=1\right] \mid A,W,S\right]\right] \mid S=0\right]$$

assumption 3 then allows intervention on A and M $\Longrightarrow$

$$= \mathbb{E}\sum_m \mathbb{E}\left[\mathbb{E}\left[Y_{am}\hat{g}_{M|a^*,W,s}(m \mid W) \mid M=m,Z,A=a,W,S=1\right] \mid A=a,W,S\right]\right] \mid S=0\right]$$

By self consistency we can eliminate ”am” counterfactual subscript $\Longrightarrow$

$$= \mathbb{E}\sum_m \mathbb{E}\left[\mathbb{E}\left[Y\hat{g}_{M|a^*,W,s}(m \mid W) \mid M=m,Z,A=a,W,S=1\right] \mid A=a,W,S\right]\right] \mid S=0\right]$$

reordering integration $\Longrightarrow$

$$= \mathbb{E}\left[\mathbb{E}\sum_m \left[\mathbb{E}\left[Y\hat{g}_{M|a^*,W,s}(m \mid W) \mid M=m,Z,A=a,W,S=1\right] \mid A=a,W,S\right]\right] \mid S=0\right]$$

And we note, the conditional expectations are well-defined due to the assumption 1. □

2.5 Targeted Maximum Likelihood Estimator, TMLE

We now describe how to estimate $\Psi(P) = \mathbb{E}\left[Y_{a,\hat{g}_{M|a^*,W,s}} \mid S=0\right]$ using targeted minimum loss-based estimation (TMLE). This estimation approach uses sequential regression, updating the conditional outcome model at each stage to both solve the empirical mean of the influence curve equal approximation equal $o_P(n^{-0.5})$ (IC equation) while also lowering the empirical negative log-likelihood loss (see section 1.2.1) of the conditional outcome model at each stage. We will be estimating the parts of the distribution needed to estimate the efficient influence curve so that the empirical mean of the influence curve is 0, much in the same way as a two time-point longitudinal intervention (van der Laan and Rose 2011). Thus, the construction of the TMLE depends on the knowledge of the efficient influence curve given below and proven in theorems 3.1.2 and 3.1.3. We recommend all fitting to be performed with an ensemble learning package as described in section 1.3.3.

Define the following function (under model $\mathcal{M}_{II}$):

$$H(O) = \frac{\hat{g}_{M|a^*,W,s}(M \mid W) p_Z\left(Z \mid A=a, W, S=0\right) p_W\left(S=0 \mid W\right) I(S=1, A=a)}{p_M\left(M \mid Z, W, S=1\right) p_Z\left(Z \mid A=a, W, S=1\right) p_A\left(a \mid W, S=1\right) p_W\left(S=1 \mid W\right) P_S(S=0)} \tag{2.10}$$

If we are assuming $\mathcal{M}_I$, then define

$$H(O) = \frac{\hat{g}_{M|a^*,W,s}(M \mid W) p_Z(Z \mid A=a, W, S=0) p_{S|W}(S=0 \mid W) I(S=1)}{g_{M,r}(M \mid Z, W, S) p_Z(Z \mid W, S) p_{S|W}(S \mid W) p_S(S=0)}$$

Efficient Influence Curve

If we were using the smaller model, $\mathcal{M}_I$, we would not include A in the regression formula for either Y or M.

$$\begin{aligned}
&D^*(P) = D^*_Y(P) + D^*_Z(P) + D^*_W(P), \text{ where} \\
&D^*_Y(P) = \left(Y - \bar{Q}_Y(M, Z, A, W)\right) H(O) \\
&D^*_Z(P) = \left(\bar{Q}_M(Z, A, W, S) - \bar{Q}_Z(a, W, S)\right) \frac{I(S=0, A=a)}{p_A\left(a \mid W, S\right) p_S(S=0)}, \text{ and} \\
&D^*_W(P) = \left(\bar{Q}_Z(a, W, S) - \Psi(P)\right) \frac{I(S=0)}{p_S(S=0)} \text{ (notation is explained further below).}
\end{aligned} \tag{2.11}$$

Let $\bar{Q}^0_{Y,n}(M, Z, A, W)$ be an initial estimate of $\mathbb{E}\left[Y \mid M, Z, A, W, S=1\right]$ $\hat{g}_{M|a^*,W,s}(M|W)$ is a data-dependent stochastic intervention on M. One can estimate $\hat{g}_{M|a^*,W,s}(M|W) = \sum_{z=0}^{1} P(M=1|Z=z, W, S=s) P(Z=z|A=a^*, W, S=s)$, where $P(M=1|Z=z, W, S=s)$ can be estimated using a fit of the mean outcome M as a function of Z, W, and S and getting predicted probabilities for $M=1$ setting $S=s$ and separately setting $Z=1$ and $Z=0$, and where $P(Z=z|A=a^*, W, S=s)$ can be estimated using a fit of the mean outcome$Z$ given A, W, and S and getting predicted probabilities for $Z=1$ and for $Z=0$, setting $A=a^*$ and $S=s$ and using observed values for W. Perform a logistic regression with a weighted intercept model as below, using weights, $\widehat{H}_n(O)$ the estimated weights from 2.10.

$$logit(Y) = logit(\bar{Q}^0_{Y,n}(M, Z, W)) + \epsilon)$$

$\bar{Q}^0_{Y,n}(M, Z, A, W)$ is then updated to $\bar{Q}^*_{Y,n}(M, Z, A, W) = logit(\bar{Q}^0_{Y,n}(M, Z, W)) + \epsilon)$. We note that in this first regression the fluctuation model above depends on A, so the predictions given by $\bar{Q}^*_{Y,n}(M, Z, A, W)$ will depend on A. We then perform the stochastic intervention on $\bar{Q}^*_{Y,n}(M, Z, A, W)$ via the computation $\bar{Q}^*_{M,n}(Z, A, W, S) = \mathbb{E}_{\hat{g}_{M|a^*,W,s}}[\bar{Q}^*_n(M, Z, A, W) \mid Z, W, S]$ and use these as outcomes for a regression on the variables A, W, S. The resulting fit is the initial estimate $\bar{Q}^0_{Z,n}(A, W, S)$. Then we update this fit by performing the following weighted intercept model:

$$logit(\bar{Q}^*_{M,n}(Z,W,S)) = logit(\bar{Q}^0_{n,Z}(A,W,S)) + \epsilon$$

using weights $H_Z(A,W,S) = \frac{I(S=0,A=a)}{p_A(A|W,S)p_S(S=0)}$. These weights are not estimated because we know the treatment mechanism for A and we know the predetermined proportion in our sample from site $S=0$. The updated initial estimate will be notated

$$\bar{Q}^*_{n,\hat{g},a^*,W}(A,W,S) = expit(logit(\bar{Q}^0_{Z,N}(A,W,S)) + \epsilon)$$

We then form our estimate by plugging in to

$$\hat{\Psi}_n = \sum_{i=1}^{n} \frac{I(S=0)}{\sum_{i=1}^{n} I(S=0)} \bar{Q}^*_{Z,n}(A=a, W_i, S=0)$$

We can easily verify, upon plugging in our updated regression models, $\bar{Q}^*_{Z,n}$ and $\bar{Q}^*_{M,n}$ and the other estimated portions of the likelihood to the influence curve, that

$$\sum_{i=1}^{n} \hat{D}^*(P^*_n)(O_i) = 0$$

where we evaluate the influence curve at the data generating distribution, P^*_n, given by our initial fits, $P^0_{M,n}$, $P^0_{Z,n}P^0_{A,n}P^0_{S|W,n}$, the empirical distribution of W, and the TML updates, $\bar{Q}^*_{M,n}$ and $\bar{Q}^*_{Z,n}$. The the influence curve is given by $\hat{D}^*(P^*_n) = \hat{D}^*_Y(P^*_n) + \hat{D}^*_Z(P^*_n) + \hat{D}^*_W(P^*_n)$ and
$\hat{D}^*_W(P^*_n)(O) = \left(\bar{Q}^*_Z(A=a,W,S) - \hat{\Psi}_n\right) \frac{I(S=0)}{p_S(S=0)}$
$\hat{D}^*_Z(P^*_n)(O) = \left(\bar{Q}^*_M(Z,A,W,S) - \bar{Q}^*_Z(A,W,S)\right) \frac{I(S=0,A=a)}{p_A(A|W,S)p_S(S=0)}$
$\hat{D}^*_Y(P^*_n)(O) = \left(Y - \bar{Q}^*_n(M,Z,A,W)\right) \widehat{H}_n(O)$.

2.5.1 TMLE inference

To compute the standard error for our estimates, we compute the sample standard deviation of the influence curve approximations over our data, $\{D^*(P^*_n)(O_i)\}^n_{i=1}$ over root n, which we will denote $\widehat{\sigma}_n(D^*(P^*_n))/\sqrt{n}$. Our 95% confidence bands,

$$\hat{\Psi}_n \pm 1.96 \times \widehat{\sigma}_n(D^*(P^*_n)(O))/\sqrt{n}$$

will cover the truth asymptotically at 95% under TMLE conditions in 1.2.1 and will be asymptotically as small as possible (for any regular asymptotically linear estimator) under the alternate H, in case we know the model is restricted as in section 2.2, and otherwise will be as small as possible for either the semi-parametric model with A and/or M mechanism known or non-parametric model. The stochastic direct effect (SDE) entails setting a^* to 0 (receive mediation effects as if under treatment a^*) and taking the difference in estimates between setting the treatment intervention, a, to 1 and setting a to 0. The corresponding influence curve approximation is just a likewise difference of the influence curve approximations for each parameter. The stochastic indirect effect (SIE) entails setting $a = 1$ and then taking the difference in estimates between setting $a^* = 1$ and $a^* = 0$. The corresponding influence curve approximations are again a likewise difference of the two influence curve approximations. For each of the estimators of SDE and SIE, we used the sample standard deviation of their respective influence curve approximations, divided by $\sqrt{n}$ for the standard error estimate.

2.5.2 Robustness Analysis

We compute the second order remainder term as presented in theorem 1.2.1.

$$R_2(P_n, P_0) = (\Psi_n - \Psi(P_0)) + P_0 D^*(P)$$

where we consider P as an estimate of P_0 for a lighter notation.

Theorem 2.5.1. *For model $\mathcal{M}_{II}$, we have the following:*

$$\begin{aligned}
&R_2(P_n, P_0) \\
=&\mathbb{E}_{P_0}(\bar{Q}_Y - \bar{Q}_{Y,0})(\bar{M})\Big[h_1(O)(g_{M,0} - g_M)(M \mid \bar{Z}) + h_2(O)(p_{Z,0} - p_Z)(Z \mid \bar{A}) \\
&+ h_3(O)(p_{A,0} - p_A)(A \mid \bar{W}) + h_4(O)(p_{S|W,0} - p_{S|W})(S \mid W)\Big] \qquad (2.12) \\
&+ \mathbb{E}_{P_0} h_5(O)(\bar{Q}_{Z,0} - \bar{Q}_Z)(\bar{A})(p_{A,0} - p_A)(A \mid \bar{W}) \qquad (2.13) \\
\leq& k \sum_{i=1}^{4} \|\bar{Q}_Y - \bar{Q}_{Y,0}\|_{L^2(P_0)} \|f_{i,0} - f_i\|_{L^2(P_0)} + k\|\bar{Q}_{Z,0} - \bar{Q}_Z\|_{L^2(P_0)} \|f_{3,0} - f_3\|_{L^2(P_0)}
\end{aligned}$$

where we substituted the following: $f_{1,0}(o) = g_{M,0}(m \mid z, x, w, s)$, $f_{2,0} = p_{Z,0}(z \mid a, w, 1)$, $f_{3,0} = p_{A,0}(x = a \mid w, s)$, $f_{4,0}(o) = p_{S|W,0}(x \mid w, s)$. Dropping the subscript, 0, indicates the estimated counterpart. Also, h_i is a bounded function by the positivity assumption (see section 2.4) and thus the last inequality holds with a sufficiently large k.

Corollary. *Assume:*

- *A1*

$$\begin{aligned}
\|\bar{Q}_{Y,0} - \bar{Q}_Y\|_{L_0^2(P_0)} \|p_{M,0} - p_M\|_{L_0^2(P_0)} =& \\
\|\bar{Q}_{Y,0} - \bar{Q}_Y\|_{L_0^2(P_0)} \|p_{Z,0} - p_Z\|_{L_0^2(P_0)} =& \\
\|\bar{Q}_{Y,0} - \bar{Q}_Y\|_{L_0^2(P_0)} \|p_{A,0} - p_A\|_{L_0^2(P_0)} =& \\
\|\bar{Q}_{Y,0} - \bar{Q}_Y\|_{L_0^2(P_0)} \|p_{S|W,0} - p_{S|W}\|_{L_0^2(P_0)} =& o_P(1/\sqrt{n})
\end{aligned}$$

- *A2*

$$\|\bar{Q}_{Z,0} - \bar{Q}_Z\|_{L_0^2(P_0)} \|p_{A,0} - p_A\|_{L_0^2(P_0)} = o_P(1/\sqrt{n})$$

Then $\sqrt{n} R_2(P, P_0) \xrightarrow{p} 0$

The proof is immediate when applying the cauchy-schwarz inequality.

Remark. Such conditions are guaranteed asymptotically when using the highly adaptive lasso to fit the regressions if the true regressions are of finite sectional variation norm and are left-hand continuous with right-hand limits (van der Laan 2016).

Remark. If A1 and A2 are satisfied, the TMLE and EE estimators will be consistent. If A1 is satisfied and we know the treatment mechanism, as in an RCT, then the TMLE and EE estimators are consistent.

The proof is given in section 3, Theorem 3.2.6 where we also prove robustness properties are the same for the restricted model TMLE and EE estimator.

2.6 Estimating Equation Estimator or One-Step Estimator

Next, we describe another estimating equation (EE) estimator of $\Psi(P)$, which solves the efficient influence curve equation, i.e., $PnD^*(P_n^0) = 0$ via Newton's method in one step. Hence, it is sometimes referred to as the one-step estimator, where P_n^0 is an initial estimate of the data generating system the same as that for TMLE. As for the stabilized IPTW estimator, we adjust the initial estimate we would obtain in the TMLE algorithm, not the model fits. So our estimate is not a plug-in estimator and therefore is not guaranteed to obey parameter bounds. We first form an initial estimate

$$\Psi_n^0 = \sum_{i=1}^{n} \frac{\mathbb{I}(S_i = 0)}{\sum_{i=1}^{n} \mathbb{I}(S = 0)} \bar{Q}_{Z,n}^0(A_i = a, W_i, S_i)$$

.

identical to the TMLE initial estimate. Then we update this estimate by adding the empirical mean of the approximated influence curve.

$$\hat{\Psi}_n^1 = \hat{\Psi}_n^0 + \sum_{i=1}^{n} D^*(P_n^0)(O_i)$$

This then leads to a second order expansion

$$\hat{\Psi}_n^1 - \Psi(P_0) = (P_n - P_0)D^*(P_n^0)(O) + R_2(P_n, P_0)$$

where $R_2(P_n, P_0) = \hat{\Psi}_n^0 - \Psi(P_0) + P_0 D^*(P_n^0)(O)$. The requirements for consistency and asymptotic efficiency on $D^*(P_n^0)$ and R_2 are then identical to those in section 1.2.1. Thus, predicated on the identical assumptions, robustness properties of the EE estimator are the same as for that of TMLE and the proof is virtually identical so we will omit it. We will see in simulations, however, that the EE estimator can be unstable due to it not being a plug-in estimator like the TMLE.

The standard errors of the EE estimate is computed as the sample standard deviation of $D^*(P_n^0)$ divided by $\sqrt{n}$. To obtain inference for the SDE or SIE we just follow the same instructions given in section 2.5.1.

2.7 Simulations

2.7.1 Overview

We compare finite sample performance of our three estimators in estimating the transport SDE and transport SIE using simulation. We show estimator performance in terms of absolute bias, efficiency, 95% confidence interval (CI) coverage, root mean squared error (RMSE), and percent of estimates lying outside the bounds of the parameter space across 1,000 simulations. For calculating the efficiency and the 95% CI coverage, we use both the IC and the bootstrap.

We consider three data-generating mechanisms (DGMs) within the structural causal model described in section 2.2. The DGMs are detailed in Table 2.6. DGM 1 is intended to break when Y and M models are misspecified, especially. DGM 2 is intended to break when Y and Z models are misspecified, especially. DGM 3 is intended to break when Y and S models are misspecified, especially.

Table 2.6: Simulation data-generating mechanisms.

Data Generating Mechanism 1	
$W_1 \sim bernoulli$	$P(W_1 = 1) = 0.5$
$W_2 \sim bernoulli$	$P(W_2 = 1) = expit(0.4 + 0.2W_1)$
$S \sim bernoulli$	$P(S = 1) = expit(3W_2 - 1)$
$A \sim bernoulli$	$P(A = 1) = 0.5$
$Z \sim bernoulli$	$P(Z = 1) = expit(-3A + -0.2S + 2W_2 + 0.2AW_2 - 0.2AS + 0.2W_2S + 2AW_2S - 0.2)$
$M \sim bernoulli$	$P(M = 1) = expit(1Z + 6W_2Z - 2W_2 - 2)$
$Y \sim bernoulli$	$P(Y = 1) = expit(log(1.2) + log(40)Z - log(30)M - log(1.2)W_2 - log(40)W_2Z)$
Data Generating Mechanism 2	
$W_1 \sim bernoulli$	$P(W_1 = 1) = 0.5$
$W_2 \sim bernoulli$	$P(W_2 = 1) = expit(0.4 + 0.2W_1)$
$S \sim bernoulli$	$P(S = 1) = expit(3W_2 - 1)$
$A \sim bernoulli$	$P(A = 1) = 0.5$
$Z \sim bernoulli$	$P(Z = 1) = expit(-0.1A + -0.2S + 0.2W_2 + 5AW_2 + 0.14AS + 0.2W_2S - 0.2AW_2S - 1)$
$M \sim bernoulli$	$P(M = 1) = expit(1Z + 3ZW_2 + 0.2ZS - 0.2W_2S + 2W_2Z + 0.2S - 0.2ZW_2S - W_2 - 2)$
$Y \sim bernoulli$	$P(Y = 1) = expit(-6Z + 0.2ZW_2 + 2ZM + 2W_2M - 2W_2 + 4M + 1ZW_2M - 0.2)$
Data Generating Mechanism 3	
$W_1 \sim bernoulli$	$P(W_1 = 1) = 0.5$
$W_2 \sim bernoulli$	$P(W_2 = 1) = expit(0.4 + 0.2W_1)$
$S \sim bernoulli$	$P(S = 1) = expit(3W_2 - 1)$
$A \sim bernoulli$	$P(A = 1) = 0.5$
$Z \sim bernoulli$	$P(Z = 1) = expit(-3A + 2S + 2W_2 + 0.2AW_2 - 0.2AS + 0.2W_2S + 2AW_2S - 0.2)$
$M \sim bernoulli$	$P(M = 1) = expit(3Z - 0.2ZW_2 + 0.2ZS - 0.2W_2S + 2W_2Z + 0.2S - 0.2ZW_2S - W_2 - 2)$
$Y \sim bernoulli$	$P(Y = 1) = expit(-6Z + 0.2ZW_2 + 2ZM + 2W_2M - 0.2W_2 + 4M + 1ZW_2M - 0.2)$

2.7.2 Results

Main Points of the Simulations

Due to the points below, TMLE is the best choice for estimating this parameter.

1. The efficient TMLE and EE estimators for the restricted model gave consistently large gains in efficiency over the TMLE and EE estimators not respecting these restrictions and even bigger gains in efficiency over the stabilized iptw estimator. The iptw was more efficient for the situation of all correctly specified models, except the model for Y. When estimating SIE, however, it lost in this situation every other time to the efficient TMLE estimator. It is also notable that TMLE tended to be more efficient than the EE estimator for this situation as well.

2. The EE estimator shows some instability in finite samples, even catastrophically giving almost all estimates outside the parameter bounds for DGM 1 under Y and M misspecified for estimating SDE. There-in lies the advantage of TMLE being a plug-in or substitution estimator, which always gives an answer within the parameter bounds.

3. Stabilized iptw is not guaranteed to be robust when only getting Y and A models correct, as expected. For n = 5000 and DGM 2, it never covered the SDE true parameter value where as TMLE and EE remained consistent for both their respective efficient and inefficient estimators.

4. The influence curve based inference for when only Y and A models were well-specified can be liberal and give poor coverage for TMLE and EE estimators even though these estimators are consistent. We show the bootstrap picks up the true variance of the estimators and corrects the coverage but in reality, when parametric assumptions fail and we need data adaptive estimation to make good predictions, the bootstrap will fail and therefore not be a theoretically sound option. Therefore, for the future, extra targeting as per Benkeser et al., 2017 , would be a valued added feature to obtain more reliable inference. Such corrective methodology is not theoretically sound for the EE method as shown in Benkeser et al., 2017.

Table 2.7: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 1 under well-specified models for sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 1, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.015	86.77	161.49	0.807	0.959	0.114	0
SDE_EE_eff	0.016	86.77	$9.71 * 10^9$	0.812	0.919	0.112	0
SDE_tmle	0.028	201.37	292.39	0.82	0.941	0.252	0
SDE_EE	0.024	216.18	$4.84*10^{11}$	0.923	0.924	0.210	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							

Continued on next page

Table 2.7 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SIE_tmle_eff	-0.005	103.450	52.71	0.833	0.961	0.044	0
SIE_EE_eff	-0.004	102.77	$2.21*10^{11}$	0.843	0.907	0.038	0
SIE_tmle	-0.007	126.57	545.44	0.791	0.948	0.056	0
SIE_EE	-0.004	118.11	158.02	0.829	0.910	0.039	0
DGM 1, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.0005	96.77	101.58	0.945	0.950	0.039	0
SDE_EE_eff	0.0005	96.81	101.20	0.946	0.948	0.039	0
SDE_tmle	0.0002	225.77	236.31	0.933	0.941	0.091	0
SDE_EE	0.0003	227.70	226.59	0.948	0.936	0.090	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0	101.96	110.93	0.938	0.940	0.008	0
SIE_EE_eff	0	102.35	107.99	0.939	0.939	0.008	0
SIE_tmle	0.0001	125.37	166.92	0.920	0.949	0.011	0
SIE_EE	0.0001	126.16	131.84	0.925	0.940	0.011	0
DGM 1, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0001	100.39	100.59	0.949	0.949	0.013	0
SDE_EE_eff	-0.0001	100.39	100.59	0.948	0.949	0.013	0
SDE_tmle	0.001	227.08	227.72	0.96	0.959	0.028	0
SDE_EE	0.001	227.21	227.14	0.96	0.958	0.028	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.0001	101.12	101.44	0.932	0.924	0.002	0
SIE_EE_eff	-0.0001	101.23	101.43	0.932	0.925	0.002	0
SIE_tmle	-0.0001	130.34	131	0.939	0.940	0.003	0
SIE_EE	-0.0001	130.50	130.94	0.940	0.941	0.003	0

Table 2.8: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 1 under well-specified models only for Y and A models sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 1, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.029	60.77	125.16	0.529	0.879	0.139	0
SDE_EE_eff	0.029	60.38	121.08	0.539	0.873	0.134	0
SDE_tmle	0.026	87.39	167.61	0.609	0.922	0.177	0
SDE_EE	0.029	93.83	134.66	0.693	0.878	0.159	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.009	54.10	200.16	0.616	0.932	0.076	0
SIE_EE_eff	-0.010	56.47	166.06	0.626	0.919	0.075	0
SIE_tmle	-0.012	69.12	244.45	0.572	0.923	0.089	0
SIE_EE	-0.010	77.87	172.43	0.669	0.912	0.079	0
DGM 1, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.005	53.490	100.99	0.643	0.879	0.064	0
SDE_EE_eff	-0.005	53.66	98.63	0.643	0.879	0.063	0
							Continued on next page

Table 2.8 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SDE_tmle	-0.007	95.11	141.58	0.747	0.895	0.092	0
SDE_EE	-0.006	97.42	125.71	0.843	0.920	0.078	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.001	64.72	97.48	0.826	0.948	0.019	0
SIE_EE_eff	0.0002	65.52	95.96	0.839	0.952	0.018	0
SIE_tmle	-0.0004	96.56	142.93	0.797	0.932	0.027	0
SIE_EE	-0.001	101.28	118.11	0.889	0.951	0.023	0
DGM 1, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0001	48.68	93.50	0.677	0.942	0.019	0
SDE_EE_eff	-0.0002	48.71	93.31	0.676	0.938	0.019	0
SDE_tmle	-0.0001	94.62	150.35	0.768	0.947	0.030	0
SDE_EE	-0.0002	94.81	124.45	0.854	0.948	0.025	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.0001	61.13	77.84	0.875	0.947	0.006	0
SIE_EE_eff	0.0001	61.18	75.11	0.888	0.945	0.005	0
SIE_tmle	0.0002	94.19	117.94	0.885	0.955	0.008	0
SIE_EE	0.0001	94.53	10	0.941	0.957	0.007	0

Table 2.9: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 1 under well-specified models for all but Y and S models, sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 1, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.196	99.74	130.55	0.348	0.496	0.229	0
SDE_EE_eff	0.23	119.11	$3.85*10^{13}$	0.275	0.230	0.254	0
SDE_tmle	0.195	143.81	234.82	0.556	0.812	0.290	0
SDE_EE	0.226	182.88	$5.62*10^{13}$	0.487	0.494	0.291	0.100
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.002	221.25	500.77	0.591	0.857	0.070	0
SIE_EE_eff	-0.011	264.85	1,190.67	0.704	0.642	0.061	0
SIE_tmle	-0.005	249.36	509.35	0.583	0.825	0.073	0
SIE_EE	-0.011	324.61	$1.23 * 120^{13}$	0.700	0.665	0.075	0
DGM 1, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.11	140.17	175	0.528	0.577	0.149	0
SDE_EE_eff	0.141	294.11	171.95	0.577	0.546	0.172	0
SDE_tmle	0.120	254.54	308.58	0.524	0.571	0.212	0
SDE_EE	0.141	454.66	307.74	0.514	0.475	0.235	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.026	290.33	468.23	0.429	0.673	0.040	0
SIE_EE_eff	0.013	638.03	393.79	0.826	0.664	0.034	0
SIE_tmle	0.022	384.49	542.07	0.606	0.803	0.039	0

Continued on next page

Table 2.9 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SIE_EE	0.013	739.29	525.56	0.810	0.728	0.041	0
DGM 1, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.045	226.69	190.97	0.717	0.512	0.054	0
SDE_EE_eff	0.078	519.45	183.26	0.910	0.064	0.083	0
SDE_tmle	0.051	490.16	587.79	0.854	0.924	0.093	0
SDE_EE	0.082	803.98	652.80	0.860	0.834	0.118	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.036	332.17	319.04	0.074	0.106	0.037	0
SIE_EE_eff	0.028	883.10	278.05	0.554	0.129	0.029	0
SIE_tmle	0.035	410.83	458.58	0.136	0.157	0.037	0
SIE_EE	0.027	1,059.64	688.37	0.740	0.391	0.031	0

Table 2.10: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 1 under well-specified models for all but the Y and Z models sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 1, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.195	92.71	116.32	0.395	0.524	0.222	0
SDE_EE_eff	0.216	113.89	117.45	0.453	0.469	0.241	0
SDE_tmle	0.076	261.30	292.91	0.824	0.865	0.323	0
SDE_EE	0.082	326.80	293.93	0.921	0.874	0.321	0.100
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.025	155.53	155.39	0.740	0.688	0.055	0
SIE_EE_eff	-0.018	181.09	119.44	0.794	0.665	0.046	0
SIE_tmle	-0.023	167.73	219.28	0.664	0.756	0.069	0
SIE_EE	-0.016	219.30	160.33	0.733	0.671	0.060	0
DGM 1, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.182	102.42	114.90	0.016	0.032	0.188	0
SDE_EE_eff	0.203	112.52	107.23	0.009	0.006	0.208	0
SDE_tmle	0.041	322.39	327.28	0.930	0.924	0.144	0
SDE_EE	0.039	334.96	319.05	0.942	0.929	0.138	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.006	200.08	124.34	0.903	0.808	0.011	0
SIE_EE_eff	-0.003	212.20	126.88	0.926	0.843	0.010	0
SIE_tmle	-0.002	241.17	229.07	0.883	0.872	0.016	0
SIE_EE	-0.003	283.03	231.65	0.892	0.862	0.017	0
DGM 1, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.181	104.08	117.69	0	0	0.181	0
SDE_EE_eff	0.202	111.38	105.15	0	0	0.202	0
SDE_tmle	0.050	327.02	332.07	0.782	0.791	0.064	0
SDE_EE	0.047	328.78	317.31	0.805	0.793	0.061	0

Continued on next page

Table 2.10 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.003	179.01	95.78	0.965	0.718	0.004	0
SIE_EE_eff	-0.0004	181.99	101.47	0.994	0.923	0.002	0
SIE_tmle	0.001	222.07	192.19	0.975	0.948	0.004	0
SIE_EE	0	247.54	198.36	0.987	0.946	0.004	0

Table 2.11: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 1 under well-specified models for all but Y and M models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 1, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.133	245.31	239.69	0.782	0.837	0.283	0
SDE_EE_eff	-1.216	1,423.80	1,532.61	0.895	0.890	1.984	51.100
SDE_tmle	-0.213	429.80	338.75	0.718	0.790	0.418	0
SDE_EE	-1.259	1,611.01	$1.13*10^{13}$	0.953	0.949	2.078	56.300
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.001	233.32	208.17	0.735	0.712	0.141	0
SIE_EE_eff	0.075	394.06	367.59	0.834	0.830	0.215	0.700
SIE_tmle	-0.044	166.40	205.14	0.539	0.575	0.139	0
SIE_EE	0.018	382.64	295.24	0.664	0.627	0.227	1.300
DGM 1, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.260	303.14	180.27	0.482	0.163	0.272	0
SDE_EE_eff	-1.181	1,040.61	1,021.97	0.080	0.066	1.272	76.400
SDE_tmle	-0.309	670.04	318.17	0.823	0.439	0.349	0
SDE_EE	-1.238	1,290.55	1,219.61	0.360	0.305	1.354	77.100
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.095	305.14	220.10	0.712	0.467	0.108	0
SIE_EE_eff	0.137	442.19	306.25	0.806	0.424	0.156	0
SIE_tmle	0.060	334.84	297.52	0.892	0.808	0.094	0
SIE_EE	0.112	591.83	473.35	0.965	0.828	0.163	0
DGM 1, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.272	280.08	149.17	0	0	0.273	0
SDE_EE_eff	-1.110	852.67	735.38	0	0	1.118	99.210
SDE_tmle	-0.272	698.69	321.77	0.147	0.001	0.276	0
SDE_EE	-1.111	1,093.33	877.83	0	0	1.121	98.410
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.104	249.26	173.89	0	0	0.106	0
SIE_EE_eff	0.135	336.39	227.03	0	0	0.137	0
SIE_tmle	0.103	370.51	279.70	0.036	0.016	0.106	0
SIE_EE	0.137	535.91	413.33	0.054	0.016	0.141	0

Table 2.12: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 1 under well-specified models for all but the Y model–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 1, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.045	157.49	170.41	0.872	0.893	0.143	0
SDE_EE_eff	0.050	232.07	$5.58*10^{13}$	0.894	0.872	0.149	0
SDE_tmle	0.010	282.54	314.26	0.876	0.912	0.273	0
SDE_EE	0.022	355.18	$4.90*10^{13}$	0.950	0.902	0.258	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.019	161.13	174.95	0.722	0.777	0.046	0
SIE_EE_eff	-0.019	177.84	$7.31*10^{13}$	0.751	0.710	0.044	0
SIE_tmle	-0.020	175.47	241.36	0.638	0.742	0.055	0
SIE_EE	-0.017	211.04	$3.08*10^{13}$	0.705	0.668	0.051	0
DGM 1, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.013	185.58	128.33	0.983	0.918	0.053	0
SDE_EE_eff	0.014	240.63	137.63	0.990	0.915	0.056	0
SDE_tmle	-0.0001	339.55	291.55	0.975	0.939	0.112	0
SDE_EE	0.002	370.42	288.90	0.988	0.939	0.112	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.002	217.44	131.94	0.930	0.880	0.010	0
SIE_EE_eff	-0.003	251	128.78	0.926	0.842	0.010	0
SIE_tmle	-0.003	255.86	222.71	0.883	0.879	0.015	0
SIE_EE	-0.004	308.32	222.67	0.885	0.833	0.016	0
DGM 1, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0002	191.06	118.82	0.999	0.965	0.014	0
SDE_EE_eff	-0.0004	240.26	126.67	0.999	0.968	0.015	0
SDE_tmle	0	354.81	288.79	0.984	0.954	0.035	0
SDE_EE	-0.0002	374.95	291.49	0.992	0.957	0.035	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0	186.27	102.20	0.995	0.914	0.002	0
SIE_EE_eff	0	237.87	103.75	0.997	0.919	0.002	0
SIE_tmle	0.0001	239.35	186.41	0.986	0.942	0.004	0
SIE_EE	0.0002	302.49	216.48	0.991	0.945	0.004	0

Table 2.13: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 2 under well-specified models for sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 2, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.003	102.26	139.52	0.970	0.992	0.064	0
SDE_EE_eff	0.004	102.13	$1.37*10^{13}$	0.974	0.989	0.060	0
SDE_tmle	0.007	293.07	375.10	0.852	0.945	0.218	0
SDE_EE	0.005	316.03	$8.83*10^{13}$	0.957	0.930	0.186	0
							Continued on next page

Table 2.13 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.003	93.01	132.95	0.878	0.944	0.033	0
SIE_EE_eff	-0.003	93.03	113.13	0.878	0.924	0.033	0
SIE_tmle	-0.001	99.85	158.57	0.865	0.952	0.041	0
SIE_EE	-0.002	99.09	$7.10*10^{12}$	0.871	0.915	0.036	0
DGM 2, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0003	101.37	105.16	0.946	0.957	0.028	0
SDE_EE_eff	-0.0003	101.40	104.51	0.946	0.957	0.028	0
SDE_tmle	-0.004	319.08	327.43	0.929	0.936	0.089	0
SDE_EE	-0.004	321.97	316.56	0.947	0.935	0.088	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.001	99.56	104.59	0.926	0.929	0.012	0
SIE_EE_eff	-0.001	99.58	101.11	0.926	0.929	0.012	0
SIE_tmle	-0.0005	101.74	106.99	0.928	0.935	0.012	0
SIE_EE	-0.0005	101.94	103.39	0.932	0.930	0.012	0
DGM 2, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0003	100.17	100.42	0.955	0.957	0.008	0
SDE_EE_eff	-0.0003	100.21	100.42	0.955	0.957	0.008	0
SDE_tmle	0.001	321.02	321.92	0.945	0.942	0.027	0
SDE_EE	0.001	321.27	321	0.946	0.942	0.027	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.0001	99.98	100.46	0.942	0.942	0.004	0
SIE_EE_eff	-0.0001	100.01	100.46	0.942	0.942	0.004	0
SIE_tmle	-0.0001	101.75	101.99	0.942	0.942	0.004	0
SIE_EE	-0.0001	101.69	101.95	0.942	0.943	0.004	0

Table 2.14: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 2 under well-specified models for Y and A models only–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 2, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.001	90.70	108.20	0.918	0.988	0.062	0
SDE_EE_eff	0.001	90.72	108.25	0.918	0.988	0.062	0
SDE_tmle	0.007	191.08	235.80	0.861	0.956	0.141	0
SDE_EE	0.004	207.43	191.55	0.957	0.951	0.121	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.002	51.42	132.19	0.575	0.927	0.082	0
SIE_EE_eff	0.002	51.31	128.88	0.577	0.920	0.080	0
SIE_tmle	0.002	96.11	158.41	0.721	0.892	0.099	0
SIE_EE	0.002	109.50	151.53	0.779	0.897	0.098	0
DGM 2, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.0002	99.14	105.22	0.946	0.969	0.028	0

Continued on next page

Table 2.14 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SDE_EE_eff	0.0002	99.14	105.22	0.946	0.969	0.028	0
SDE_tmle	-0.002	222.27	220.01	0.941	0.947	0.061	0
SDE_EE	-0.002	224.49	199.30	0.971	0.952	0.055	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.0004	51.46	110.33	0.622	0.934	0.032	0
SIE_EE_eff	-0.0005	51.46	109.03	0.625	0.933	0.032	0
SIE_tmle	0.001	110.16	133.88	0.850	0.921	0.044	0
SIE_EE	0.0003	114.19	136.01	0.871	0.926	0.047	0
DGM 2, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0	98.78	100.01	0.956	0.960	0.008	0
SDE_EE_eff	0	98.80	100.03	0.956	0.960	0.008	0
SDE_tmle	0.001	228.37	231.43	0.939	0.941	0.020	0
SDE_EE	0.001	228.50	203.24	0.966	0.942	0.018	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.0005	54.95	100.51	0.727	0.938	0.010	0
SIE_EE_eff	0.0004	54.95	100.46	0.732	0.939	0.010	0
SIE_tmle	0.001	129.68	136.04	0.879	0.920	0.014	0
SIE_EE	0.001	130.08	137.03	0.891	0.930	0.014	0

Table 2.15: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 2 under well-specified models for all models except Y and S models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 2, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.048	198.25	193.84	0.903	0.921	0.146	0
SDE_EE_eff	-0.102	315.28	$3.45*10^{13}$	0.946	0.875	0.181	0
SDE_tmle	-0.057	317.69	365.58	0.819	0.880	0.245	0
SDE_EE	-0.102	554.59	$8.79*10^{12}$	0.960	0.811	0.340	1.100
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.142	362.76	392.68	0.589	0.652	0.200	0
SIE_EE_eff	0.189	708.70	$9.86*10^{13}$	0.940	0.517	0.246	0
SIE_tmle	0.118	368.02	420.49	0.606	0.748	0.201	0
SIE_EE	0.184	749.61	$2.09*10^{12}$	0.944	0.682	0.267	0
DGM 2, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.162	309.07	214.95	0.395	0.121	0.170	0
SDE_EE_eff	-0.219	453.68	234.34	0.553	0.029	0.227	0
SDE_tmle	-0.173	519.71	423.35	0.712	0.599	0.206	0
SDE_EE	-0.226	758.85	619.39	0.817	0.613	0.281	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.220	529.14	274.22	0.016	0.012	0.222	0
SIE_EE_eff	0.277	1,080.48	309.93	0.203	0.006	0.280	0
SIE_tmle	0.213	600.82	434.83	0.060	0.052	0.219	0

Continued on next page

Table 2.15 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SIE_EE	0.278	1,253.70	726.29	0.442	0.132	0.292	0
DGM 2, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.164	285.91	168.14	0	0	0.165	0
SDE_EE_eff	-0.220	415.64	191.91			0.22	0
SDE_tmle	-0.164	511.22	399.63	0.019	0.010	0.167	0
SDE_EE	-0.218	699.90	586.82	0.029	0.023	0.224	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.222	507.95	186.55	0	0	0.223	0
SIE_EE_eff	0.278	1,033.84	216.14	0	0	0.279	0
SIE_tmle	0.222	600.12	343.91	0	0	0.222	0
SIE_EE	0.278	1,201.77	675.26	0	0	0.279	0

Table 2.16: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 2 under well-specified models for all but Y and Z models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 2, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.028	73.84	84.89	0.881	0.976	0.053	0
SDE_EE_eff	0.031	81.31	95.66	0.884	0.969	0.058	0
SDE_tmle	0.146	892.01	513.89	0.914	0.847	0.360	0
SDE_EE	0.576	1,247.94	1,341.48	0.935	0.923	0.966	24.300
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.046	165.05	120.15	0.815	0.728	0.086	0
SIE_EE_eff	-0.029	249.13	155.58	0.819	0.747	0.105	0
SIE_tmle	-0.068	413.29	276.43	0.799	0.695	0.132	0
SIE_EE	-0.003	724.66	686.48	0.825	0.837	0.467	5
DGM 2, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.031	74.57	78.71	0.647	0.693	0.038	0
SDE_EE_eff	0.033	79.85	84.01	0.658	0.697	0.040	0
SDE_tmle	0.179	1,002.95	559.87	0.991	0.735	0.224	0
SDE_EE	0.555	1,171.51	1,116.10	0.584	0.545	0.626	6.800
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.015	200.43	101.69	0.960	0.728	0.019	0
SIE_EE_eff	0.017	329.17	147.24	0.993	0.808	0.026	0
SIE_tmle	-0.040	469.27	531	0.853	0.667	0.064	0
SIE_EE	0.075	931.81	763.34	0.909	0.817	0.156	0.200
DGM 2, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.032	75.49	77.35	0	0	0.032	0
SDE_EE_eff	0.033	80.22	81.92	0	0	0.034	0
SDE_tmle	0.188	973.02	463.63	0.223	0.003	0.192	0
SDE_EE	0.545	1,112.09	1,008.79	0	0	0.552	0

Continued on next page

Table 2.16 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.014	192.60	92.53	0.380	0.004	0.014	0
SIE_EE_eff	0.018	307.60	121.67	0.767	0.029	0.019	0
SIE_tmle	-0.040	492.45	143.03	0.185	0	0.040	0
SIE_EE	0.066	911.65	686.60	0.447	0.231	0.072	0

Table 2.17: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 2 under well-specified models for all but Y and M models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 2, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.028	122.23	210.47	0.825	0.985	0.114	0
SDE_EE_eff	0.055	561.86	$2.01*10^{14}$	0.905	0.998	0.724	8
SDE_tmle	0.034	797.09	552.54	0.834	0.825	0.412	0
SDE_EE	0.026	1,816.94	$2.62*10^{14}$	0.974	0.972	1.374	29.500
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.030	283.99	172.61	0.842	0.780	0.092	0
SIE_EE_eff	-0.030	534.83	$7.88*10^{13}$	0.885	0.908	0.345	3.300
SIE_tmle	-0.030	252.15	207.16	0.686	0.734	0.110	0
SIE_EE	0.034	517.82	$8.40*10^{14}$	0.839	0.830	0.415	3.100
DGM 2, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.038	79.10	179.89	0.506	0.918	0.061	0
SDE_EE_eff	0.077	208.62	728.54	0.299	0.970	0.195	0
SDE_tmle	0.033	1,209.81	582.41	0.991	0.968	0.138	0
SDE_EE	0.077	1,870.98	1,799.55	0.987	0.990	0.430	3.200
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.036	213.86	103.92	0.525	0.230	0.041	0
SIE_EE_eff	-0.018	252.81	228.74	0.395	0.324	0.097	0
SIE_tmle	-0.035	290.63	159.72	0.681	0.399	0.043	0
SIE_EE	-0.017	354.25	346.99	0.588	0.565	0.102	0
DGM 2, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.039	72.52	169.93	0.039	0.211	0.042	0
SDE_EE_eff	0.074	116.09	615.71	0.073	0.715	0.090	0
SDE_tmle	0.040	1,191.15	425.74	1	0.780	0.055	0
SDE_EE	0.077	1,623.22	1,278.49	0.955	0.889	0.135	0
SDE_iptw	0.026	710.95	429.68	0.992	0.874	0.046	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.037	130.59	63.97	0.002	0	0.037	0
SIE_EE_eff	-0.020	126.70	105.49	0.178	0.129	0.034	0
SIE_tmle	-0.036	176.50	85.45	0.005	0	0.037	0
SIE_EE	-0.020	164.23	147.83	0.232	0.206	0.034	0
SIE_iptw	-0.030	134.37	83.04	0	0	0.031	0

Table 2.18: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 2 under well-specified models for all but the Y models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 2, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.016	117.52	135.93	0.928	0.991	0.071	0
SDE_EE_eff	0.017	134.03	$3.23*10^{13}$	0.971	0.994	0.069	0
SDE_tmle	-0.005	437.29	464.11	0.856	0.888	0.296	0
SDE_EE	0.003	510.08	$1.11*10^{14}$	0.971	0.887	0.270	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.015	154.70	134.11	0.741	0.759	0.061	0
SIE_EE_eff	-0.016	158.75	111.49	0.752	0.711	0.059	0
SIE_tmle	-0.013	170.37	205.62	0.649	0.798	0.076	0
SIE_EE	-0.015	185.12	$8.34*10^{11}$	0.700	0.667	0.072	0
DGM 2, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.001	119.51	107.23	0.981	0.978	0.026	0
SDE_EE_eff	-0.001	122.98	106.84	0.990	0.978	0.026	0
SDE_tmle	-0.002	527.31	463.49	0.974	0.950	0.115	0
SDE_EE	-0.0005	539.36	468.44	0.981	0.957	0.116	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.0002	207.84	113.29	0.993	0.962	0.012	0
SIE_EE_eff	0.0003	204.81	112.47	0.993	0.962	0.012	0
SIE_tmle	-0.0001	243.79	185.49	0.979	0.951	0.020	0
SIE_EE	-0.0001	240.52	177.94	0.978	0.944	0.021	0
DGM 2, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0003	111.40	100.18	0.963	0.936	0.009	0
SDE_EE_eff	-0.0003	111.55	100.18	0.968	0.937	0.009	0
SDE_tmle	0.002	508.10	414.57	0.984	0.947	0.034	0
SDE_EE	0.002	508.50	414.73	0.984	0.947	0.034	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.0001	195.72	101.34	0.999	0.941	0.004	0
SIE_EE_eff	-0.0001	189.44	101.82	0.999	0.942	0.004	0
SIE_tmle	-0.0002	231.77	163.22	0.992	0.954	0.006	0
SIE_EE	-0.0002	220.13	155.51	0.993	0.958	0.005	0

Table 2.19: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 3 under well-specified models for sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 3, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.020	93.10	158.60	0.823	0.955	0.115	0
SDE_EE_eff	-0.017	93.22	$5.82*10^{10}$	0.828	0.896	0.099	0
SDE_tmle	-0.030	126.80	239.56	0.809	0.951	0.199	0
SDE_EE	-0.020	139.54	$7.75*10^{11}$	0.890	0.875	0.140	0

Continued on next page

Table 2.19 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.011	90.59	159.14	0.803	0.929	0.114	0
SIE_EE_eff	0.008	90.40	$1.59*10^{11}$	0.809	0.905	0.104	0
SIE_tmle	0.018	90.81	173.84	0.742	0.907	0.152	0
SIE_EE	0.009	95.89	$2.02*10^{11}$	0.782	0.886	0.107	0
DGM 3, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.002	99.42	106.45	0.931	0.948	0.043	0
SDE_EE_eff	-0.002	99.45	106.04	0.931	0.947	0.043	0
SDE_tmle	-0.003	168.58	169.76	0.933	0.939	0.080	0
SDE_EE	-0.003	177.72	170.88	0.960	0.941	0.085	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.0005	100.48	107.39	0.928	0.937	0.045	0
SIE_EE_eff	-0.0005	100.48	106.91	0.927	0.937	0.045	0
SIE_tmle	-0.001	107.84	116.48	0.926	0.937	0.050	0
SIE_EE	-0.001	108.51	114.77	0.930	0.938	0.048	0
DGM 3, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.001	100.04	126.93	0.941	0.962	0.015	0
SDE_EE_eff	-0.0005	100.02	113.84	0.943	0.955	0.014	0
SDE_tmle	0.001	214.33	238.29	0.961	0.965	0.032	0
SDE_EE	0.0003	215.45	222.18	0.964	0.964	0.031	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0	100.07	118.92	0.939	0.938	0.015	0
SIE_EE_eff	-0.0002	100.07	100.78	0.940	0.936	0.014	0
SIE_tmle	-0.0004	108.04	125.46	0.938	0.946	0.016	0
SIE_EE	-0.0005	108.09	108.55	0.944	0.944	0.014	0

Table 2.20: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 3 under well-specified models for only the Y and A models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 3, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.013	71.12	119.65	0.744	0.896	0.097	0
SDE_EE_eff	-0.013	71.11	119.59	0.745	0.896	0.097	0
SDE_tmle	-0.011	95.20	159.52	0.781	0.942	0.124	0
SDE_EE	-0.012	95.46	131.92	0.844	0.920	0.106	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.005	29.99	132.19	0.390	0.899	0.050	0
SIE_EE_eff	0.005	30.01	131.03	0.389	0.898	0.050	0
SIE_tmle	0.003	38.27	150.41	0.438	0.915	0.054	0
SIE_EE	0.005	38.70	132.89	0.468	0.901	0.051	0
DGM 3, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
							Continued on next page

Table 2.20 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SDE_tmle_eff	-0.002	69.65	106.13	0.796	0.955	0.042	0
SDE_EE_eff	-0.003	69.61	106.09	0.797	0.954	0.042	0
SDE_tmle	-0.002	95.76	142.10	0.792	0.934	0.057	0
SDE_EE	-0.002	95.89	121.23	0.868	0.946	0.048	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.001	27.56	106.69	0.371	0.930	0.023	0
SIE_EE_eff	-0.0004	27.57	106.65	0.368	0.930	0.023	0
SIE_tmle	-0.001	37.31	113.59	0.464	0.930	0.024	0
SIE_EE	-0.0004	37.46	108.94	0.489	0.933	0.023	0
DGM 3, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0004	66.47	110.54	0.804	0.962	0.013	0
SDE_EE_eff	-0.0005	66.45	110.18	0.805	0.961	0.013	0
SDE_tmle	-0.0004	93.90	152.31	0.802	0.966	0.018	0
SDE_EE	-0.0005	93.90	125.67	0.890	0.963	0.015	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.0002	26.08	109.30	0.382	0.937	0.007	0
SIE_EE_eff		26.090	106.07	0.385	0.940	0.007	0
SIE_tmle	-0.0002	35.98	113.49	0.497	0.929	0.007	0
SIE_EE	0.0001	36	107.81	0.491	0.949	0.007	0

Table 2.21: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 3 under well-specified models for all but the Y and S models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 3, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.213	74.55	170.93	0.230	0.654	0.266	0
SDE_EE_eff	-0.254	88.47	$1.52*10^{13}$	0.115	0.274	0.279	0
SDE_tmle	-0.180	119.92	267.80	0.519	0.819	0.297	0
SDE_EE	-0.238	150.22	$6.71*10^{13}$	0.453	0.574	0.284	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.124	43.16	105.06	0.274	0.492	0.185	0
SIE_EE_eff	0.164	53.82	$2.61 * 10^8$	0.194	0.120	0.193	0
SIE_tmle	0.120	44.87	109.68	0.273	0.478	0.186	0
SIE_EE	0.162	58.43	$5.01*10^{13}$	0.226	0.178	0.192	0
DGM 3, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.142	93.50	153.26	0.236	0.334	0.160	0
SDE_EE_eff	-0.197	202.02	148.83	0.287	0.284	0.219	0
SDE_tmle	-0.136	204.33	296.39	0.477	0.717	0.195	0
SDE_EE	-0.191	371.54	301.70	0.301	0.280	0.260	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.088	60.27	94.55	0.111	0.350	0.101	0
SIE_EE_eff	0.139	73.44	42.17	0.008	0	0.144	0

Continued on next page

Table 2.21 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SIE_tmle	0.087	63.54	112.34	0.169	0.472	0.102	0
SIE_EE	0.139	79.94	53.53	0.020	0	0.144	0
DGM 3, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.056	203.22	244.19	0.477	0.548	0.065	0
SDE_EE_eff	-0.070	875.83	342.74	0.962	0.596	0.082	0
SDE_tmle	-0.062	518.76	616	0.813	0.889	0.111	0
SDE_EE	-0.074	1,340.79	935.49	0.862	0.865	0.162	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.125	147.45	105.41	0	0	0.126	0
SIE_EE_eff	0.133	67.89	35	0	0	0.134	0
SIE_tmle	0.122	158.62	145.54	0	0	0.124	0
SIE_EE	0.133	74.73	47.12	0	0	0.134	0

Table 2.22: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 3 under well-specified models for for all but the Y and Z models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 3, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.155	79.58	135.64	0.422	0.658	0.195	0
SDE_EE_eff	-0.150	117.23	136.56	0.580	0.714	0.188	0
SDE_tmle	0.005	264.09	290.02	0.806	0.874	0.287	0
SDE_EE	0.028	380.69	299.92	0.939	0.880	0.292	0.300
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.047	111.60	126.11	0.640	0.639	0.086	0
SIE_EE_eff	0.038	170.70	110.67	0.863	0.755	0.070	0
SIE_tmle	0.046	120.45	151.05	0.626	0.640	0.094	0
SIE_EE	0.032	195.13	146.58	0.691	0.639	0.088	0
DGM 3, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.103	111.77	124.79	0.307	0.383	0.112	0
SDE_EE_eff	-0.106	158.80	114.52	0.526	0.266	0.114	0
SDE_tmle	0.116	397.58	376.82	0.799	0.735	0.192	0
SDE_EE	0.120	481.43	371.48	0.942	0.746	0.191	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.002	141.83	125.67	0.977	0.954	0.022	0
SIE_EE_eff	0.0002	214.39	101.32	0.993	0.953	0.018	0
SIE_tmle	-0.008	151.37	186.26	0.892	0.923	0.035	0
SIE_EE	-0.004	269.77	193.99	0.973	0.938	0.036	0
DGM 3, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.098	105.11	103.48	0	0	0.099	0
SDE_EE_eff	-0.104	144.61	97.28	0	0	0.104	0
SDE_tmle	0.128	387.96	351.80	0.210	0.178	0.135	0

Continued on next page

Table 2.22 – continued from previous page

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
SDE_EE	0.127	440.62	351.63	0.282	0.183	0.134	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.005	122.10	87.11	0.959	0.826	0.007	0
SIE_EE_eff	-0.001	180.44	76.76	1	0.932	0.005	0
SIE_tmle	-0.014	143.41	146.80	0.602	0.626	0.017	0
SIE_EE	-0.005	230.11	164.35	0.992	0.931	0.011	0

Table 2.23: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 3 under well-specified models for all but the Y and M models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 3, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.038	130.75	184.19	0.744	0.913	0.157	0
SDE_EE_eff	-0.035	198.80	$8.27*10^{13}$	0.805	0.873	0.184	0.200
SDE_tmle	0.055	197.50	318.16	0.574	0.859	0.338	0
SDE_EE	-0.006	409.96	$5.23*10^{14}$	0.876	0.876	0.414	2.200
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.097	108.17	131.11	0.573	0.583	0.196	0
SIE_EE_eff	0.078	203.97	$1.18*10^{14}$	0.656	0.628	0.197	0.100
SIE_tmle	0.109	109.89	121.39	0.533	0.512	0.204	0
SIE_EE	0.082	207.74	$6.31*10^{14}$	0.622	0.575	0.211	0.100
DGM 3, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.001	170.79	155.19	0.958	0.948	0.058	0
SDE_EE_eff	0.011	252.72	167.82	0.991	0.963	0.059	0
SDE_tmle	0.079	443.35	490.67	0.569	0.642	0.266	0
SDE_EE	-0.010	909.43	847.53	0.984	0.707	0.427	1.800
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.017	110.82	158.64	0.871	0.949	0.066	0
SIE_EE_eff	-0.031	308.72	178.99	0.985	0.929	0.081	0
SIE_tmle	-0.010	125.30	181.72	0.876	0.948	0.076	0
SIE_EE	-0.028	328.96	226.06	0.962	0.928	0.098	0
DGM 3, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	0.004	157.81	131.85	0.975	0.931	0.018	0
SDE_EE_eff	0.013	215.88	125.17	0.994	0.844	0.021	0
SDE_tmle	0.013	628.84	586.62	0.942	0.945	0.076	0
SDE_EE	0.015	777.41	723.73	0.952	0.920	0.091	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.017	107.69	110.40	0.738	0.761	0.024	0
SIE_EE_eff	-0.029	248.66	133.17	0.992	0.644	0.035	0
SIE_tmle	-0.017	121.09	124.46	0.789	0.806	0.025	0
SIE_EE	-0.029	271.57	175.39	0.983	0.775	0.039	0

Table 2.24: Simulation results comparing estimators of $\Psi_{TransportSDE}$ and $\Psi_{TransportSIE}$ for DGP 3 under well-specified models for all but the Y models–sample sizes 100, 500 and 5000

Estimator	Bias	Efficiency		95%CI Cov		RMSE	% Out of Bds
		IC	Bootstrapping	IC	Bootstrapping		
DGM 3, N=100							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.067	114.03	177.02	0.731	0.893	0.158	0
SDE_EE_eff	-0.095	153.08	$9.73*10^{15}$	0.766	0.817	0.161	0
SDE_tmle	0.019	189.07	312.17	0.598	0.876	0.314	0
SDE_EE	-0.063	335.43	$2.51*10^{14}$	0.857	0.854	0.335	1.600
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.093	107.56	142.73	0.533	0.582	0.200	0
SIE_EE_eff	0.094	172.64	$1.18*10^{14}$	0.612	0.562	0.185	0
SIE_tmle	0.101	109.75	138.89	0.516	0.532	0.203	0
SIE_EE	0.096	180.93	$2.42*10^{14}$	0.604	0.545	0.193	0
DGM 3, N=500							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.007	147.33	140.13	0.954	0.949	0.054	0
SDE_EE_eff	-0.025	215.05	133.32	0.971	0.923	0.056	0
SDE_tmle	0.066	411.77	462.31	0.607	0.674	0.238	0
SDE_EE	-0.025	729.40	644.29	0.992	0.752	0.314	1
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	-0.009	113.07	157.65	0.898	0.955	0.060	0
SIE_EE_eff	0.004	270.99	147.05	0.985	0.959	0.058	0
SIE_tmle	-0.005	126.18	184.44	0.896	0.958	0.068	0
SIE_EE	0.003	293.04	190.52	0.970	0.950	0.074	0
DGM 3, N=5000							
Transport stochastic direct effect ($\Psi_{TransportSDE}$)							
SDE_tmle_eff	-0.0002	147.13	118.60	0.984	0.947	0.015	0
SDE_EE_eff	-0.001	250.94	124.08	0.992	0.950	0.014	0
SDE_tmle	0.006	574.32	554.59	0.925	0.910	0.076	0
SDE_EE	-0.002	716.64	643.04	0.947	0.910	0.087	0
Transport stochastic indirect effect ($\Psi_{TransportSIE}$)							
SIE_tmle_eff	0.0001	130.82	108.84	0.984	0.952	0.014	0
SIE_EE_eff	0.001	220.90	108.07	1	0.939	0.015	0
SIE_tmle	-0.0001	146.59	126.90	0.955	0.939	0.017	0
SIE_EE	0.001	240.50	147.62	0.999	0.938	0.020	0

2.8 Conclusion

In this paper, we defined and identified parameters that transport stochastic direct and indirect mediating effects from a source population ($S = 1$) to a new, target population ($S = 0$). Identification of such parameters rely on the typical sequential randomization and positivity assumptions of other stochastic mediation effects (Rudolph, Sofrygin, Zheng, et al. 2017; Zheng and M. v. d. Laan 2017; VanderWeele and Tchetgen Tchetgen 2017) as well as a common outcome model assumption, described previously for transport estimators (Rudolph and van der Laan 2017), which can be tested nonparametrically (A. R. Luedtke, Carone, and M. J. v. d. Laan 2015). Such parameters enable the prediction of mediating effects in new populations based on data about the mediation mechanism in a source population and the differing distributions of compositional characteristics between the two populations. Thus, transport SDE and SIE parameters contribute to understanding how and why interventions may work differently and/or have differing effects when applied to new populations.

We proposed five estimators for such effects: A stabilized weighted IPTW estimator, TMLE and one-step estimator (EE) using the efficient influence curve for both the restricted and unrestricted models. As expected, when we simulated by generated data from the restricted model, the restricted TMLE and EE had very considerable gains in efficiency. Overall, TMLE matches the coverage and MSE of the one-step estimator with greater stability in finite samples in that it never strays outside the parameter bounds. In a couple of instances, the EE gave almost 100% of its estimates outside the realm of possibility.
We also saw both the TMLE and EE are consistent when only the outcome model and the treatment mechanism is known, while IPTW is not. However, in such circumstances we might get the influence curve wrong and thus invalid inference when using the approximated influence curve variance for the standard errors. This was corrected with the bootstrap here because we used logistic regressions which can be non-parametrically bootstrapped and therefore, recover the true variance. In practice, when we use highly adaptive prediction methods that are necessary to reduce bias, the non-parametric bootstrap does not give proper standard error estimates. Therefore, we can look to the future to employ the doubly robust inference procedures as in Benkeser et al, 2017 where we perform extra targeting of the variance of the influence curve in order to both use great prediction methodology and obtain valid influence curve approximations for the inference.

The estimators we propose are limited in that they consider a stochastic intervention on mediator, M, that is assumed known and estimated from observed data. However, we plan to extend them to a true, unknown stochastic intervention in the future. The author has already derive the efficient influence curve for this parameter (see section 3.1.10). Another limitation is that the parameters are only identified if one assumes a common outcome model across the source and target populations. There will be some research questions for which it is not possible to establish evidence for or against this assumption, as in questions about predicting a long-term outcome in a new population. However, when the research question instead focuses on establishing the extent to which mechanisms are shared across populations, and the full set of data $O = (S, W, A, Z, M, YS)$ is observed for both populations, one can empirically test whether there is evidence against such a shared outcome model (A. R. Luedtke, Carone, and M. J. v. d. Laan 2015).

In the main text, we focused on transporting mediation estimates where an instrument, A, was statically intervened on and mediator M was stochastically intervened on. Moreover, we were primarily concerned with a statistical model that imposed instrumental variable assumptions such as the exclusion restriction assumption. However, we describe how each estimator can be easily modified to accommodate statistical models that do not impose instrumental variable assumptions, allowing for a direct effect of A on M and of A on Y, in scenarios where A is randomly assigned only conditionally on W, S, and for data generating mechanisms that do not include an intermediate variable, Z. Thus, our transport mediation estimators can be applied to a wide-range of common data generating mechanisms.

Theory, Derivations and Implementation

3.1 Tutorial: Deriving Efficient Influence Functions for Large Models

This section aims to provide the reader with a useful tutorial on how to derive efficient influence functions for non-parametric and semi-parametric models, while providing some necessary background for the reader so as to understand the core concepts involved in the process. It is the author's aim that this paper unifies the derivation procedure for a very broad class of parameters in a simple way so as to draw the broader statistics community into embracing statistical techniques for large models. It is also the aim of this paper for it to be self-contained, only indicating places where the reader might explore concepts in more detail but such exploration is not at all needed.

3.1.1 The Hilbert Space

The efficient influence function can be seen as an element of a Hilbert space, which generalizes familiar geometrical properties to allow for infinite dimensional spaces.

Definition 3.1.1. A Hilbert space, $\mathcal{H}$, has an inner product, denoted by $\langle \cdot, \cdot \rangle$, which takes as arguments any two elements of $\mathcal{H}$ and obeys the following:

1. $\langle x, y \rangle = \overline{\langle y, x \rangle}$ where $\bar{a}$ is the complex conjugate of a. However, for this paper, we are only considering real-valued inner products, so x and y are simply reversible in the inner product as in, $\langle x, y \rangle = \langle y, x \rangle$.

2. $\langle x + z, y \rangle = \langle x, y \rangle + \langle z, y \rangle$

3. The norm $\| \cdot \|$ of any $x \in \mathcal{H}$ is given by $\langle x, x \rangle = \|x\|^2$. The norm must obey the natural notion of distance as mathematically defined here:

   (a) $\|x + y\| \leq \|x\| + \|y\|$, the triangle inequality

   (b) $|a|\|x\| = \|ax\|$

   (c) $\|x\| = 0 \iff x = 0$

4. $a\langle x, y \rangle = \langle ax, y \rangle = \langle x, ay \rangle$ for scalar a.

A Hilbert space is complete with respect to the norm, which means the space includes the limit of all cauchy sequences under the norm. Cauchy sequences are sequences where the elements get closer and closer together, which is a fundamental distinction but more fundamental than we need in order to proceed with clarity. For more background on the basics of Hilbert spaces, the reader may consult Folland,1999. Here are two examples of Hilbert spaces, the second of which forms the basis of this paper (no pun intended):

Example 3.1.1. $\mathbb{R}^2$
The points on the cartesian plane form a 2-dimensional Hilbert space and it is equipped with an inner product more familiarly known as the dot product. If $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$, then $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2$.

This example is sufficient to convey a few of the key geometrical properties of Hilbert spaces we will use.

- **Orthogonality:**

  If the inner product of any two elements is 0, we say they are orthogonal. In $\mathbb{R}^2$ we can see this fits our visual notion of such.

- **Unique Projection:** We notate the projection of (x, y) on the subspace, $\mathbf{X} = \{(x, 0) | x \in \mathbb{R}\}$, as follows: $\prod((x, y) \| \mathbf{X})$. We see, just by regarding the shadow of (x,y) on the x-axis, that the projection is $(x, 0)$ and it is unique. We have a more general formula for projecting any vector on a subspace but this example suffices to illustrate that any projection must satisfy the following two properties:

- **Two Properties of Projections**

  1. The projected item must be in the space onto which it is projected: (x, 0) is in $\{(x, 0) | x \in \mathbb{R}\}$), which it obviously is.
  2. The projected element minus its projection must be perpendicular to the projection. This means the projection is the closest element in the space to the projected element. This is easy to verify for this basic example because $(x, y) - (x, 0) = (0, y)$ and $(0, y) \perp (x, 0)$ because the dot product $\langle (0, y), (x, 0) \rangle = (0, y) \cdot (x, 0) = 0$. We can see in the plane that these two vectors are perpendicular. Such a geometrical interpretation of projection also follows for infinite dimensional Hilbert spaces.

- **Direct Sum Decomposition:** Coming from the fact we have unique projections, we can decompose $\mathbb{R}^2$ into 2 orthogonal subspaces, $\mathbf{X} \oplus \mathbf{Y} = \{(x, 0) | x \in \mathbb{R}\} \oplus \{(0, y) | y \in \mathbb{R}\}$. Any $(x, y) \in \mathbb{R}^2$ can be written as unique sum of projections, $\prod((x, y) \| \mathbf{X}) + \prod((x, y) \| \mathbf{Y})$. More generally, if $\mathbf{Z}$ were any subspace such as any arbitrary line through the origin, then its orthogonal complement, i.e., the perpendicular line through the origin, $\mathbf{Z}^\perp$ would also decompose $\mathbb{R}^2$ as $\mathbf{Z} \oplus \mathbf{Z}^\perp$ and $(x, y) = \prod((x, y) \| \mathbf{Z}) + \prod((x, y) \| \mathbf{Z}^\perp)$. If a Hilbert space has direct sum decomposition, $\mathcal{H} = \mathbf{H_1} \oplus \mathbf{H_2} \oplus ... \oplus \mathbf{H_m}$, then all $h \in \mathcal{H}$ can be written as the unique sum $h = \prod(h \| \mathbf{H_1}) + \prod(h \| \mathbf{H_2}) + ... + \prod(h \| \mathbf{H_m})$.

Example 3.1.2. $L_0^2(P)$

$L_0^2(P)$ is the hilbert space of mean 0 functions of finite variance with respect to P, i.e. for all $f \in L_0^2(P)$, $\mathbb{E}_P f(O) = 0$ and $\mathbb{E}_P f(0)^2 < \infty$. The inner product of two elements, f and g in $L_0^2(P)$ is defined as $\langle f, g \rangle = \mathbb{E}_P[f(O)g(O)]$. Thus two elements are considered orthogonal if their covariance is 0. $L_0^2(P)$ is an infinite dimensional Hilbert space we will focus upon exclusively for this tutorial. The reader can consult Folland, 1999, section 5.5 for more detail on Hilbert spaces.

A Note on Integration and Measure Theory

A measure, ν, is a non-negative mapping defined on a σ-algebra, which we will consider as a set of subsets from a larger set. The trio, consisting of larger set, σ-algebra and measure, define a **measure space**, denoted by $(\mathcal{X}, \mathcal{A}, \nu)$. Let the larger set $\mathcal{X} = \mathbb{R}$ and let ν be the Lebesgue measure, which simply measures the length of any interval, (a, b), i.e., $\nu((a, b)) = b - a$. This is the measure used for introductory integration. The σ-algebra we consider for Lebesgue measure is naturally the borel σ-algebra, $\mathcal{B}$, which is the set of all countable unions and intersections of intervals of the form (a, b). We could have also used closed or half-open intervals to generate $\mathcal{B}$ as well. $\mathcal{B}$ also includes singleton sets of points because $\{a\} = \cap_{i=1}^{\infty}(a - 1/i, a + 1/i)$, i.e., the countable intersection of ever smaller open intervals about a.

Naturally we should have the following equivalence: $\nu(\{a\}) = \nu\left(\cap_{i=1}^{\infty}(a - 1/i, a + 1/i)\right) = \lim_{i \to \infty} \nu(a - 1/i, a + 1/i) = \lim_{i \to \infty} 2/i = 0$, since the set $\{a\}$ has length 0. In order that the measure of a limit of nested intersections is a limit of the measures of the sets (and likewise for nested unions), we could not have included all sets of real numbers in $\mathcal{A}$. Though this fact is surprising and intriguing in its own right, we need not delve into it further. For more about the necessity of σ-algebras and a complete mathematical construction of measures, the interested reader may consult Folland, 1999, chapters 1 and 2.

The examples below cover the situations we will encounter, essentially binary or continuous conditional distributions.

1. **Counting measure**: Let $\mathcal{X} = \{0, 1\}$ and consider σ-algebra $\mathcal{A} = \{\{0\}, \{1\}, \{0, 1\}\}$. The "measure space", $(\mathcal{X}, \mathcal{A}, \nu)$, is thusly defined via $\nu(\{0\}) = \nu(\{1\}) = 1$ and $\nu(\{0, 1\}) = 2$. For $\mathcal{X} = \mathbb{N}$, the counting numbers and $\mathcal{A}$ the set of all subsets of $\mathbb{N}$, the counting measure does the same thing in that it counts the number of elements in a set.

2. **Lebesgue measure, 2-d:** We might have ν on the σ-algebra generated by countable unions and intersections of all boxes in $\mathbb{R}^2$ as in 2-d college calculus. Here $\mathcal{B}$ is generated by countable unions and intersections of boxes on the plane and the measure space $(\mathbb{R}^2, \mathcal{B}, \nu)$ is defined by ν giving each 2-d box a measure equal to its area.

3. **Lebesgue with counting measure:** Let $\mathcal{X} = \mathbb{R} \cup \{0, 1\}$ and $\mathcal{A}$ = all sets generated by countable unions and intersections of sets of the form $\{(a, b), z)\}$ where z can be 0 or 1. In this case, ν puts a weight of $b - a$ on each of these sets, which will define Lebesgue

measure isolated to when $z = 1$ or $z = 0$. We might do the same, using $\mathcal{X} = \mathbb{R}^2 \cup \{0, 1\}$ where ν maps each 2-d box to its area or the equivalent for $\mathcal{X} = \mathbb{R}^d \cup \{0, 1\}$.

Integral Notation

ν is said to **dominate** P ($P << \nu$) or is a **dominating** measure of P if whenever $\nu(A)$ is 0, so is $P(A)$ for two measure spaces, $(\mathcal{X}, \mathcal{A}, \nu)$ and $(\mathcal{X}, \mathcal{A}, P)$. This leads to P having a unique Radon-Nikodym derivative (Folland 1999) of P with respect to ν, otherwise known as the density of P, notated with the lowercase, p. For a measure space, $(\mathcal{X}, \mathcal{A}, P)$, we write, for a set $A \in \mathcal{A}$, $P(A) = \int_A p(x) d\nu(x)$, which is sometimes written as $P(A) = \int_A dP(x)$. One might connect this with our intro calculus notation for a continuous 1-dimensional random variable, X, and Lebesgue measure, ν, where $\frac{dP}{d\nu}(x) = \frac{dP}{dx}(x) = p(x)$, a standard derivative. Then we would have $P(A) = \int_A \frac{dP}{dx}(x) dx$ as in the fundamental theorem of calculus. However, the intro calculus notion of derivative and integral breaks down if random variable X is discrete, say, or a combination of discrete and continuous variables, so the Radon-Nikodym derivative is much more general and less confining. We will always use the symbol, ν, as the dominating measure in this tutorial.

It is best to illustrate, via some basic examples, the computational fluidity measure theory provides. We will use these basic ideas throughout the tutorial:

1. Let Y be the outcome with continuous conditional distribution, $P_Y(Y \mid X)$ for a random variable, X. The dominating measure of $P_Y(Y \mid X)$, will be Lebesque measure, ν, and the density is written $p_Y(y \mid x)$. The mean of Y given X is given by $\mathbb{E}[Y \mid X]$ which we notate as $\int y p_Y(y \mid X) d\nu(y) = \int y p_Y(y \mid X) dy$ as we might be most familiar from intro calculus. Here we think of integrating as a limiting process of finer and finer reimann sums.

2. Let Y be a binary outcome conditional on X with binary conditional distribution, $P_Y(Y \mid X)$. The dominating measure of P_Y will be the counting measure, ν. The mean of Y given X is given by $\int y p_Y(y \mid X) d\nu(y) = 1 p_Y(1 \mid X) d\nu(1) + 0 p_Y(0 \mid X) d\nu(0) = p_Y(1 \mid X)$ as we expect for a binary. Notice, $d\nu(y)$ is the same as $v(\{y\}) = 1$ for $y = 0$ or 1. In other words, for the counting measure $d\nu$ and ν are interchangeable for a set of one element and the integral wrt a counting measure is just a sum. That is, for a discrete random variable, Y, taking values $\{y_i\}_{i=1}^m$, where m might be infinite, as in a Poisson distribution, we can write the conditional mean of $Y \mid X$ as $\int y p_Y(y \mid X) d\nu(y) = \sum_{i=1}^m y_i p_Y(y_i \mid X) d\nu(y_i)$ where $d\nu(y_i) = 1 = \nu(y_i)$. In other words, this sum is as fine-grain as we can get and hence, is equivalent to the integral.

3. **Multiple Integrals**

   Consider random variable $O = (X, Y) \sim P$ with density, p. The density factors as $p(o) = p_Y(y \mid x) p_X(x)$, where p_Y and p_X are the conditional densities. Consider function f defined by $f(x, y)$ for some formula basic formula like $exp(x + y)$ or a polynomial.

$$\begin{aligned}
\mathbb{E}f(X,Y) &= \int f(x,y)p(x,y)d\nu(x,y) \\
&= \int f(x,y)p_Y(y \mid x)p_X(x)d\nu(x,y) \\
&\text{note the equivalence with a double integral: we will use this frequently} \\
&= \int \underbrace{\int f(x,y)p_Y(y \mid x)d\nu(y)}_{\text{can integrate here}} p_X(x)d\nu(x) \\
&= \underbrace{\int \int f(x,y)p_Y(y \mid x)d\nu(y)p_X(x)d\nu(x)}_{\text{can integrate outside first wrt x}}
\end{aligned}$$

If Y is, say, binary and X is continuous or for joint distribution of X and Y, we technically cannot use the same symbol, ν, for all of their dominating measures, but we will not worry about that and abuse the notation for convenience. This doesn't affect our computation in that for the double integral we will understand which dominating measure (for our purposes either counting measure or Lebesgue measure) we are considering by the variable we are integrating with respect to. It is also notable that whether we integrate the expression via the inner integral then the outer or vice-versa, both come out the same as integrating the single integral directly. This is the substance of the **fubini-tonelli** theorem (Folland 1999), which the reader may look into further.

Remark. Computations in this tutorial will be with respect to densities of single variables and only involve the counting measure as the dominating measure.

4. **Common tricks we will use:** Consider the previous item with continuous conditional distribution of Y given X and X binary.

$$\begin{aligned}
&\int yp_Y(y \mid 1)d\nu(y) \\
=&\int \int yp_Y(y \mid x)d\nu(y)\frac{x}{p_X(x)}p_X(x)d\nu(x)
\end{aligned}$$

$$\begin{aligned}
&\int y(p_Y(y \mid 1) - p_Y(y \mid 0))d\nu(y) \\
=&\int \int yp_Y(y \mid x)d\nu(y)\frac{2x-1}{p_X(x)}p_X(x)d\nu(x)
\end{aligned}$$

The reader may verify these facts.

5. **Instructive Advertisement for Measure Theory:**

Though we never need to consider this case, it is instructive for the reader so as to understand the nice generality afforded by measure theory in integrating as well as the notion of a unique density (the radon-nikodym derivative) corresponding to a probability distribution and its dominating measure. This takes us beyond what we need for our computations but will provide confidence in using the notation. Let the distribution Y be given by the distribution function,

$$F(y) = \begin{cases} y/2 & 0 \leq y < 1/2 \\ y/2 + 1/2 & 1/2 \leq y \leq 1 \end{cases}$$

Notice, F is not continuous. We have thusly defined a measure space, $([0,1], \mathcal{B}_{[0,1]}, P)$ where $P((a,b)) = \frac{b-a}{2} + \frac{1}{2}\mathbb{I}(1/2 \in (a,b))$. Say our dominating measure is $\nu((a,b)) = b - a + \mathbb{I}(1/2 \in (a,b))$. Then our unique radon-nikodym derivative is the density $p(y) = 1/2$ for $0 \leq y \leq 1$.

To see this, notice for the latter density that we have:

$$\begin{aligned} \int p(x)d\nu(x) &= \int_{[0,1/2)} p(x)d\nu(x) + \int_{\{1/2\}} p(x)d\nu(x) + \int_{(1/2,1]} p(x)d\nu(x) \\ &= 1/4 + p(1/2) \times \nu(\{1/2\}) + 1/4 = 1 \end{aligned}$$

Hence we are forced into defining the density so that $p(1/2) = 1/2$ for the total probability to be 1. We also see "area under the density" interpretation for probability of a set fails because the area under the density is 1/2, not 1, if we use Lebesgue measure.

If $\nu((a,b)) = b - a + \frac{1}{2} \times I(1/2 \in (a,b))$ then

$$p(y) = \begin{cases} 1/2 & 0 \leq y < 1/2 \\ 1 & y = 1/2 \\ 1/2 & 1/2 < y \leq 1 \end{cases}.$$

To see this, notice for the latter density we have:

$$\begin{aligned} \int p(x)d\nu(x) &= \int_{[0,1/2)} p(x)d\nu(x) + \int_{\{1/2\}} p(x)d\nu(x) + \int_{(1/2,1]} p(x)d\nu(x) \\ &= 1/4 + p(1/2) \times \nu(\{1/2\}) + 1/4 = 1 \end{aligned}$$

Hence we are forced into defining the density so that $p(1/2) = 1$. Thus for any probability measure P and accompanying dominating measure, ν, we have a unique radon-nikodym derivative we can use for integrating. The general result is proven in Folland, 1999.

3.1.2 Tangent Spaces and Factorization of Densities

Now that we have taken care of some necessary notational considerations we are ready to illustrate the general technique of deriving efficient influence curves. We therefore discuss some important objects in efficiency theory.

Tangent Space for Nonparametric Model

First, we consider the model, $\mathcal{M}$, to be the set of all possible distributions for our true distribution. Since we assume nothing about this set of models we will call it non-parametric. We will consider observed data, which for a single observation is written as, $O \in \mathbb{R}^d$, and $O \sim P \in \mathcal{M}$. The density of P factors as follows:

$$p(o) = \prod_{i=1}^{d} p_{O_i}(o_i \mid \bar{o}_{i-1})$$

where $o = \bar{o}_d = (o_d, ..., o_1)$, where the reader may note that we order the variables moving backward in time from left to right, when we write them. We will generally establish a time ordering of variables and use the subscript notation to represent the conditional densities. So p_{O_i} is the conditional density of o_i given the previous variables, $\bar{o}_{i-1}$.
Pulling from van der Vaart, 2001, we define a path through P as a 1-dimensional submodel that passes through P at $\epsilon = 0$ in the direction, S.

$$\{P_\epsilon \in \mathcal{M}, p_\epsilon = (1 + \epsilon S)p \text{ s.t. } \int S(o)p(o)d\nu(o) = 0, \int S^2(o)p(o)d\nu(o) < \infty \text{ and } P_{\epsilon=0} = P\}$$

The tangent space, T, at a distribution, P, is the closure in the $L_0^2(P)$ norm of the set of scores, S for the all the paths through P. This turns out to be the entirety of the Hilbert space $L_0^2(P)$ since $L_0^2(P)$ is already complete. We write:

$$T = \overline{\{S | \mathbb{E}_P S(O) = 0, \mathbb{E}_P S(O)^2 < \infty\}} = L_0^2(P)$$

where the overbar represents the closure of the set.

1. The reader may quickly verify that for a given submodel, $S = \frac{d}{d\epsilon} logp_\epsilon \Big|_{\epsilon=0}$. Thus scores retain the intuitive notion of derivative of log likelihood as with parametric models. The only difference is here, we have infinitely many score directions that span an infinite dimensional space.

2. Another useful observation is that every element of the submodel in a non-parametric model for our d-dimensional data, O, has a density that also factors as follows: $p_\epsilon(o) = \prod_{i=1}^{d} p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1})$, where $\bar{o}_{i-1} = (o_{i-1}, ..., o_1)$ where $p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1}) = p_{O_i}(o_i \mid \bar{o}_{i-1})$ at $\epsilon = 0$. This implies

$$\begin{aligned} S(o) &= \sum_{i=1}^{d} \frac{d}{d\epsilon} \log p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1}) \Big|_{\epsilon=0} \\ &= \sum_{i=1}^{d} S_{O_i}(\bar{o}_i) \end{aligned}$$

and the reader may also verify S_{O_i} and S_{O_j} have covariance 0, i.e., $S_{O_i} \perp S_{O_j}$ in $L_0^2(P)$ for $i \neq j$.

3. $S_{O_i} \in T_{O_i} = \overline{\{g \mid E[g(O) \mid O_{i-1}] = 0, E[g^2(O)] \leq \infty\}}$ and T_{O_i} forms a subspace of T. **EXERCISE:** The reader may verify that $T_{O_i} \perp T_{O_j}$ for $i \neq j$. That is, all elements of T_{O_i} have covariance 0 with those of T_{O_j}.
4. The projection of S on T_{O_i} is given by $\prod(S \mid T_{O_i}) = \mathbb{E}[S(O) \mid \bar{O}_i] - \mathbb{E}[S(O) \mid \bar{O}_{i-1}]$. **EXERCISE:**The reader may verify that this is indeed a projection by verifying the projection is in the set upon which it is projected and that $(S - \prod(S \mid T_{O_i})) \perp \prod(S \mid T_{O_i})$, i.e. has covariance 0 with respect to P. This exercise is good preparation for the rest of the tutorial.
5. $T = T_{O_d} \oplus ... \oplus T_{O_1}$. Any score, S, is thusly a unique sum of its projections on the d tangent subspaces and those projections are given by $S_{O_i} = \frac{d}{d\epsilon} \log p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1})\Big|_{\epsilon=0}$.

We thus have the following convenient identity we will call upon for all derivations of efficient influence curves. Noting the introductory calculus fact by the chain rule, $\frac{d}{dx} \log f(x) = \frac{\frac{df}{dx}(x)}{f(x)}$, we arrive at the following identity:

A Key Identity

$$\begin{aligned}
\frac{d}{d\epsilon} p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1})\Big|_{\epsilon=0} &= p_{O_i}(o_i \mid \bar{o}_{i-1}) \frac{d}{d\epsilon} \log p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1})\Big|_{\epsilon=0} \\
&= S_{O_i}(o) p_{O_i}(o_i \mid \bar{o}_{i-1}) \\
\implies \frac{d}{d\epsilon} p_{O_i,\epsilon}(o_i \mid \bar{o}_{i-1})\Big|_{\epsilon=0} &= \left(\mathbb{E}[S(O) \mid \bar{O}_i = \bar{o}_i] - \mathbb{E}[S(O) \mid \bar{O}_{i-1} = \bar{o}_{i-1}]\right) p_{O_i}(o_i \mid \bar{o}_{i-1})
\end{aligned} \tag{3.14}$$

Parametric connection

Consider a parametric model containing elements P_θ for 1-dimensional θ. Let γ be differentiable with respect to ϵ at $\epsilon = 0$ and $\gamma(0) = \theta$. Let $r = \gamma'(0)$ and regard the path through P_θ defined by $P_{\gamma(\epsilon)}$. If the likelihood, p_θ is differentiable wrt θ, we have for any given o:

$$\text{taylor series} \implies \text{for small } \epsilon$$

$$p_{\gamma(\epsilon)}(o) = p_{\theta + r\epsilon + O(\epsilon^2)}(o) = p_\theta(o) + \frac{dp_\theta}{d\theta}(o) r\epsilon + O(\epsilon^2) \approx p_\theta(o)\left(1 + \epsilon r \frac{d}{d\theta} \log p_\theta(o)\right)$$

We can see the score as the mean 0 function next to the ϵ similarly to the paths for the non-parametric case. Such is really a result of the chain rule where we have $\frac{d}{d\epsilon} \log p_{\gamma(\epsilon)}\Big|_{\epsilon=0} = r\frac{d}{d\theta} \log p_\theta = S_\theta$, the familiar "derivative of log-likelihood" score we know from parametric statistics. Our scores form a 1-dimensional tangent space, $\{r\frac{d}{d\theta}\log p_\theta \text{ s.t. } r \in \mathbb{R}\}$, a subspace of $L_0^2(P_\theta)$, assuming $r\frac{d}{d\theta}\log p_\theta$ is of finite variance. The reader may verify the fact $r\frac{d}{d\theta}\log p_\theta$ has mean 0 with respect to P_θ. Very similar reasoning follows for k-dimensional parametric models, where we will have a k-dimensional tangent space as a subspace of $L_0^2(P_\theta)$, $\{r^T\nabla_\theta \log p_\theta \text{ s.t. } r \in \mathbb{R}^k\}$, that is, all linear combinations of the k partial derivatives.

The Efficient Influence Curve

Consider a parameter mapping on the model, $\mathcal{M}$, which, for simplicity, we will consider as a mapping to the reals given by $\Psi(P)$. We can borrow from van der Vaart, 2000, who defines the pathwise derivative as a continuous linear map from T to the reals given by

$$\lim_{\epsilon\to 0}\left(\frac{\Psi(P_\epsilon)-\Psi(P)}{\epsilon}\right) \longrightarrow \dot{\Psi}_P(S) \tag{3.15}$$

We note to the reader, we imply a direction, S, when we write P_e, which has density $p(1+\epsilon S)$, but generally leave it off the notation as understood.

By the riesz representation theorem (Riesz 1909) for Hilbert Spaces, if the functional defined in (3.15) is a bounded and linear functional on the tangent space, T, it can be written in the form of an inner product $\langle D^*_\Psi(P), S\rangle_{L_0^2(P)} = \int D^*_\Psi(P)(o)S(o)p(o)d\nu(o)$ where $D^*_\Psi(P)$ is a unique element of T, which we call the canonical gradient or **efficient influence curve**. The efficient influence curve is defined at a distribution ,P, according to the parameter mapping, Ψ, and is a function of the data, O.

It is possible to have a gradient not in T if T is a proper subspace $L_0^2(P)$, i.e., it is possible to have a $D(P) \in L_0^2(P)$ such that for all $S \in T$, $\dot{\Psi}_P(S) = \langle D, S\rangle$.
EXERCISE: Prove this element has a larger variance than $D^*(P)$ by using the basic properties of inner products and the uniqueness of $D^*(P)$ in T. Because all regular asymptotically linear estimators have a corresponding gradient, this proves the efficient influence curve has a variance that is the general cramer-rao lower bound for any regular asymptotically linear estimator (van der Vaart 2000).

Parametric connection

Again, returning to our parametric model, define the parameter mapping as $\Psi(P_{\gamma(\epsilon)}) = \gamma(\epsilon)$, for which we let $\gamma'(0) = r$, i.e., assuming differentiability of the parameter mapping in the ordinary sense of introductory calculus. Now we can notice, using the $L_0^2(P)$ norm, $\|f\|^2 = \int f(o)^2 p_\theta(o)d\nu(o)$, which implies the following:

$$r = \frac{\int r(\frac{d}{d\theta}\log p_\theta(o))^2 p_\theta(o)d\nu(o)}{\|\frac{d}{d\theta}\log p_\theta\|^2}$$

$$
\begin{aligned}
&= \int \frac{\frac{d}{d\theta}\log p_\theta(o)}{\|\frac{d}{d\theta}\log p_\theta\|^2} r \underbrace{\frac{d}{d\theta}\log p_\theta(o)}_{\text{the score } S_\theta} p_\theta(o) d\nu(o) \\
&= \int \frac{\frac{d}{d\theta}\log p_\theta(o)}{\|\frac{d}{d\theta}\log p_\theta\|^2} S_\theta(o) d\nu(o) \\
&= \left\langle \frac{\frac{d}{d\theta}\log p_\theta}{\|\frac{d}{d\theta}\log p_\theta\|^2}, S_\theta \right\rangle
\end{aligned}
$$

And thus the efficient influence curve is given by $\frac{\frac{d}{d\theta}\log p_\theta(o)}{\|\frac{d}{d\theta}\log p_\theta\|^2}$, whose variance we can see is the inverse of the Fisher Information, $1/\|\frac{d}{d\theta}\log p_\theta\|^2$, which we know to be the cramer-rao lower bound and attainable via maximum likelihood estimation, under regularity assumptions.

Remark. For a note on regularity, see Kale, 1985, where Hodges classic example of irregularity is discussed.

The General Technique

The general approach to derive the efficient influence curve for a given parameter will be to compute the derivative of the parameter mapping along a path, i.e. compute $\dot{\Psi}_P(S)$ above via taking a derivative and write it as an inner product with the score, S, via use of the key identity (3.14). Since this functional will be bounded and linear for the parameters we encounter, then by the previous paragraph, this will tell us exactly what the efficient influence curve is. Precisely the efficient influence curve will be the function with the score, S, in the inner product, which means the efficient influence curve will be the function multiplied by the score in the integral with respect to P. We will start with easy examples and grow progressively more involved, including influence curves for new parameters derived by the author.

3.1.3 Example 1: $\int F(x)^2 dx$

Let $\Psi(P) = \int F(x)^2 dx$, the parameter mapping for $P \in \mathcal{M}$, the set of continuous distributions, where F is the CDF.

$$
\begin{aligned}
\frac{d}{de}\Psi(P_e)\bigg|_{e=0} &= \frac{d}{de}\int \left(\int_0^x p_e(z)dz\right)^2 dx\bigg|_{e=0} \\
&= \int 2\int \mathbb{I}(z \le x)p(z)dz \frac{d}{de}\int \mathbb{I}(z \le x)p_e(z)dz\bigg|_{e=0} dx \\
&\overset{(3.14)}{=} \int 2F(x)\int \mathbb{I}(z \le x)(\mathbb{E}[S(Z) \mid z] - \mathbb{E}S(Z))p(z)dzdx \\
&= \int 2F(x)\int \mathbb{I}(z \le x)\mathbb{E}[S(Z) \mid z]p(z)dzdx
\end{aligned}
$$

$$- \int 2F(x) \int \mathbb{I}(z \leq x)\mathbb{E}S(Z)p(z)dzdx$$

reverse integration order to write as an integral wrt the density, p

$$= \int \int 2F(x)\mathbb{I}(z \leq x)dxS(z)p(z)dz - \mathbb{E}[S(Z) \int 2F(x)^2dx]$$

$$= \mathbb{E}[S(Z) \int 2F(x)(\mathbb{I}(Z \leq x) - F(x))dx]$$

So the efficient IC is $2 \int F(x)(\mathbb{I}(Z \leq x) - F(x))dx$

3.1.4 Example 2: Treatment Specific Mean

This influence curve is very well-known and can be derived in many ways but it will serve as a good flagship example for the general technique.

STEP 1

Define the data and distribution as well as the factoring: $O = (W, A, Y) \sim P$. P has density, $p(o) = p_Y(y \mid a, w)p_A(a \mid w)p_W(w)$. We will assume A is binary. We also employ the notation, $\bar{Q}(A, W) = \mathbb{E}[Y \mid A, W]$.

STEP 2

Define the parameter as a mapping from $\mathcal{M}$ to the real numbers. $\Psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y \mid A = 1, W]]$

STEP 3

Take derivative of parameter mapping along path in the score direction at P. Write the derivative in terms of a derivative of $p_{Y,e}(y \mid a, w)$ and $p_{W,e}(w)$. Then employ (3.14). We will be very thorough in our steps here.

$$\begin{aligned}
\left.\frac{d}{de}\right|_{e=0} \Psi(P_e) &= \mathbb{E}_{P_e}[\mathbb{E}_{P_e}[Y \mid A = 1, W]] \\
&\overset{dom.convergence}{=} \int \int y \left.\frac{d}{de}\right|_{e=0} (p_{Y,e}(y \mid a = 1, w)d\nu(y)p_{W,e}(w))d\nu(w) \\
&= \int \int y \left.\frac{d}{de}\right|_{e=0} p_{Y,e}(y \mid a = 1, w)d\nu(y)p_W(w)d\nu(w) + \int \int y p_Y(y \mid a = 1, w)d\nu(y) \left.\frac{d}{de}\right|_{e=0} p_{W,e}(w)d\nu(w) \\
&= \int \int \int y \left.\frac{d}{de}\right|_{e=0} p_{Y,e}(y \mid a, w)d\nu(y) \frac{a p_A(a \mid w)}{p_A(a \mid w)} d\nu(a)p_W(w)d\nu(w) \quad (3.16) \\
&\quad + \int \int y p_Y(y \mid a = 1, w)d\nu(y) \left.\frac{d}{de}\right|_{e=0} p_{W,e}(w)d\nu(w) \quad (3.17)
\end{aligned}$$

Now (3.14) establishes the following identities:

$$\begin{aligned}
\frac{d}{d\epsilon} p_{Y^\epsilon}(y \mid a, w)|_{\epsilon=0} &= \left(\mathbb{E}[S(o) \mid w, a, y] - \mathbb{E}[S(W, A, Y) \mid w, a]\right) p_Y(w \mid a, w) \\
&= \left(S(w, a, y) - \mathbb{E}[S(W, A, Y) \mid w, a]\right) p_Y(w \mid a, w) \qquad (3.18) \\
\frac{d}{d\epsilon} p_{W^\epsilon}(w)|_{\epsilon=0} &= \left(\mathbb{E}[S(W, A, Y) \mid w] - \mathbb{E}S(W, A, Y)\right) p_W(w) \qquad (3.19)
\end{aligned}$$

Now we continue from (3.16) and (3.17):

$$\overset{(3.18)\text{ and }(3.19)}{=} \int\int\int y\Big[\mathbb{E}[S(O) \mid w, a, y] - \mathbb{E}[S(O) \mid w, a]p_Y(y \mid w, a)\Big] d\nu(y) \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$+ \int\int yp_Y(y \mid a = 1, w) d\nu(y) \Big[\mathbb{E}[S(O) \mid w] - \mathbb{E}[S(O)]p_W(w)\Big] d\nu(w)$$

Splitting up the first integral and noting $\mathbb{E}[S(O) \mid w, a, y] = S(O)$:

$$= \int\int\int yS(O)p_Y(y \mid w, a) d\nu(y) \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$- \int\int \underbrace{\int y\mathbb{E}[S(O) \mid w, a]p_Y(y \mid w, a) d\nu(y)}_{\text{integrate wrt y}} \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$+ \int \underbrace{\int yp_Y(y \mid a = 1, w) d\nu(y)}_{\text{integrate wrt y}} \Big[\mathbb{E}[S(O) \mid w] - \mathbb{E}[S(O)]p_W(w)\Big] d\nu(w)$$

integrate the 2nd and 3rd integrals wrt y

$$= \int\int\int yS(O)p_Y(y \mid w, a) d\nu(y) \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$- \int\int \bar{Q}(w, a)\mathbb{E}[S(O) \mid w, a] \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$+ \int \bar{Q}(1, w) \Big[\mathbb{E}[S(O) \mid w] - \mathbb{E}[S(O)]p_W(w)\Big] d\nu(w)$$

replacing expectations with integrals we get:

$$= \int\int\int yS(o)p_Y(y \mid w, a) d\nu(y) \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$- \int\int \bar{Q}(1, w) \int S(o)p_Y(y \mid w, a) d\nu(y) \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)$$

$$+ \int \bar{Q}(1, w) \int S(o)p_{YA}(y, a \mid w) d\nu(y, a) p_W(w) d\nu(w) - \int S(o)p(o) d\nu(o) \int \bar{Q}(1, w) p_W(w) d\nu(w)$$

Note the first term becomes a single integral as discussed section 3.1.1

$$= \int yS(o) \frac{a}{p_A(a \mid w)} \underbrace{p_Y(y \mid w, a)p_A(a \mid w)p_W(w)}_{p(o)} d\nu(o)$$

$$- \int\int\int \bar{Q}(1, w)S(o) \underbrace{p_Y(y \mid w, a)) d\nu(y) \frac{ap_A(a \mid w)}{p_A(a \mid w)} d\nu(a) p_W(w) d\nu(w)}_{\frac{a}{p_A(a|w)} p(o) d\nu(y) d\nu(a) d\nu(w)}$$

$$+ \int \bar{Q}(1, w) \int S(o) \underbrace{p_{YA}(y, a \mid w) d\nu(y, a) p_W(w) d\nu(w)}_{p(o) d\nu(y,a) d\nu(w)} - \underbrace{\int S(o)p(o) d\nu(o) \int \bar{Q}(1, w) p_W(w) d\nu(w)}_{\int S(o)p(o)\Psi(P) d\nu(o)}$$

the second and third terms become single integrals (see section 3.1.1) yielding:

$$= \int S(o) \frac{a}{p_A(a \mid w)} yp(o) d\nu(o) - \int S(o) \frac{a}{p_A(a \mid w)} \bar{Q}(w, a) \underbrace{p_Y(y \mid w, a)p_A(a \mid w)p_W(w)}_{p(o)} d\nu(o)$$

$$+ \int S(o)\bar{Q}(1, w)p(o) d\nu(o) - \int S(o)\Psi(P)p(o) d\nu(o)$$

$$= \int S(o)\left[\frac{a}{p_A(a \mid w)}(y - \bar{Q}(w,a)) + \bar{Q}(1,w) - \Psi(P)\right]p(o)d\nu(o)$$

Now we notice the last expression is an $L_0^2(P)$ inner product of the score, S and the function defined by the formula:

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

and $D^*(P)$ is therefore the efficient influence curve, assuming $1/p_A(a \mid w)$ does not blow up anywhere to make derivative functional unbounded.

Remark. If one follows the guidelines of section 3.1.1, the derivation takes care of itself. One should keep one's mind's eye on making sure the full density is under the integral, meaning all factors of the likelihood, so as to have a properly defined $L_0^2(P)$ inner product. There is also the trick of multiplying by $\frac{ap_A(a|w)}{p_A(a|w)}dv(a)$ within the integral so as to be able to write this full density.

Regarding Semi-Parametric Models With Known Treatment Mechanism

Our parameter mapping does not depend on the treatment mechanism, g, and also $T_A \perp T_Y \oplus T_W$ which, means our efficient influence curve must therefore be in $T_Y \oplus T_W$ for the nonparametric model. Therefore, our efficient influence curve will have two orthogonal components in T_Y and T_W respectively. We have no component in T_A, which is why we need not perform a TMLE update of the initial prediction, g_n, of g_0. Such also teaches us that for the semi-parametric model, where the treatment mechanism is known, the efficient influence function will remain the same.

3.1.5 Example 3: Efficient Influence Curve of TE Variance, VTE

Let $P \in \mathcal{M}$, non-parametric for the same data structure as in section 3.1.4. Then define $b_P(W) = \mathbb{E}_P[Y \mid A = 1, W] - \mathbb{E}_P[Y \mid A = 0, W]$. We note, this also covered in Levy, 2018 tech report on the VTE (Levy et al. 2018).

Theorem 3.1.1. *Let $\Psi(P) = var_P(b(W))$. The efficient influence curve for Ψ at P is given by:*

$$\boldsymbol{D^\star(P)(W,A,Y) = 2\left(b(W) - \mathbb{E}b(W)\right)\left(\frac{2A-1}{p_A(A|W)}\right)\left(Y - \bar{Q}(A,W)\right) + \left(b(W) - \mathbb{E}b\right)^2 - \Psi(P)}$$

where $\bar{Q}(A,W) = \mathbb{E}(Y|A,W)$

Proof.

$$\begin{aligned}
&\frac{d}{d\epsilon}\Psi(P_\epsilon)(S)\bigg|_{\epsilon=0} \\
=&\frac{d}{d\epsilon}\mathbb{E}_{P_\epsilon}\bigg(b_{P_\epsilon}(W)-\mathbb{E}_{P_\epsilon}b_{P_\epsilon}(W)\bigg)^2\bigg|_{\epsilon=0} \\
=&\frac{d}{d\epsilon}\int\bigg(b_{P_\epsilon}(w)-\mathbb{E}_{P_\epsilon}b_{P_\epsilon}(W)\bigg)^2 p_\epsilon(o)d\nu(o)\bigg|_{\epsilon=0} \\
=&\int 2\bigg(b_{P_\epsilon}(w)-\mathbb{E}_{P_\epsilon}b_{P_\epsilon}(W)\bigg)\frac{d}{d\epsilon}\bigg(b_{P_\epsilon}(w)-\mathbb{E}_{P_\epsilon}b_{P_\epsilon}(W)\bigg)p(o)d\nu(o)\bigg|_{\epsilon=0} \\
&+\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)^2\frac{d}{d\epsilon}p_{W,\epsilon}(w)\bigg|_{\epsilon=0}d\nu(w)
\end{aligned}$$

note that $\displaystyle\int 2\bigg(b_{P_\epsilon}(w)-\mathbb{E}_{P_\epsilon}b_{P_\epsilon}(W)\bigg)\frac{d}{d\epsilon}\left(\mathbb{E}_{P_\epsilon}b_{P_\epsilon}(W)\right)p(o)d\nu(o)\bigg|_{\epsilon=0}=0$ so we have:

$$\begin{aligned}
\overset{(3.19)}{=}&\int 2\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\frac{d}{d\epsilon}b_{P_\epsilon}(w)p(o)d\nu(o)\bigg|_{\epsilon=0} \\
&+\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)^2\left(\mathbb{E}[S(W,A,Y)\mid w]-\mathbb{E}S(W,A,Y)\right)p_W(w)d\nu(w) \\
=&2\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\frac{d}{d\epsilon}\bigg[\int\bigg(yp_{Y\epsilon}(y|a=1,w)-yp_{Y\epsilon}(y|a=0,w)\bigg)d\nu(y)\bigg]p_W(w)d\nu(w)\bigg|_{\epsilon=0} \\
&+\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)^2\int S(o)p_{Y,A}(y,a\mid w)d\nu(y,a)p_W(w)d\nu(w) \\
&-\int S(o)\Psi(P)p(o)d\nu(o) \\
=&2\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\underbrace{\int\bigg(y\frac{d}{d\epsilon}p_{Y\epsilon}(y|a,w)\bigg|_{\epsilon=0}\frac{2a-1}{p_A(a|w)}p_A(a|w)d\nu(y,a)}_{\text{trick}}p_W(w)d\nu(w) \qquad (3.20) \\
&+\int\bigg[\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)^2-\Psi(P)\bigg]S(o)p(o)d\nu(o)
\end{aligned}$$

Now continuing with the term (3.20).

$$\begin{aligned}
\overset{(3.18)}{=}&2\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\bigg[\int y\bigg(\mathbb{E}_P[S(O)\mid y,a,w] \\
&-\mathbb{E}_P[S(O)\mid a,w]\bigg)p_Y(y\mid a,w)\frac{2a-1}{p_A(a|w)}p_A(a|w)d\nu(y,a)\bigg]p_W(w)d\nu(w) \\
&\text{splitting into separate integrals} \\
=&2\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\underbrace{\int S(o)yp_Y(y\mid a,w)\frac{2a-1}{p_A(a|w)}p_A(a|w)d\nu(y,a)}_{\text{an integral wrt a,y}}p_W(w)d\nu(w) \\
&-2\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\int\underbrace{\int yp_y(y|a,w)d\nu(y)}_{\bar{Q}(a,w)}\mathbb{E}_P[S(O)\mid a,w]\frac{2a-1}{p_A(a|w)}p_A(a|w)d\nu(a)p_W(w)d\nu(w) \\
&\text{replace expectations with integrals} \\
&2\int\bigg(b_P(w)-\mathbb{E}_P b_P(W)\bigg)\underbrace{\int S(o)yp_Y(y\mid a,w)\frac{2a-1}{p_A(a|w)}p_A(a|w)d\nu(y,a)}_{\text{an integral wrt a,y}}p_W(w)d\nu(w)
\end{aligned}$$

$$-2\int\Big(b_P(w)-\mathbb{E}_Pb_P(W)\Big)\int\bar{Q}(a,w)\int S(o)p_Y(y\mid a,w)d\nu(y)\frac{2a-1}{p_A(a|w)}p_A(a|w)d\nu(a)p_W(w)d\nu(w)$$

$$\overset{fubini}{=}2\int\Big(b_P(w)-\mathbb{E}_Pb_P(W)\Big)\frac{(2a-1)}{p_A(a|w)}yS(o)p(o)d\nu(o)-2\int\Big(b_P(w)-\mathbb{E}_Pb_P(W)\Big)\frac{(2a-1)}{p_A(a|w)}\bar{Q}(a,w)S(o)p(o)d\nu(o)$$

$$=2\int\Big(b_P(w)-\mathbb{E}_Pb_P(W)\Big)\frac{(2a-1)}{p_A(a|w)}(y-\bar{Q}(a,w))S(o)p(o)d\nu(o)$$

And we can see the unique reisz representer (the function in the $L_0^2(P)$ inner product with the score, S) is given by

$$2\left(b(W)-\mathbb{E}b(W)\right)\left(\frac{2A-1}{p_A(A|W)}\right)(Y-\bar{Q}(A,W))+\left(b(W)-\mathbb{E}b\right)^2-\Psi(P)$$

completing the proof. □

Remark. From here on out we will avoid the double and triple integrals and take them as understood because otherwise the notation is too clumsy.

3.1.6 Example 4: Affect Among the Treated

We have the identical data structure as before. However, to avoid confusion and maintain notation, we will factor the density as follows:
$p(w,a,y)=p_Y(y\mid a,w)g(a\mid w)p_W(w)$ so $g(a\mid w)$ takes the place of $p_A(a\mid w)$. We will use P_A to be the marginal density of A, which is binary. Thus the score $\frac{d}{de}p_{A,e}\Big|_{e=0}=S_{Amarginal}(a)p_A(a)$ as in the step before establishing, the key identity, (3.14). But then we see the obvious that the score for a binary marginal is just $\mathbb{I}(A=a)-p_A(a)$, so we get $\frac{d}{de}p_{A,e}\Big|_{e=0}=(\mathbb{I}(A=a)-p_A(a))p_A(a)$ and this can be used below when we take derivatives. The author will start the reader with a couple of crucial steps. First

$$\Psi(P)=\mathbb{E}_P[(\mathbb{E}_P[Y\mid 1,W]-\mathbb{E}_P[Y\mid 0,W])\mid A=1]$$

The efficient influence curve is given in van der Laan and Rose, 2011 as

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

The reader is encouraged to derive this fact after being given a few first steps as follows: We write the parameter mapping as an integral for a path along score, S, whose notation is supressed here as usual. S will appear later when we apply (3.14).
$\Psi(P_e)=\int y(p_{Y,e}(y\mid 1,w)-p_{Y,e}(y\mid 0,w))\frac{g_e(0|w)p_{W,e}(w)}{p_{A,e}(0)}dv(y,w)$
and when you differentiate at $e=0$ you get four terms: $\frac{d}{de}\int y(p_{Y,e}(y\mid 1,w)-p_{Y,e}(y\mid 0,w))\frac{g(0|w)p_W(w)}{p_A(0)}dv(y,w)\Big|_{e=0}$
$\frac{d}{de}\int y(p_Y(y\mid 1,w)-p_Y(y\mid 0,w))\frac{g_e(0|w)p_W(w)}{p_A(0)}dv(y,w)\Big|_{e=0}$

$\frac{d}{de} \int y(p_Y(y \mid 1, w) - p_Y(y \mid 0, w)) \frac{g(0|w) p_{W,e}(w)}{p_A(0)} dv(y, w) \Big|_{e=0}$

$\frac{d}{de} \int y(p_Y(y \mid 1, w) - p_Y(y \mid 0, w)) \frac{g(0|w) p_W(w)}{p_{A,e}(0)} dv(y, w) \Big|_{e=0}$

Any density that is being differentiated must be rewritten in its full conditional form, i.e., without any specific numbers in the conditional so you have $p_{Y,e}(y \mid a, w), p_{A,e}(a \mid w), p_{W,e}(w)$ and $p_{A,e}(a)$. Thus we apply the usual trick to do so:

$\frac{d}{de} \int y(p_{Y,e}(y \mid a, w) \frac{(2a-1)g(a|w)}{g(a|w)} \frac{g(0|w) p_W(w)}{p_A(0)} dv(y, a, w) \Big|_{e=0}$

$\frac{d}{de} \int p_Y(y \mid a, w)(\bar{Q}(1, w) - \bar{Q}(0, w)) g_e(a \mid w) \frac{a}{g(a|w)} \frac{p_W(w)}{p_A(0)} dv(y, a, w) \Big|_{e=0}$

$\frac{d}{de} \int p_Y(y \mid a, w)(\bar{Q}(1, w) - \bar{Q}(0, w)) g(a \mid w) \frac{g(0|w) p_{W,e}(w)}{p_A(0)} dv(y, a, w) \Big|_{e=0}$

$\frac{d}{de} \int y p_Y(y \mid a, w)(\bar{Q}(1, w) - \bar{Q}(0, w)) p_{W|A}(w \mid a) \frac{a p_A(a)}{p_{A,e}(a)} dv(y, a, w) \Big|_{e=0}$

Now the reader is ready to proceed and carefully integrate, using (3.14).

3.1.7 Example 5: Efficient Influence Curve for Transporting Stochastic Direct and Indirect Effects Non-parametric Model

Here we consider data of the form $O = (YS, M, Z, A, W, S)$ where we consider M, Z, A, S as binaries and W as a vector of covariates. YS indicates we only see an outcome for when $S = 1$, i.e., for when the site of our population is taken from site 1. The observed data likelihood factors as below, assuming the non-parametric model.

$$p(O) = p_{Y \times S}(Y \times S \mid M, Z, A, W, S) g_M(M \mid Z, A, W, S) p_Z(Z \mid A, W, S) g_A(A \mid W, S) p_{W|S}(W \mid S) p_S(S)$$

We perform an intervention on A for a population at both sites, S = 1 and 0. Z can be considered an intermediate confounder and M, a mediator. Here we consider a data adaptive parameter where $\hat{g}_{M|a^*,W,s}(m \mid W) = \sum_z \hat{g}_M(M \mid z, a^*, W, s)(M \mid W)$ is the stochastic intervention on M marginalized over Z and defined for a fixed value of $A = a^*$ and $S = s$. $\hat{g}_{M|a^*,W,s}$ can be considered as estimated from the data and thus, it can be considered as a given. That is, it defines the parameter below data adaptively.

Theorem 3.1.2. *Consider a non-parametric model or semiparametric model with one or both the treatment and mediator mechanisms known (mechanisms for A and M). Consider the parameter defined by*

$$\Psi(P) = \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sum_m \Bigg[\mathbb{E} Y \hat{g}_{M|a^*,W,s}(m \mid W) \mid M = m, W, Z, A = a, S = 1\Bigg] \mid A = a, W, S\Bigg] \mid S = 0\Bigg]$$

where the expectations are taken with respect to P. Then the efficient influence curve is given by

$$D^*(P)(O) = D^*_Y(P)(O) + D^*_Z(P)(O) + D^*_W(P)(O)$$

where

$$
\begin{aligned}
D_Y^*(P)(O) &= (Y - \mathbb{E}\left[Y \mid M, Z, A, W\right]) * \\
&\quad \frac{\hat{g}_{M|a^*,W,s}\left(M \mid W\right) p_Z\left(Z \mid A, W, S=0\right) p_{S|W}\left(S=0 \mid W\right) I(S=1, A=a)}{g_M\left(M \mid Z, A, W, S\right) p_Z\left(Z \mid A, W, S\right) g_A\left(A \mid W, S\right) p_{S|W}\left(S \mid W\right) P_S(S=0)} \\
D_Z^*(P)(O) &= \left(\bar{Q}_M(Z, A, W) - \bar{Q}_Z(A, W, S)\right) \frac{I(S=0, A=a)}{g_A\left(A \mid W, S\right) p_S(S=0)} \\
D_W^*(P)(O) &= \left(\bar{Q}_Z(A=a, W, S) - \Psi(P)\right) \frac{I(S=0)}{p_S(S=0)}
\end{aligned}
$$

Proof. (3.14) implies the following, replacing our usual score name, S, currently occupied by the site variable, S, with γ:

$$\frac{d}{d\epsilon}\left(p_{Y,\epsilon}(Y \times S \mid M, Z, A, W, S)\right)\Big|_{\epsilon=0} = \left(\gamma(O) - \mathbb{E}\left[\gamma(O) \mid M, Z, A, W, S\right]\right) p_Y(Y \times S \mid M, Z, A, W, S) \tag{3.21}$$

$$\frac{d}{d\epsilon}\left(p_{Z,\epsilon}(Z \mid A, W, S)\right)\Big|_{\epsilon=0} = \left(\mathbb{E}\left[\gamma(O) \mid Z, A, W, S\right] - \mathbb{E}\left[\gamma(O) \mid A, W, S\right]\right) p_Z(Z \mid A, W, S) \tag{3.22}$$

$$\frac{d}{d\epsilon}\left(p_{W|S,\epsilon}(W \mid S)\right)\Big|_{\epsilon=0} = \left(\mathbb{E}\left[\gamma(O) \mid W, S\right] - \mathbb{E}\left[\gamma(O) \mid S\right]\right) p_{W|S}(W \mid S) \tag{3.23}$$

Our parameter of interest is given by

$$\Psi(P) = \int y p_Y(y \mid m, z, a, w, s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid a, w, s=0\right) p_{W|S}\left(w \mid s=0\right) d\nu(y, m, z, w)$$

We then take the pathwise derivative for a path along score, γ. We can note to the reader that this derivative is unaffected by knowledge of the treatment mechanism, $E[A \mid S, W]$, or the mediator mechansim, $E[M \mid Z, A, W, S]$, due to the estimand not depending on these models as well as the fact that scores, γ_A and γ_M are orthogonal (have 0 covariance) to $\gamma_Y, \gamma_Z, \gamma_W)$ in the Hilbert Space $L^2(P)$. This is why for a semi-parametric model where the M and/or A mechanisms are known, the efficient influence curve will be the same as that for the non-parametric model.

$$
\begin{aligned}
\frac{d}{d\epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}\int y p_{Y,\epsilon}(y \mid m, z, a, w, s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_{Z,\epsilon}\left(z \mid a, w, s=0\right) p_{W|S,\epsilon}\left(w \mid s=0\right) d\nu(y, m, z, w)\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\int y p_{Y,\epsilon}(y \mid m, z, a, w, s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid a, w, s=0\right) p_{W|S}\left(w \mid s=0\right) d\nu(y, m, z, w)\Big|_{\epsilon=0} \\
&+ \frac{d}{d\epsilon}\int y p_Y(y \mid m, z, a, w, s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_{Z,\epsilon}\left(z \mid a, w, s=0\right) p_{W|S}\left(w \mid s=0\right) d\nu(y, m, a, z, w)\Big|_{\epsilon=0} \\
&+ \frac{d}{d\epsilon}\int y p_Y(y \mid m, z, a, w, s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid a, w, s=0\right) p_{W|S\epsilon}\left(w \mid s=0\right) d\nu(y, m, z, w)\Big|_{\epsilon=0}
\end{aligned} \tag{3.24}
$$

The first term in 3.24:

$$
\begin{aligned}
&\frac{d}{d\epsilon}\int y p_{Y,\epsilon}(y \mid m, z, a, w, s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid x=a, w, s=0\right) p_{W|S}\left(w \mid s=0\right) d\nu(y, m, z, w)\Big|_{\epsilon=0} \\
&= \int y \frac{d}{d\epsilon} p_{Y,\epsilon}((y \times s) \mid m, z, x, w, s)\Big|_{\epsilon=0} \hat{g}_{M|a^*,W,s}\left(m \mid w\right) \frac{g_M\left(m \mid z, w, s\right)}{g_M\left(m \mid z, w, s\right)} p_Z\left(z \mid x=a, w, s=0\right) \frac{p_Z\left(z \mid x, w, s\right)}{p_Z\left(z \mid x, w, s\right)}
\end{aligned}
$$

$$
\begin{aligned}
&* \frac{I(s=1,x=a) g_A\left(x \mid w,s\right)}{g_A\left(x \mid w,s\right)} p_{W|S}\left(w \mid s=0\right) \frac{p_{W|S}\left(w \mid s\right)}{p_{W|S}\left(w \mid s\right)} \frac{p_S(s)}{p_S(s=1)} d\nu(y,m,z,x,w,s) \\
\overset{(3.21)}{=} &\int y\left(\gamma(o) - \mathbb{E}\left[\gamma(o) \mid m,z,x,w,s\right]\right) p_Y\left((y,s) \mid m,z,w,s\right) \hat{g}_{M|a^*,W,S}\left(m \mid w\right) \frac{g_M\left(m \mid z,w,s\right)}{g_M\left(m \mid z,w,s\right)} p_Z\left(z \mid x=a,w,s=0\right) \\
&* \frac{p_Z\left(z \mid x,w,s\right)}{p_Z\left(z \mid x,w,s\right)} \frac{I(s=1,x=a) g_A\left(x \mid w,s\right)}{g_A\left(x \mid w,s\right) P_S(s=1)} p_{W|S}\left(w \mid s=0\right) \frac{p_{W|S}\left(w \mid s\right)}{p_{W|S}\left(w \mid s\right)} p_S(s) d\nu(y,m,z,x,w,s) \\
= &\int \gamma(o)\Big(y - \mathbb{E}\Big[y \mid m,z,x,w,s\Big]\Big) \times \\
&\frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a,w,s=0) p_{S|W}(s=0 \mid w) I(s=1,x=a)}{g_M(m \mid z,w,s=1) p_Z(z \mid x=a,w,s=1) g_A(a \mid w,s=1) p_{S|W}(s=1 \mid w) p_S(s=0)} p(o) d\nu(o) \\
= &\left\langle \gamma, D_Y^*(P) \right\rangle_{L_0^2(P)}
\end{aligned}
$$

where

$$
D_Y^*(P)(O) = \left(Y - \mathbb{E}\left[Y \mid M,Z,A,W\right]\right) \frac{\hat{g}_{M|a^*,W,s}\left(M \mid W\right) p_Z\left(Z \mid A,W,S=0\right) p_{S|W}\left(S=0 \mid W\right) I(S=1,A=a)}{g_M\left(M \mid Z,A,W,S\right) p_Z\left(Z \mid A,W,S\right) g_A\left(A \mid W,S\right) p_{S|W}\left(S \mid W\right) P_S(S=0)}
$$

The reader may notice $D_Y^*(P)(O)$ is not a mean 0 function of $Y \mid M,Z,W$ because it also depends on the variable, A. Hence, it is not an element of the tangent space under the restricted model where the mechanism for M and Y do not depend directly on A. Therefore, $D^*(P)(O)$ has an extra orthogonal component in addition to the efficient influence curve for the restricted model so any efficiently constructed estimator based on this influence curve will not be efficient for the restricted semi-parametric model. The second term in 3.24:

$$
\begin{aligned}
&\frac{d}{d\epsilon}\int y p_Y(y \mid m,z,w,s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_{Z,\epsilon}\left(z \mid a,w,s=0\right) p_{W|S}\left(w \mid s=0\right) d\nu(y,m,z,w)\Big|_{\epsilon=0} \\
= &\int y p_Y(y \mid m,z,x,w) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) \frac{d}{d\epsilon} p_{Z,\epsilon}\left(z \mid x,w,s\right)\Big|_{\epsilon=0} \frac{I(s=0)I(x=a)}{g_A\left(x \mid w,s\right) p_S(s=0)} \\
&* g_A\left(x \mid w,s\right) p_{W|S}\left(w \mid s\right) p_S(s) d\nu(y,m,z,x,w,s) \\
\overset{(3.22)}{=} &\int y p_Y(y \mid m,x,z,w) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) \left(\mathbb{E}\left[\gamma(o) \mid z,x,w,s\right] - \mathbb{E}\left[\gamma(o) \mid x,w,s\right]\right) p_Z(z \mid x,w,s) \\
&* \frac{I(s=0)I(x=a)}{g_A\left(x \mid w,s\right) p_S(s=0)} g_A\left(x \mid w,s\right) p_{W|S}\left(w \mid s\right) p_S(s) d\nu(y,m,z,x,w,s) \\
= &\int \gamma(o)\Big(\mathbb{E}_{\hat{g}_{M|a^*,W,s}}\Big(\mathbb{E}\Big[Y \mid M,A,Z,W\Big] \mid z,x,w,s=1\Big) - \\
&\mathbb{E}_{P_{Z|A,W,S}}\Big[\mathbb{E}_{\hat{g}_{M|a^*,W,s}}\Big(\mathbb{E}\Big[Y \mid M,Z,AW\Big] \mid Z,A,W,S=1\Big) \mid x,w,s\Big]\Big) * \frac{I(s=0,x=a)}{g_A\left(x \mid w,s\right) p_S(s=0)} p(o) d\nu(o) \\
= &\left\langle \gamma, D_Z^*(P) \right\rangle_{L_0^2(P)}
\end{aligned}
$$

We substitute

$$
\begin{aligned}
\bar{Q}_M(z,x,w) &= \mathbb{E}_{\hat{g}_{M|a^*,W,s}}\left(\mathbb{E}\left[Y \mid M,A,Z,W\right] \mid z,x,w\right) \\
\bar{Q}_Z(x,w,s) &= \mathbb{E}_{P_{Z|A,W,S}}\Big[\mathbb{E}_{\hat{g}_{M|a^*,W,s}} \bar{Q}_M(Z,A,W) \mid x,w,s\Big]
\end{aligned}
$$

and since x represents the treatment, A, in the integrals above, we get
$\mathbf{D^*_Z(P)(O)} = \left(\mathbf{\bar{Q}_M(Z,A,W)} - \mathbf{\bar{Q}_Z(A,W,S)}\right) \frac{\mathbf{I(S=0,A=a)}}{\mathbf{g_A(A|W,S)p_S(S=0)}}$
The third term in 3.24:

$$
\frac{d}{d\epsilon}\int y p_Y(y \mid m,z,a,w,s=1) \hat{g}_{M|a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid a,w,s=0\right) p_{W|S,\epsilon}\left(w \mid s=0\right) d\nu(y,m,z,w)\Big|_{\epsilon=0}
$$

$$
\begin{aligned}
&= \int y p_Y(y \mid m,a,z,w) \hat{g}_{M\mid a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid a,w,s\right) \frac{d}{d\epsilon} p_{W\mid S,\epsilon}\left(w \mid s\right)\Big|_{\epsilon=0} \frac{I(s=0)}{p_S(s=0)} p_S(s) d\nu(y,m,z,x,w,s) \\
&\overset{(3.23)}{=} \int y p_Y(y \mid m,a,z,w) \hat{g}_{M\mid a^*,W,s}\left(m \mid w\right) p_Z\left(z \mid a,w,s\right) \left(\mathbb{E}\left[\gamma(o) \mid w,s\right] - \mathbb{E}\left[\gamma(o) \mid s\right]\right) \\
&\quad * p_{W\mid S}(w \mid s) \frac{I(s=0)}{p_S(s=0)} p_S(s) d\nu(y,m,z,x,w,s) \\
&= \int S(o) \left(\bar{Q}_Z(x=a,w,s) - \Psi(P)\right) \frac{I(s=0)}{p_S(s=0)} p(o) d\nu(o) \\
&= \langle \gamma, D^*_W \rangle_{L^2_0(P)}
\end{aligned}
$$

where $\mathbf{D^*_W(P)(O) = \left(\bar{Q}_Z(A=a,W,S) - \Psi(P)\right) \frac{I(S=0)}{p_S(S=0)}}$
Thus the efficient influence curve is the sum of its orthogonal components:

$$D^*(P)(O) = D^*_Y(P)(O) + D^*_Z(P)(O) + D^*_W(P)(O)$$

□

Regarding Semi-Parametric Models With Known Treatment and Mediator Mechanisms

Our parameter mapping does not depend on the treatment mechanism g or the mediator mechanism, g_M. Also, T_A, the tangent space of mean 0 functions of A given W, S, as well as T_M, the tangent space of mean 0 functions of M given Z, A, W, S are both perpendicular to the subspace of the tangent space containing D^*_W, D^*_Z and D^*_Y. Thus we would not perform a TMLE update of the initial fits for the mediator mechanism, $E[M \mid Z, A, W, S]$, and the treatment mechanism, $E[A \mid W, S]$, just as we would not do so for the treatment mechanism at any time point for a longitudinal TMLE for the treatment specific mean (van der Laan and Rose 2011). If our stochastic intervention was not data adaptive in the sense that our parameter depends on our fit of $\hat{g}_{M\mid a^*,W,s}$ but rather we defined $\hat{g}_{M\mid a^*,W,s}$ as the fixed true mechanism we were estimating from the data, then our parameter mapping would depend on g_M and we would have a component in the efficient influence curve in T_M.

3.1.8 Example 6: Efficient Influence Curve for Transporting Stochastic Direct and Indirect Effects Restricted Model

Now we will derive the efficient influence curve for same parameter as the previous section, except, we will assume the restricted semi-parametric model where M and Y mechanism do not depend directly on the instrument, A.

Theorem 3.1.3. *The efficient influence curve for our restricted model, where M and Y do not depend directly on A, is given by*

$$D^*(P)(O) = D^*_{Y,r}(P)(O) + D^*_Z(P)(O) + D^*_W(P)(O)$$

where

$$D^*_{Y,r}(P)(O) = \left(y - \mathbb{E}\left[y \mid m, z, w\right]\right) \frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a_0, w, s = 0) p_{S|W}(s = 0 \mid w) I(s = 1)}{g_{M,r}(m \mid z, w, s) p_Z(z \mid w, s) p_{S|W}(s \mid w) p_S(s = 0)}$$

Proof. We can note that our only task here is to project $D^*_Y(P)$, our component of the influence curve in T_Y, onto the subspace of T_Y given by
$T_{Y,r} = \overline{\{\gamma(O \mid YS, M, Z, W, S) : \mathbb{E}(\gamma(O) \mid YS, M, Z, W, S) = 0\}}$.
Consider observed data density argument $o = (ys, m, z, x, w, s)$ and the corresponding random variable, $O = (YS, M, Z, A, W, S)$, where we will retain the variable ordering in all the notation. p_{YS} is the conditional density ys given m, z, a, w, s and p_M is the conditional density of m given z, x, w, s, i.e., given all varibles to the right and we will stay with that convention in naming other densities. Note, a is fixed here (the intervention on A) and x is the variable for the treatment in the density (playing the role of random variable A). $p_{YS,r}$ is the conditional density ys given m, z, w and $p_{M,r}$ is the conditional density of m given z, w, s in the restricted model, i.e. we don't put the instrument, a, in the conditional statement. $\bar{M} = m, z, x, w, s$, i.e. all past variables from most recent backward.
Notice the following:

$$\begin{aligned} p_{A|\bar{Y}S,r}(x \mid ys, m, z, w, s = 1) &= \frac{p_{\bar{A},r}(x, ys, m, z, w, s = 1)}{p_{O/A}(ys, m, z, w, s = 1)} \\ &= \frac{p_{Y,r}(y \mid m, z, w) p_{M,r}(m \mid z, w, s = 1) p_{\bar{Z}}(z, x, w, s)}{p_{Y,r}(y \mid m, z, w) p_{M,r}(m \mid z, w, s = 1) p_{\bar{Z}}(z, w, s = 1)} \\ &= \frac{p_{\bar{Z}}(z, x, w, s = 1)}{p_{\bar{Z}/A}(z, w, s = 1)} \end{aligned} \tag{3.25}$$

$$\begin{aligned} p_{A,YS,r}(x, ys \mid m, z, w, s = 1) &= \frac{p_{\bar{Y},r}(ys, x, m, z, w, s = 1)}{p_{\bar{M},r}(m, z, w, s = 1)} \\ &= \frac{p_{Y,r}(y \mid m, z, w) p_{\bar{Z}}(z, x, w, s = 1)}{p_{\bar{Z}/A}(z, w, s = 1)} \end{aligned} \tag{3.26}$$

Thus from 3.25 and 3.26 and referencing item 4 in section 3.1.2:

$$\begin{aligned} &\prod(D^*_Y \| T_{Y,r}) \\ &= \mathbb{E}(D^*_Y(O) \mid YS, M, Z, W, S) - \mathbb{E}(D^*_Y(O) \mid M, Z, W, S) \\ &= \mathbb{E}(D^*_Y(O) \mid YS, M, Z, W, S) \\ &= \int \left(y - \mathbb{E}\left[y \mid m, z, w\right]\right) \times \\ &\frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s = 0) p_{S|W}(s = 0 \mid w) I(s = 1, x = a)}{g_{M,r}(m \mid z, w, s = 1) p_Z(z \mid a, w, s = 1) g_A(a \mid w, 1) p_{S|W}(1 \mid w) p_S(0)} p_{A|\bar{Y}S,r}(x \mid ys, m, z, w, s) d\nu(x) \\ &- \int \left(y - \mathbb{E}\left[y \mid m, z, w\right]\right) \times \\ &\frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s = 0) p_{S|W}(s = 0 \mid w) I(s = 1, x = a)}{g_{M,r}(m \mid z, w, s = 1) p_Z(z \mid a, w, s = 1) g_A(a \mid w, 1) p_{S|W}(1 \mid w) p_S(0)} p_{A|\bar{Y}S,r}(x, ys \mid m, z, w, s) d\nu(x, ys) \end{aligned}$$

remembering we are integrating wrt x and all else is fixed in the first integral
All is fixed but x and ys in the second integral. Since I(s=1), ys = 1 and s = 1

$$= \int \left(y - \mathbb{E}\left[y \mid m, z, w \right] \right) \times$$

$$\frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s=0) p_{S|W}(s=0 \mid w) I(s=1, x=a)}{g_{M,r}(m \mid z, w, s=1) p_Z(z \mid a, w, s=1) g_A(a \mid w, 1) p_{S|W}(1 \mid w) p_S(0)} p_{A|Y\bar{S},r}(x \mid ys, m, z, w, s=1) d\nu(x)$$

$$- \int \left(y - \mathbb{E}\left[y \mid m, z, w \right] \right) \times$$

$$\frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s=0) p_{S|W}(s=0 \mid w) I(s=1, x=a)}{g_{M,r}(m \mid z, w, s=1) p_Z(z \mid a, w, s=1) g_A(a \mid w, 1) p_{S|W}(1 \mid w) p_S(0)} p_{A|Y\bar{S},r}(x, ys \mid m, z, w, s=1) d\nu(x, ys)$$

use (3.25) and (3.26) for the 1st and 2nd integrals respectively, which kills the 2nd integral:

$$= \int \left(y - \mathbb{E}\left[y \mid m, z, w \right] \right) \times$$

$$\frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s=0) p_{S|W}(s=0 \mid w) I(s=1, x=a)}{g_{M,r}(m \mid z, w, s=1) p_Z(z \mid a, w, s=1) g_A(a \mid w, 1) p_{S|W}(1 \mid w) p_S(0)} \frac{p_{\bar{Z}}(z, x, w, s=1)}{p_{\bar{Z}}(z, w, s=1)} d\nu(x)$$

$$- \underbrace{\int \left(y - \mathbb{E}\left[y \mid m, z, w \right] \right) p_{Y,r}(y \mid m, z, w) d\nu(y)}_{\text{is } 0} \times$$

$$\int \frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s=0) p_{S|W}(s=0 \mid w) I(s=1, x=a)}{g_{M,r}(m \mid z, w, s=1) p_Z(z \mid a, w, s=1) g_A(a \mid w, 1) p_{S|W}(1 \mid w) p_S(0)} \frac{p_{\bar{Z}}(z, x, w, s=1)}{p_{\bar{Z}/A}(z, w, s=1)} d\nu(x)$$

$$= \left(y - \mathbb{E}\left[y \mid m, z, w \right] \right) \frac{\hat{g}_{M|a^*,W,s}(m \mid w) p_Z(z \mid a, w, s=0) p_{S|W}(s=0 \mid w) I(s=1)}{g_{M,r}(m \mid z, w, s) p_{Z|W,S}(z \mid w, s) p_{S|W}(s \mid w) p_S(s=0)}$$

And the proof is complete since the other components of the unrestricted model's influence curve will remain the same. The reader may note that $p_{Z|W,S}(z \mid w, s) = p_Z(z \mid 1, w, s) g_A(1 \mid w, s) + p_Z(z \mid 0, w, s) g_A(0 \mid w, s)$, so we need not perform any additional regressions for this restricted model.

□

3.1.9 Example 7: Efficient Influence Curve for Transporting Stochastic Direct, Fixed Parameter, Non-parametric Model

According to our general technique of section 3.1.2, our observed data is of the form, $O_6, O_5, ..., O_1 = YS, M, Z, A, W, S$, and thus our we will have corresponding orthogonal tangent spaces $T_{YS}, T_M, T_Z, T_A, T_W, T_S$. The orthogonality and the fact our parameter mapping does not depend on the treatment mechanism g_A, tells us the efficient influence curve for the unrestricted model, which is non-parametric, will be the same as for the model with a known treatment mechanism.

Let us define our parameter by the mapping from the observed data model to the real numbers by $\Psi_f(P)$ and retain the identical definition as in theorem 3.1.2 but bear in mind we are including the true $\hat{g}_{M|a^*,W,s^*}$ in the definition. Therefore our parameter of interest depends on the true models for P_Z and P_M. Thus the efficient influence curve for this parameter in both the unrestricted and restricted models will have components in the tangent space subspace, T_M and an additional component in T_Z to what we had before for the data adaptive parameter. In other words, this parameter is fixed, not data adaptive as in the previous two examples.

Theorem 3.1.4. *The efficient influence curve for the unrestricted model at distribution, P, is given by $D^*_f(P) = D^*_{f,Y}(P) + D^*_{f,M}(P) + D^*_{f,Z}(P) + D^*_{f,W}(P)$ where*

$$\begin{aligned}
D^*_{f,Y}(P) &= D^*_Y(P) \\
D^*_{f,M}(P) &= (M - g_M(1 \mid Z, A, W, S))\frac{(\bar{Q}_{a,0}(1, W) - \bar{Q}_{a,0}(0, W))p_{S|W}(0 \mid W)\mathbb{I}(A = a^*, S = s^*)}{g_A(A \mid W, S)p_{S|W}(S \mid W)P(S = 0)} \\
D^*_{f,Z}(P) &= D^*_Z(P) \\
&\quad + (Z - p_Z(1 \mid A, W, S))\frac{(\bar{Q}^Z_{a,0}(1, A, W, S) - \bar{Q}^Z_{a,0}(0, A, W, S))p_{S|W}(0 \mid W)\mathbb{I}(A = a^*, S = s^*)}{g_A(A \mid W, S), p_{S|W}(S \mid W)P(S = 0)} \\
D^*_{f,W}(P) &= D^*_W(P)
\end{aligned}$$

D^*_Y, D^*_Z and D^*_W are the same as for the data adaptive parameter and we define

$$\begin{aligned}
\bar{Q}(M, Z, A, W) &= \mathbb{E}[Y \mid M, Z, A, W] \\
\bar{Q}_{a,0}(M, W) &= \sum_z \bar{Q}(M, z, a, W)p_Z(z \mid a, W, 0) \\
\bar{Q}^Z_{a,0}(Z, A, W, S) &= \sum_m \bar{Q}_{a,0}(m, W)(M, Z, a, W)p_M(m \mid Z, A, W, S)
\end{aligned}$$

Proof. According to the general approach of section 3.1.2, we will compute a pathwise derivative of the parameter mapping. From equation (3.14) we obtain

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} g_{M,\epsilon}(m \mid z, x, w, s) = (\mathbb{E}[\gamma(O) \mid m, z, x, w, s] - \mathbb{E}[\gamma(O) \mid z, x, w, s])\, g_M(m \mid z, x, w, s) \tag{3.27}$$

where γ is the score along which the pathwise derivative is being computed. Everything stays identical to theorem 3.1.2, except we will have the following extra piece of the derivative:

$$\begin{aligned}
&\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int y p_Y(y \mid m, z, a, w, s = 1) \sum_c \left[g_{M,\epsilon}(m \mid c, a^*, w, s^*)p_{Z,\epsilon}(c \mid a^*, w, s^*)\right] p_Z(z \mid a, w, 0)p_W(W \mid 0)d\nu(o) \\
&= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int y p_Y(y \mid m, z, a, w, s = 1) \sum_c \left[g_{M,\epsilon}(m \mid c, a^*, w, s^*)p_Z(c \mid a^*, w, s^*)\right] p_Z(z \mid a, w, 0)p_W(W \mid 0)d\nu(o) && (3.28) \\
&+ \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int y p_Y(y \mid m, z, a, w, s = 1) \sum_c \left[g_M(m \mid c, a^*, w, s^*)p_{Z,\epsilon}(c \mid a^*, w, s^*)\right] p_Z(z \mid a, w, 0)p_W(W \mid 0)d\nu(o) && (3.29)
\end{aligned}$$

To compute 3.28 we have

$$\begin{aligned}
&\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int y p_Y(y \mid m, z, a, w) \sum_c \left[g_{M,\epsilon}(m \mid c, a^*, w, s^*)p_Z(c \mid a^*, w, s^*)\right] p_Z(z \mid a, w, 0)p_W(w \mid 0)d\nu(y, m, z, w) \\
=&\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int \bar{Q}_{a,0}(m, w) \sum_c \left[g_{M,\epsilon}(m \mid c, a^*, w, s^*)p_Z(c \mid a^*, w, s^*)\right] p_W(w \mid 0)d\nu(m, w) \\
=&\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int \bar{Q}_{a,0}(m, w)g_{M,\epsilon}(m \mid z, x, w, s)p_Z(z \mid x, w, s)\mathbb{I}(x = a^*, s = s^*)*
\end{aligned}$$

$$\begin{aligned}
&\frac{p_{A,S|W}(x,s \mid w)p_{S|W}(0 \mid w)p_W(w)}{p_{A,S|W}(x,s \mid w)P(s=0)}d\nu(m,z,x,w,s) \\
\overset{(3.27)}{=}&\int \gamma(o)\left(\bar{Q}_{a,0}(m,w) - \left(\bar{Q}_{a,0}(1,w)g_M(1 \mid z,x,w,s) + \bar{Q}_{a,0}(0,w)g_M(0 \mid z,x,w,s)\right)\right) * \\
&\frac{\mathbb{I}(x=a^*, s=s^*)p_{S|W}(0 \mid w)}{p_{A,S|W}(x,s \mid w)P(s=0)}p(o)d\nu(o) \\
=&\int \gamma(o)(m - g_M(1 \mid z,x,w,s))\frac{\mathbb{I}(x=a^*, s=s^*)(\bar{Q}_{a,0}(1,w) - \bar{Q}_{a,0}(0,w))p_{S|W}(0 \mid w)}{g_A(x \mid s,w)p_{S|W}(s \mid w)P(s=0)}p(o)d\nu(o) \\
=&\langle \gamma, D^*_{f,M}(P)\rangle_{L^2(P)}
\end{aligned}$$

To compute 3.29 we have

$$\begin{aligned}
&\frac{d}{d\epsilon}\bigg|_{\epsilon=0}\int yp_Y(y \mid m,z,a,w)\sum_c\left[g_M(m \mid c,a^*,w,s^*)p_{Z,\epsilon}(c \mid a^*,w,s^*)\right]p_Z(z \mid a,w,0)p_W(w \mid 0)d\nu(m,z,w) \\
=&\frac{d}{d\epsilon}\bigg|_{\epsilon=0}\int \bar{Q}_{a,0}(m,w)\sum_c\left[g_M(m \mid c,a^*,w,s^*)p_{Z,\epsilon}(c \mid a^*,w,s^*)\right]p_W(W \mid 0)d\nu(m,w) \\
=&\frac{d}{d\epsilon}\bigg|_{\epsilon=0}\int \bar{Q}_{a,0}(m,w)g_M(m \mid z,x,w,s)p_{Z,\epsilon}(z \mid x,w,s)\mathbb{I}(x=a^*, s=s^*)* \\
&\frac{p_{A,S|W}(x,s \mid w)p_{S|W}(0 \mid w)p_W(W)}{p_{A,S|W}(x,s \mid w)P(s=0)}d\nu(m,z,x,w,s) \\
\overset{(3.22)}{=}&\int \gamma(o)\Big(\bar{Q}^Z_{a,0}(z,x,w,s) - \Big(\bar{Q}^Z_{a,0}(1,x,w,s)p_Z(1 \mid x,w,s) + \bar{Q}^Z_{a,0}(0,x,w,s)p_Z(0 \mid x,w,s)\Big)\Big) * \\
&\frac{\mathbb{I}(x=a^*, s=s^*)p_{S|W}(0 \mid w)}{p_{A,S|W}(x,s \mid w)P(s=0)}p(o)d\nu(o) \\
=&\int \gamma(o)(z - p_Z(1 \mid x,w,s))\frac{\mathbb{I}(x=a^*, s=s^*)(\bar{Q}^Z_{a,0}(1,x,w,s) - \bar{Q}^Z_{a,0}(0,x,w,s))p_{S|W}(0 \mid w)}{g_A(x \mid w,s)p_{S|W}(s \mid w)P(s=0)}p(o)d\nu(o) \\
=&\langle \gamma, D^*_{f,Z}(P)\rangle_{L^2(P)}
\end{aligned}$$

by the general approach in section 3.1.2 we have finished the proof.

□

3.1.10 Example 8: Efficient Influence Curve for Transporting Stochastic Direct, Fixed Parameter, Restricted Model

We will now derive the efficient influence as per the previous section parameter but we will assume the M and Y mechanisms do not directly depend on A, i.e., A is an instrument.

Theorem 3.1.5. *The efficient influence curve for the unrestricted model at distribution, P, is given by $D^*_f(P) = D^*_{f,Y,r}(P) + D^*_{f,M,r}(P) + D^*_{f,Z}(P) + D^*_{f,W}(P)$ where*

$$\begin{aligned}
D^*_{f,Y,r}(P) &= D^*_{Y,r}(P) \\
D^*_{f,M,r}(P) &= (M - g_M(1 \mid Z,A,W,S))\frac{(\bar{Q}_{a,0}(1,W) - \bar{Q}_{a,0}(0,W))p_{S|W}(0 \mid W)\mathbb{I}(S=s^*)p_Z(Z \mid a^*,W,S)}{p_Z(Z \mid W,S)p_{S|W}(S \mid W)P(S=0)} \\
D^*_{f,Z}(P) &= D^*_Z(P) + (Z - p_Z(1 \mid A,W,S))* \\
&\quad \frac{(\bar{Q}^Z_{a,0}(1,A,W,S) - \bar{Q}^Z_{a,0}(0,A,W,S)p_{S|W}(0 \mid W)\mathbb{I}(A=a^*, S=s^*)}{g_A(A \mid W,S), p_{S|W}(S \mid W)P(S=0)} \\
D^*_{f,W}(P) &= D^*_W(P)
\end{aligned}$$

where $D^*_{Y,r}$ remains the same as for the restricted model and the data adaptive parameter in theorem 3.1.3 because this portion of the influence curve is not affected by the scores in T_M due to it being orthogonal to T_M. $D^*_{f,Z}$ and $D^*_{f,W}$ are the same as for the fixed parameter and unrestricted model because T_M is orthogonal to T_Z, T_W and T_S.

Proof. We will utilize the following facts, very similarly to equations 3.25 and 3.26:

$$p_{A,YS,r}(a, ys \mid m, z, w, s) = \frac{p_{YS,r}(ys \mid m, z, w, s) p_{\bar{Z}}(z, x, w, s)}{p_{\bar{Z}/A}(z, w, s)} \tag{3.30}$$

$$p_{A,YS,M,r}(a, ys, m \mid z, w, s) = \frac{p_{YS,r}(ys \mid m, z, w, s) p_{M,r}(m \mid z, w, s) p_Z(z \mid x, w, s)}{p_{\bar{Z}/A}(z, w, s)} \tag{3.31}$$

We will project onto the tangent space $D^*_{f,M}(P)$ onto the tangent space of mean zero function of O given Z, W, S.

$$\begin{aligned}
D^*_{f,M,r}(P) = &\prod(D^*_{f,M}(P) \mid T_{A,YS,M}) \\
&= \mathbb{E}[D^*_{f,M}(P)(O) \mid M, Z, W, S] - \mathbb{E}[D^*_{f,M}(P)(O) \mid Z, W, S] \\
&\overset{(3.31)}{=} \mathbb{E}[D^*_{f,M}(P)(O) \mid M, Z, W, S] \\
&= \int (m - g_M(1 \mid z, x, w, s)) \frac{\mathbb{I}(x = a^*, s = s^*)(\bar{Q}_{a,0}(1, w) - \bar{Q}_{a,0}(0, w)) p_{S|W}(0 \mid w)}{g_A(x \mid s, w) p_{S|W}(s \mid w) P(s = 0)} * \\
&\quad p_{A,YS,r}(a, ys \mid m, z, w, s) d\nu(a, ys) \\
&\overset{(3.30)}{=} \int (m - g_M(1 \mid z, x, w, s)) \frac{\mathbb{I}(s = s^*)(\bar{Q}_{a,0}(1, w) - \bar{Q}_{a,0}(0, w)) p_Z(z \mid a*, w, s) p_{S|W}(0 \mid w)}{p_{Z|W,S}(z \mid w, s) p_{S|W}(s \mid w) P(s = 0)}
\end{aligned}$$

Since x plays the role of A in the integrand, so as to not confuse a lower case a with the fixed values, the proof is complete. □

3.1.11 Example 9: Influence Mean Under Stochastic Intervention for Longitudinal Data

Let us assume we have longitudinal data of the form $L(0)$ = baseline confounders, $A(0)$, treatment given at baseline, followed by time varying confounders, $L(1)$ and treatment at time point 1, $A(1)$ so that our observed data is $O = (L(0), A(0), L(1), A(1), ..., L(K), A(K), Y)$ where Y is the outcome. We will use the shorthand notation $\bar{L}(j) = (L(0), ..., L(j))$ and likewise for $\bar{A}(j)$ so that $O = (\bar{L}(K), \bar{A}(K), Y)$, for the treatment or exposure variable. Note, $a(-1)$ and $l(-1)$ are null and there is no treatment mechanism at time $K + 1$. We define the conditional probability distributions, $P_{L(i)}$, the conditional distribution of $L(i)$ given the past as well as $P_{A(i)}$, the conditional distribution of $A(i)$ given the past. The corresponding respective densities to these conditional distributions have the same subscripted notation, $p_{L(i)}$ and $g_{A(i)}$, where we use the letter g to distinguish the treatment mechanism densities it from the conditional densities of the confounders, $L(i)$. $\bar{L}(i) = (L(i), ..., L(0))$, the confounder history through time, i, and likewise for $\bar{A}(i)$, the treatment history through time, i. As usual we use lower case letters for these equivalent variables when using integral notation.

Theorem 3.1.6. *The efficient influence curve for the mean under stochastic intervention given by*

$$g(\bar{A}) = \prod_{i=0}^{K} g_i^*(A(i) \mid \bar{L}(i), \bar{A}(i-1))$$

is given by the function

$$\begin{aligned} D^*(P)(O) &= D^*(P)(\bar{L}(K+1), \bar{A}(K)) && (3.32) \\ &= \sum_{j=0}^{K+1} \left(\prod_{i=0}^{j-1} \frac{g_i^*(A(i) \mid \bar{L}(i), \bar{A}(i-1))}{g_i(A(i) \mid \bar{L}(i), \bar{A}(i-1))} \right) \left(\bar{Q}_{L(j)}(\bar{L}(i), \bar{A}(i-1)) - \mathbb{E}_{P_{L(j)}}[\bar{Q}_{L(j)} \mid \bar{L}(i-1), \bar{A}(i-1)] \right) && (3.33) \end{aligned}$$

where starting with $Y = \bar{Q}_{L(K+1)}$
we set $\mathbb{E}_{P_{g^*}} \left[\bar{Q}_{L(K+1)} \mid \bar{L}(K), \bar{A}(K-1) \right] = \bar{Q}_{L(K)}(\bar{L}(K), \bar{A}(K-1))$
and we continue to recursively set
$\bar{Q}_{L(i)} \left(\bar{L}(i), \bar{A}(i-1) \right) = \mathbb{E}_{P_{g^*}} \left[\bar{Q}_{L(i+1)} \mid \bar{L}(i), \bar{A}(i-1) \right]$

Proof. we have for a path, P_ϵ through P:

$$\frac{d}{d\epsilon} \Psi(P_\epsilon) \bigg|_{\epsilon=0} = \frac{d}{d\epsilon} \bigg|_{\epsilon=0} \int y \prod_{i=0}^{K+1} p_{L(i),\epsilon} \left(l(i) \mid \bar{a}(i-1), \bar{l}(i-1) \right) g_i^* \left(a(i) \mid \bar{a}(i-1), \bar{l}(i) \right) dv(o)$$

Regarding Semi-Parametric Models With Known Treatment Mechanism

We already notice that we will have only parts of the score corresponding to $p_{L(i),\epsilon}$ parts of the likelihood and not the treatment mechanism since the parameter does not depend on these factors. This will automatically make the efficient influence curve only in the part of the tangent space defined by the mean 0 functions of $L(i)$ given the past, i.e., the efficient influence curve will only have components of the form $\frac{d}{d\epsilon} logp_{L(i),\epsilon} \Big|_{\epsilon=0}$ and thus, since these components are orthogonal to the mean 0 functions of $A(i)$ given the past, i.e.$\frac{d}{d\epsilon} logp_{A(i),\epsilon} \Big|_{\epsilon=0}$, the efficient influence curve for the model with known treatment mechanism will be the same as for the non-parametric model. Now we can shorten things with subscripts indicative of the variable the conditional probabilities are functions of. We can notice that (3.14) implies

$$\frac{d}{d\epsilon} p_{L(i),\epsilon}(l(i) \mid \bar{a}(i-1), \bar{l}(i-1)) \bigg|_{\epsilon=0} = p_{L(i)}(l(i) \mid \bar{a}(i-1), \bar{l}(i-1)) * \left(\mathbb{E}\left[S(O) \mid \bar{a}(j-1), \bar{l}(j) \right] - \mathbb{E}\left[S(O) \mid \bar{a}(j-1), \bar{l}(j-1) \right] \right)$$

$$\begin{aligned} &\frac{d}{d\epsilon} \bigg|_{\epsilon=0} \int y \prod_{i=0}^{K+1} p_{L(i),\epsilon} \left(l(i) \mid \bar{a}(i-1), \bar{l}(i-1) \right) \prod_{i=0}^{K} g^*_{A(i)} \left(a(i) \mid \bar{a}(i-1), \bar{l}(i) \right) dv(o) \\ \overset{(3.14)}{=} &\sum_{j=0}^{K+1} \int y \prod_{i=j}^{K+1} p_{L(i)}(l(i) \mid \bar{a}(i-1), \bar{l}(i-1)) g^*_{A(i-1)}(a(i-1) \mid \bar{a}(i-2), \bar{l}(i-1)) \mathbb{E}\left[S(O) \mid \bar{a}(j-1), \bar{l}(j) \right] \\ &\int \prod_{i=0}^{j-1} p_{L(i)}(l(i) \mid \bar{a}(i-1), \bar{l}(i-1)) g^*_{A(i)}(a(i) \mid \bar{a}(i-1), \bar{l}(i)) dv(o) \end{aligned}$$

$$
\begin{aligned}
&-\sum_{j=0}^{K+1}\int y\prod_{i=j}^{K+1}p_{L(i)}(l(i)\mid\bar{a}(i-1),\bar{l}(i-1))g^*_{A(i-1)}(a(i-1)\mid\bar{a}(i-2),\bar{l}(i-1))\mathbb{E}\left[S(O)\mid\bar{a}(j-1),\bar{l}(j-1)\right]\\
&\prod_{i=0}^{j-1}p_{L(i)}(l(i)\mid\bar{a}(i-1),\bar{l}(i-1))g^*_{A(i)}(a(i)\mid\bar{a}(i-1),\bar{l}(i))d\nu(o)\\
=&\int\sum_{j=0}^{K+1}\bar{Q}_{L(j)}(\bar{l}(j),\bar{a}(j-1))(\bar{l}(j),\bar{a}(j-1))p_{L(j)}(l(j)\mid\bar{a}(j-1),\bar{l}(j-1))\mathbb{E}_P\left[S(O)\mid\bar{a}(j-1),\bar{l}(j)\right]\\
&\prod_{i=0}^{j-1}p_{L(i)}(l(i)\mid\bar{a}(i-1),\bar{l}(i-1))g^*_{A(i)}(a(i)\mid\bar{a}(i-1),\bar{l}(i))d\nu(o)\\
&-\sum_{j=0}^{K+1}\int\mathbb{E}_{P_{L(j)}}[\bar{Q}_{L(j)}(\bar{l}(j),\bar{a}(j-1))\mid\bar{l}(i-1),\bar{a}(i-1)]\mathbb{E}_{P_{g^*}}\left[S(O)\mid\bar{a}(j-1),\bar{l}(j-1)\right]\\
&\prod_{i=0}^{j-1}p_{L(i)}(l(i)\mid\bar{a}(i-1),\bar{l}(i-1))g^*_{A(i)}(a(i)\mid\bar{a}(i-1),\bar{l}(i))d\nu(o)\\
\overset{fubini}{=}&\sum_{j=0}^{K+1}\int\bar{Q}_{L(j)}(\bar{l}(j),\bar{a}(j-1))(\bar{l}(j),\bar{a}(j-1))p_{L(j)}(l(j)\mid\bar{a}(j-1),\bar{l}(j-1))\mathbb{E}_P\left[S(O)\mid\bar{a}(j-1),\bar{l}(j)\right]\\
&\prod_{i=0}^{j-1}p_{L(i)}(l(i)\mid\bar{a}(i-1),\bar{l}(i-1))g^*_{A(i)}\frac{g_{A(i)}}{g_{A(i)}}(a(i)\mid\bar{a}(i-1),\bar{l}(i))d\nu(o)\\
&-\sum_{j=0}^{K+1}\int\mathbb{E}_{P_{L(j)}}[\bar{Q}_{L(j)}(\bar{l}(j),\bar{a}(j-1))\mid\bar{l}(i-1),\bar{a}(i-1)]\mathbb{E}_{P_{g^*}}\left[S(O)\mid\bar{a}(j-1),\bar{l}(j-1)\right]\\
&\prod_{i=0}^{j-1}p_{L(i)}(l(i)\mid\bar{a}(i-1),\bar{l}(i-1))g^*_{A(i)}\frac{g_{A(i)}}{g_{A(i)}}(A(i)\mid\bar{a}(i-1),\bar{l}(i))d\nu(o)\\
\overset{fubini}{=}&\sum_{j=0}^{K+1}\mathbb{E}\left(\prod_{i=0}^{j-1}\frac{g^*_{A(i)}}{g_{A(i)}}(A(i)\mid\bar{A}(i-1),\bar{L}(i))\right)\left(\bar{Q}_{L(j)}\left(\bar{L}(j),\bar{A}(j-1)\right)-\mathbb{E}_P[\bar{Q}_{L(j)}\mid\bar{L}(j-1),\bar{A}(j-1)]\right)S(O)\\
=&\left\langle\sum_{j=0}^{K+1}\left(\prod_{i=0}^{j-1}\frac{g^*_{A(i)}}{g_{A(i)}}\right)\left(\bar{Q}_{L(j)}-\mathbb{E}_P[\bar{Q}_{L(j)}\mid\cdot,\cdot]\right),S\right\rangle_{L^2_0(P)}
\end{aligned}
$$

And the proof is complete by the riesz representation theorem.

□

3.1.12 Example 10: Survival Under a Dynamic Rule

We can also perform a similar analysis with right censored survival data. In this case, we observe an event time, $\tilde{T} = min(C, T)$ and Δ where $\Delta = 1$ indicates the death was observed, i.e., that $\tilde{T} = T$. Otherwise we observe the censoring time, C. We also have observed confounders, W, and a treatment assignment, A, given at baseline. Thus our observed data is of the form:

$$O = (W, A, \tilde{T}, \Delta) \sim \mathcal{M},\ \text{non-parametric}$$

Our parameter mapping is defined as

$$\Psi(P) = \mathbb{E}\prod_{t=0}^{t_0}\left(1 - \mathbb{E}_P\left[dN(t) \mid A = d(W), W, N(t-1) = A_2(t-1) = 0\right]\right)$$

where $A_2(t)$ indicates whether the subject was censored at time t or before and $N(t)$ is an indicator of whether the subject has died or not. The ordering of the variables is as follows for some discretization of time which, WLOG, we just set to $0, 1, 2, ...etc$ of time: $W =$ confounders, $A =$ treatment assignment, $A_2(0) =$ indicator of censoring in which case, C $=$ 1, $dN(1) =$ indicator of failure at time 1, $A_2(1)$, then $dN(2)$, $A_2(2)$,etc. We note that this is an alternate form of the observed data structure for discretized time

To place this in the framework of our general method, we can notice we have conditional densities of death, given the past. Define, $dN(t)$ as the indicator of death at time t. Then the conditional density of death at time, t, given the past is denoted $p_{dN(t)}$. Therefore, by (3.14) we get

$$\frac{d}{d\epsilon} p_{dN(i),\epsilon}(dN(t) \mid N(t) = 0, A, W)\Big|_{\epsilon=0} = p_{dN(i)}(dN(t) \mid pa(A_2(t))) * (\mathbb{E}\left[S(O) \mid pa(A_2(t))\right] - \mathbb{E}\left[S(O) \mid pa(dN(t))\right]) \tag{3.34}$$

$$\text{and } \frac{d}{d\epsilon} p_{W),\epsilon}(w)\Big|_{\epsilon=0} = p_W(w) * (\mathbb{E}\left[S(O) \mid w\right] - \mathbb{E}S(O)) \tag{3.35}$$

where $pa(A_2(t))$ are all the preceding variables to the censoring mechanism at time, t, including $dN(t)$ Using the same principles as previously described we can differentiate the parameter mapping along a path defined by the score, S, at the truth, P, as follows. We will proceed by differentiating the parameter mapping as in the previous section and once we have written the derivative as an inner product of a function with the score, that function will be our efficient influence curve. We note $s_c(t \mid A, W)$ is the probability of being censored after time, $t - 1$, having received treatment A and with confounders, W. $s(t \mid A, W)$ is the conditional probability of survival past time t, given A and W. We note to the reader that survival estimates can be obtained for s_c from those who were censored at the various time points, such as with a pooled logistic regression where all participants contribute a line of data for each time point they are uncensored and a time for each of those lines of data. Similarly we can get estimates of the conditional survival hazard, $\lambda(\cdot \mid A, W)$. The regressions are then fit and we can estimate the probability of being censoring beyond time, t, as $s_c(t \mid A, W) = \prod_{c=0}^{t}(1 - \lambda_C(c \mid A, W))$ where our regression estimates $\lambda(c \mid A, W)$ for all of the discrete times, c.

Theorem 3.1.7. *The efficient influence curve for* $\Psi(P)$ *is*

$$D^*(W, A, \tilde{T}, \Delta) = \left[\sum_{t=1}^{t_0} \frac{I(A = d(W))I(\tilde{T} > t-1)s(t_0 \mid A, W)}{g(A \mid W)s_c(t-1|A, W)s(t \mid A, W)} \times (dN(t) - \lambda(t \mid A, W)) + s(t_0 \mid A = d(W), W) - \Psi(P)\right]$$

Proof.

$$\begin{aligned}
\frac{d}{d\epsilon}\Big|_{\epsilon=0} \Psi(P_\epsilon) =& \mathbb{E}_w \frac{d}{d\epsilon}\Big|_{\epsilon=0} \prod_{t=1}^{t_0} (1 - \mathbb{E}_{P_\epsilon}\left[dN(t) \mid A = d(W), W, N(t-1) = A_2(t-1) = 0\right]) + \\
& \frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathbb{E}_{P_{W,\epsilon}} \prod_{t=0}^{t_0} (1 - \mathbb{E}_P\left[dN(t) \mid A = d(W), W, N(t-1) = A_2(t-1) = 0\right]) \\
=& \mathbb{E} \frac{d}{d\epsilon}\Big|_{\epsilon=0} \prod_{t=1}^{t_0} (1 - \mathbb{E}_{P_\epsilon}\left[dN(t) \mid A = d(W), W, N(t-1) = A_2(t-1) = 0\right]) + \\
& \int \prod_{t=0}^{t_0} (1 - \mathbb{E}_P\left[dN(t) \mid a = d(w), w, n(t-1) = a_2(t-1) = 0\right]) \frac{d}{d\epsilon}\Big|_{\epsilon=0} p_{W,\epsilon} d\nu(w)
\end{aligned}$$

$$\begin{aligned}
&= \int \sum_{t=1}^{t_0} \prod_{i \neq t}^{t_0} (1 - \mathbb{E}_P\left[dN(i) \mid A = d(W), W, N(i-1) = A_2(i-1) = 0\right]) \times \\
&\quad \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \mathbb{E}_{P_\epsilon}\left[dN(t) \mid A = d(W), W, N(t-1) = A_2(t-1) = 0\right] + \\
&\quad \text{(from (3.35))} \\
&\quad \int \prod_{t=0}^{t_0} (1 - \mathbb{E}_P\left[dN(t) \mid a = d(w), w, n(t-1) = a_2(t-1) = 0\right]) \frac{d}{d\epsilon}\bigg|_{\epsilon=0} (\mathbb{E}\left[S(O) \mid w\right] - \mathbb{E}S(O)) p_W d\nu(w) \\
&= \int \sum_{t=1}^{t_0} \frac{\prod_{i=1}^{t_0} (1 - \mathbb{E}_P\left[dN(t) \mid A = d(w), w, N(i-1) = A_2(i-1) = 0\right])}{1 - \lambda(t \mid A = d(W), W)} \times \\
&\quad \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \int dn(t) p_{dN(\tau)\epsilon}(dn(t) \mid A = d(w), W = w, n(t-1) = a_2(t-1) = 0) \\
&\quad dv(dn(t)) dv(w) + \mathbb{E}\left[s(O) \prod_{t=0}^{t_0} (1 - \mathbb{E}_P\left[dN(t) \mid a = d(w), w, n(t-1) = a_2(t-1) = 0\right]) - \Psi(P)\right] \\
&= \int \sum_{t=1}^{t_0} \frac{S(t \mid A - d(W), W)}{1 - \lambda(t \mid A = d(W), W)} \times \\
&\quad \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \int dn(t) \frac{I(a = d(w)) I(N(i-1) = A_2(i-1) = 0)}{g(a \mid w) \prod_{i=0} g_{A_2(i)}(a_2(i) \mid pa(a_2(i)) \prod_{i=1}^{t-1} p_{dN(i)}(dn(i) \mid pa(dn(i))} \times \\
&\quad p_{dN(t)\epsilon}(dn(t) \mid pa(dn(t)) \mid pa(dn(i)) \\
&\quad \times \prod_{i=0}^{t-1} g_{A_2(i)}(a_2(i) \mid pa(a_2(i)) g(a \mid w) p_W(w) dv(o) \\
&\quad + \mathbb{E}\left[s(O) S(t_0 \mid a = d(W), W) - \Psi(P)\right] \\
&\overset{3.34}{=} \int \sum_{t=1}^{t_0} \frac{S(t \mid a = d(w), w)}{1 - \lambda(t \mid a = d(w), w)} dn(t) \frac{I(a = d(w)) I(N(i-1) = A_2(i-1) = 0)}{g(a \mid w) S_c(t-1 \mid a, w) S(t-1 \mid a, w)} \times \\
&\quad (\mathbb{E}_P\left[S \mid pa(a_2(t))\right] - \mathbb{E}_P\left[S \mid pa(dn(t))\right]) p_{dN(\tau)}(dn(t) \mid pa(dn(t)) \times \\
&\quad \times \prod_{k=1}^{t-1} p_{dN(i)}(dn(k) \mid pa(dn(k)) \times \prod_{k=0}^{t-1} g_{A_2(k)}(a_2(k) \mid pa(a_2(k)) g(a \mid w) p_W(w) dv(o) + \\
&\quad \mathbb{E}\left[s(O) S(t_0 \mid a = d(W), W) - \Psi(P)\right] \\
&= \int \sum_{t=1}^{t_0} \frac{I(a = d(w)) I(N(i-1) = A_2(i-1) = 0) S(t_0 \mid a, w)}{g(a \mid w) S_c(t-1 | a, w) S(t \mid a, w)} \\
&\quad \left(dn(t) \mathbb{E}_P\left[S \mid pa(a_2(t))\right] p_{dN(\tau)}(dn(t) \mid pa(dn(t)) - \lambda(t \mid a, w) \mathbb{E}_P\left[S \mid pa(dn(t))\right]\right) \\
&\quad \prod_{k=1}^{t-1} p_{dN(i)}(dn(k) \mid pa(dn(k)) \times \prod_{k=0}^{t-1} g_{A_2(k)}(a_2(k) \mid pa(a_2(k)) g(a \mid w) p_W(w) dv(o) + \\
&\quad \mathbb{E}\left[s(O) S(t_0 \mid a = d(W), W) - \Psi(P)\right]) \\
&\overset{fubini}{=} \int \sum_{t=1}^{t_0} \frac{I(a = d(w)) I(N(i-1) = A_2(i-1) = 0) S(t_0 \mid a, w)}{g(a \mid w) S_c(t-1 | a, w) S(t \mid a, w)} \times (dn(t) - \lambda(t \mid a, w)) s(o) p(o) dv(o) + \\
&\quad \mathbb{E}\left[s(O) S(t_0 \mid a = d(W), W) - \Psi(P)\right] \\
&= \mathbb{E}\left[s(O) \left[\sum_{t=1}^{t_0} \frac{I(A = d(W)) I(N(i-1) = A_2(i-1) = 0) S(t_0 \mid A, W)}{g(A \mid W) S_c(t-1 | A, W) S(t \mid A, W)} \times \left(dN(t) - \lambda(t \mid A, W)\right) + \right.\right. \\
&\quad \left.\left. S(t_0 \mid A = d(W), W) - \Psi(P)\right]\right] \\
&= \left\langle \left[\sum_{t=1}^{t_0} \frac{I(A = d(W)) I(\tilde{T} > t-1) S(t_0 \mid A, W)}{g(A \mid W) S_c(t-1 | A, W) S(t \mid A, W)} \times \left(I(\tilde{T} = t) - \lambda(t \mid A, W)\right) \right.\right. \\
&\quad \left.\left. + S(t_0 \mid A = d(W), W) - \Psi(P)\right], S(O) \right\rangle_{L_0^2(P)}
\end{aligned}$$

And we can see the influence curve in the inner product with the score and the proof is complete. Note, that we can replace $dN(t)$ with $I(\tilde{T}=t)$ because either time, t, is a censored time or the term is 0. Also, by definition of $\tilde{T}$, $I(N(i-1)=A_2(i-1)=0)=I(\tilde{T}>t-1)$ $\square$

3.1.13 Example 11: TE CDF

Theorem 3.1.8. *Assume k is lipschitz and smooth on $\mathbb{R}$. The efficient influence curve for the parameter, $\Psi_{\delta,t}$, is given by*

$$\mathbf{D}^{\star}_{\boldsymbol{\Psi}_{\delta,\mathbf{t}}}(\mathbf{P})(\mathbf{O}){=}\frac{-1}{\delta}\mathbf{k}\left(\frac{\mathbf{b}(\mathbf{W})-\mathbf{t}}{\delta}\right)*\frac{\mathbf{2A-1}}{\mathbf{g}(\mathbf{A}|\mathbf{W})}(\mathbf{Y}-\bar{\mathbf{Q}}(\mathbf{A},\mathbf{W})){+}\int\frac{\mathbf{1}}{\delta}\mathbf{k}\left(\frac{\mathbf{x}-\mathbf{t}}{\delta}\right)\mathbb{I}(\mathbf{b}(\mathbf{W})\leq\mathbf{x})\mathbf{dx}-\boldsymbol{\Psi}_{\delta,\mathbf{t}}$$

Proof. Define $\Phi(x)=1/(1+exp(x))$. We also define $b_\epsilon=\mathbb{E}_{P_\epsilon}[Y\mid A=1,W]-\mathbb{E}_{P_\epsilon}[Y\mid A=1,W]$, where P_ϵ and S, the so-called score function, are as previously defined. We will now compute the pathwise derivative functional on $L^2_0(P)$, writing it as an inner product (covariance in the Hilbert Space $L^2(P)$), of the score, S, and the efficient influence curve, a unique element of the tangent space, $L^2_0(P)$. We notate the efficient influence curve as indexed by the distribution, P, and as a function of the observed data, $O\sim P$: $D^*(P)(O)$. By dominated convergence we have

$$\Psi_{\delta,t}(P)=\underset{h\to 0}{lim}\mathbb{E}_W\int_{-1}^{1}\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\Phi(\frac{b(W)-x}{h})dx$$

$$\lim_{\epsilon\to 0}\frac{\Psi_{\delta,t}(P_\epsilon)-\Psi_{\delta,t}(P)}{\epsilon} \tag{3.36}$$

$$=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\underset{h\to 0}{lim}\mathbb{E}_W\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\left(\Phi(\frac{b_\epsilon(W)-x}{h})-\Phi(\frac{b(W)-x}{h})\right)dx$$

$$+\,\mathbb{E}_W\left(\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\mathbb{I}(b(W)\leq x)-\Psi_{t,\delta}(P)\right)S(O)dx$$

let's ignore (2) for now

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\underset{h\to 0}{lim}\mathbb{E}_W\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\left(\frac{1}{h}\Phi'(\frac{b(W)-x}{h})\left(b_\epsilon(W)-b(W)\right)\right)dx+$$

$$+\lim_{\epsilon\to 0}\frac{1}{\epsilon}\underset{h\to 0}{lim}\mathbb{E}_W\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\left(\frac{1}{2h^2}\Phi^{(2)}\left(\zeta\left(\frac{x-b(W)}{h}\right)\right)\left(b_\epsilon(W)-b(W)\right)^2\right)dx$$

$$=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\underset{h\to 0}{lim}\left(\mathbb{E}_W\int_x\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\left(\frac{1}{h}\Phi'(\frac{b(W)-x}{h})\left(b_\epsilon(W)-b(W)\right)\right)dx+R_{2,h,x}(b_\epsilon,b)\right) \tag{3.37}$$

We can note that for $h(\epsilon)$ such that $\frac{\epsilon}{h^2(\epsilon)}\to 0$ as $\epsilon\to 0$, $\frac{R_2}{\epsilon}\to 0$ because R_2 is order $\frac{\epsilon^2}{h^2}$. To see this, consider the convenient fact that $\Phi^{(2)}(x)$ is bounded.
Let's now drop $\underset{\epsilon\to 0}{lim}\frac{1}{\epsilon}$ for now and use integration by parts to compute a part of the integrand in (3.37):

$$\mathbb{E}_W\underset{a\to\infty}{lim}\int_{t-a\delta}^{t+a\delta}\frac{1}{\delta}k\left(\frac{x-t}{\delta}\right)\frac{1}{h}\Phi'(\frac{b(W)-x}{h})dx\left(b_\epsilon(W)-b(W)\right)$$

$$=\mathbb{E}_W\underset{a\to\infty}{lim}\left(\frac{-1}{\delta}k\left(\frac{x-t}{\delta}\right)\Phi(\frac{b(W)-x}{h})\Big|_{t-a\delta}^{t+a\delta}+\int_x\frac{1}{\delta^2}k'\left(\frac{x-t}{\delta}\right)\mathbb{E}_W\Phi(\frac{b(W)-x}{h})\right)\left(b_\epsilon(W)-b(W)\right)dx$$

$$
\begin{aligned}
&= \mathbb{E}_W \lim_{a\to\infty} \left(\frac{-1}{\delta} k\left(\frac{x-t}{\delta}\right) \Phi(\frac{b(W)-x}{h}) \Big|_{t-a\delta}^{t+a\delta} + \int \frac{1}{\delta^2} k'\left(\frac{x-t}{\delta}\right) \mathbb{E}_W \left[\Phi(\frac{b(W)-x}{h}) - \mathbb{I}(b(W) \le x) \right] \right) (b_\epsilon(W) - b(W))\, dx \\
&+ \mathbb{E}_W \int_x \frac{1}{\delta^2} k'\left(\frac{x-t}{\delta}\right) \mathbb{I}(b(W) \le x)\, (b_\epsilon(W) - b(W))\, dx \\
&h \to 0 \text{ and Dominated convergence } \implies \text{ 2nd term disappears. k lipschitz } \implies \\
&= \mathbb{E}_W \lim_{a\to\infty} \frac{-1}{\delta} \mathbb{I}(b(W) \le t + a\delta) k(a) + \frac{1}{\delta} \mathbb{I}(b(W) \le t - a\delta) k(-a)) \\
&+ \frac{1}{\delta} \left(k(a) - k\left(\frac{max(b(W), t - a\delta) - t}{\delta} \right)) \right) \mathbb{I}(b(W) \le t - a\delta)\, (b_\epsilon(W) - b(W)) \\
&= \mathbb{E}_W \frac{-1}{\delta} k\left(\frac{b(W)-t}{\delta}\right) (b_\epsilon(W) - b(W))
\end{aligned}
$$

We can summarize as follows:

$$
\begin{aligned}
\lim_{\epsilon\longrightarrow 0} \frac{\Psi_{\delta,t}(P_\epsilon) - \Psi_{\delta,t}(P)}{\epsilon} &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \mathbb{E}_W \left(\frac{-1}{\delta} k\left(\frac{b(W)-t}{\delta}\right) (b_\epsilon(W) - b(W)) \right) + \qquad (3.38) \\
&+ \lim_{\epsilon\to 0} \lim_{h(\epsilon)\to 0} \mathbb{E}_W \frac{R_{2,h,x}(b_\epsilon, b)}{\epsilon} \qquad (3.39) \\
&+ \mathbb{E}_W \left(\int_x \frac{1}{\delta} k\left(\frac{x-t}{\delta}\right) \mathbb{I}(b(W) \le x) - \Psi_{t,\delta}(P) \right) S(O) dx \qquad (3.40)
\end{aligned}
$$

As previously stated, the term (3.39(disappears by easy choice of h. We then compute the pathwise derivative along S at $\epsilon = 0$ to compute term 3.38:

$$
\begin{aligned}
&\lim_{\epsilon\to 0} \frac{1}{\epsilon} \mathbb{E}_W \left(\frac{-1}{\delta} k\left(\frac{b(W)-t}{\delta}\right) (b_\epsilon(W) - b(W)) \right) \\
&= \int \left(\frac{-1}{\delta} k\left(\frac{b(w)-t}{\delta}\right) \int \left(p_\epsilon(y|a=1,w) - \frac{d}{d\epsilon} p_{Y\epsilon}(y|a=0,w) \right) d\nu(y) \right) p_W(w) d\nu(w) \\
&= \int \left(\frac{-1}{\delta} k\left(\frac{b(w)-t}{\delta}\right) \int\int \frac{2a-1}{g(a|w)} y p_{Y\epsilon}(y|a,w) S_Y(o) d\nu(y) \right) g(a|w) p_W(w) d\nu(a,w) \\
&= \int \left(\frac{-1}{\delta} k\left(\frac{b(w)-t}{\delta}\right) \int\int \frac{2a-1}{g(a|w)} y p_{Y\epsilon}(y|a,w) \left(S(o) - \mathbb{E}[S|a,w] \right) d\nu(y) \right) g(a|w) p_W(w) d\nu(a,w) \\
&= \int \frac{-1}{\delta} k\left(\frac{b(w)-t}{\delta}\right) \frac{2a-1}{g(a|w)} S(o) p(o) d\nu(o) - \int \frac{1}{\delta} \bar{Q}(a,w) S(o) p(o) d\nu(o) \\
&= \mathbb{E}\left[\frac{-1}{\delta} k\left(\frac{b(W)-t}{\delta}\right) \frac{2A-1}{g(A|W)} (Y - \bar{Q}(A,W)) S(O) \right]
\end{aligned}
$$

Combining this result with term (3.40), we get

$$
\lim_{\epsilon\longrightarrow 0} \frac{\Psi_{\delta,t}(P_\epsilon) - \Psi_{t,\delta}(P)}{\epsilon} = \langle D^\star_{\Psi_{\delta,t}}(P), S \rangle_{L_0^2(P)}
$$

where

$$
\mathbf{D}^\star_{\boldsymbol{\Psi}_{\delta,\mathbf{t}}}(\mathbf{P})(\mathbf{O}) = \quad \frac{\mathbf{-1}}{\delta} \mathbf{k}\left(\frac{\mathbf{b(W) - t}}{\delta}\right) * \frac{\mathbf{2A - 1}}{\mathbf{g(A|W)}} (\mathbf{Y} - \bar{\mathbf{Q}}(\mathbf{A}, \mathbf{W})) + \int \frac{\mathbf{1}}{\delta} \mathbf{k}\left(\frac{\mathbf{x - t}}{\delta}\right) \mathbb{I}(\mathbf{b(W)} \le \mathbf{x}) \mathbf{dx} - \boldsymbol{\Psi}_{\delta,\mathbf{t}}
$$

And this is the efficient influence curve since the canonical gradient is the only gradient for a non-parametric model.

□

3.2 Remainder Term Derivations and Robustness Analysis

Our remainder term revolves around the fact we have solved the efficient influence curve equation for the TMLE updated initial estimate, P_n^*, of the true data generating distribution, P_0. We thus obtain a second order expansion by virtue of the fact $P_n^* D^\star(P_n^*) = 0$

$$\begin{aligned} \sqrt{n}\left(\Psi(P_n^*) - \Psi(P_0)\right) &= \sqrt{n}(P_0 - P_n^*)D^\star(P_n^*) + \sqrt{n}R_2(P_n^*, P_0) && (3.41) \\ \implies R_2(P_n^*, P_0) &= \sqrt{n}\left(\Psi(P_n^*) - \Psi(P_0) - P_0 D^\star(P_n^*)\right) && (3.42) \end{aligned}$$

The reader may recall the three conditions assuring asymptotic efficiency of the TMLE estimator in 1.2.1. Here we will focus on the 2^{nd} condition in section 1.2.1 regarding the remainder term for each of the three parameters of interest in this paper.

3.2.1 TML Estimator of VTE

Theorem 3.2.1. *If P_0 is the true distribution, it is necessary to estimate the true TE function b_0 with b_n^* so that $\|b_n^* - b_0\|_{L^2(P_0)} = o_P(n^{-0.25})$ in order for TMLE to be a consistent asymptotically efficient estimator under a known treatment mechanism, g_0. If g_0 is unknown, we also need $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)}\|g_n - g_0\|_{L^2(P_0)} = o_P(\frac{1}{n^{0.5}})$. That is, If the first factor is $o_P(n^{r_{\bar{Q}}})$ and the second is $o_P(n^{r_g})$, then we require $r_{\bar{Q}} + r_g \leq -0.5$*

Proof. For this discussion we will drop the subscript, n, and superscript, $\star$ in $P_n^\star$ and merely consider, P, as an estimate of the truth, P_0. We will use $b(W)$ to denote the TE function where the conditional expectation is with respect to distribution, P, ie the estimated TE function, and $b_0(W)$ to be the true TE function. Likewise, $\mathbb{E}_0$ is the expectation with respect to the true observed data distribution, P_0, and leaving the subscript, 0, off the expectation sign means the expectation is with respect to P.

$$\begin{aligned} R_2(P, P_0) &= \Psi(P) - \Psi(P_0) + P_0\left(D^\star(P)\right) \\ &= \mathbb{E}\left(b(W) - \mathbb{E}b(W)\right)^2 - \mathbb{E}_0\left(b_0(W) - \mathbb{E}_0 b_0(W)\right)^2 + \\ &\mathbb{E}_0\left[2\left(b(W) - \mathbb{E}b(W)\right)\frac{2A-1}{g(A|W)}\left(Y - \bar{Q}(A,W) + \left(b(W) - \mathbb{E}b(W)\right)^2 - \Psi(P)\right]\right. \\ &= -\mathbb{E}_0\left(b_0(W) - \mathbb{E}_0 b_0(W)\right)^2 + \\ &\mathbb{E}_0\left[2\left(b(W) - \mathbb{E}b(W)\right)\frac{2A-1}{g(A|W)}\left(Y - \bar{Q}(A,W)\right) + \left(b(W) - \mathbb{E}b(W)\right)^2\right] \\ &= \mathbb{E}_0\left[\left(b(W) - \mathbb{E}b(W)\right)^2 - \left(b_0(W) - \mathbb{E}_0 b_0(W)\right)^2\right] + \\ &\mathbb{E}_0\mathbb{E}_0\left[2\left(b(W) - \mathbb{E}b(W)\right)\frac{2A-1}{g(A|W)}\left(\bar{Q}_0(A,W) - \bar{Q}(A,W)\right)|W\right] \\ &= \mathbb{E}_0\left[\left(b(W) - \mathbb{E}b(W)\right)^2 - \left(b_0(W) - \mathbb{E}_0 b_0(W)\right)^2\right] + \\ &+ \mathbb{E}_{P_W}\left[2\left(b(W) - \mathbb{E}b(W)\right)\left(\frac{g_0(1|W)}{g(1|W)}\left(\bar{Q}_0(1,W) - \bar{Q}(1,W)\right) - \frac{g_0(0|W)}{g(0|W)}\left(\bar{Q}_0(0,W) - \bar{Q}(0,W)\right)\right)\right] \\ &= \mathbb{E}_0\left[\left(b(W) - \mathbb{E}b(W)\right)^2 - \left(b_0(W) - \mathbb{E}_0 b_0(W)\right)^2 + 2\left(b_0(W) - b(W)\right)\left(b(W) - \mathbb{E}b(W)\right)\right] \\ &+ \mathbb{E}_0\left[2\left(b(W) - \mathbb{E}b(W)\right)\left(\frac{g_0(1|W) - g(1|W)}{g(1|W)}\left(\bar{Q}_0(1,W) - \bar{Q}(1,W)\right) - \frac{g_0(0|W) - g(0|W)}{g(0|W)}\left(\bar{Q}_0(0,W) - \bar{Q}(0,W)\right)\right)\right] \\ &= \left(\mathbb{E}_0 b_0(W) - \mathbb{E}b(W)\right)^2 - \mathbb{E}_0\left(b_0(W) - b(W)\right)^2 \qquad (3.43) \end{aligned}$$

$$+\mathbb{E}_0\left[2\left(b(W)-\mathbb{E}b(W)\right)\left(\frac{g_0(1|W)-g(1|W)}{g(1|W)}\left(\bar{Q}_0(1,W)-\bar{Q}(1,W)\right)-\frac{g_0(0|W)-g(0|W)}{g(0|W)}\left(\bar{Q}_0(0,W)-\bar{Q}(0,W)\right)\right)\right]$$

We can regard the $\left(\mathbb{E}_0 b_0(W)-\mathbb{E}b(W)\right)^2$ term in 3.43 and notice that for an unknown, g_0, it is well-known that the double robustness of TMLE in estimating the causal risk difference, $\mathbb{E}_0 b_0(W)$, implies that if we estimate both g_0 and $\bar{Q}_0$ so that the product of the respective L_2 rates of convergence is $o(n^{-0.5})$, then we obtain $\sqrt{n}\left(\mathbb{E}_0 b_0(W)-\mathbb{E}b(W)\right)\overset{D}{\Longrightarrow}N\left[0,var_0(D_1^\star(P_0))\right]$ where $D_1^\star(P_0)$ is the efficient influence curve for the causal risk difference. We therefore know $\mathbb{E}_0 b_0(W)-\mathbb{E}b(W\xrightarrow{p}0$ and by slutsky's theorem,
$\sqrt{n}\left(\mathbb{E}_0 b_0(W)-\mathbb{E}b(W)\right)^2\overset{D}{\Longrightarrow}0$. Therefore this term poses no additional problem to the rest of the terms.

Now we can address the standard "double robust" term:

$$\begin{aligned}&\mathbb{E}_0\left[2\left(b(W)-\mathbb{E}b(W)\right)\left(\frac{g_0(1|W)-g(1|W)}{g(1|W)}\left(\bar{Q}_0(1,W)-\bar{Q}(1,W)\right)-\frac{g_0(0|W)-g(0|W)}{g(0|W)}\left(\bar{Q}_0(0,W)-\bar{Q}(0,W)\right)\right)\right]\\ \leq&K\mathbb{E}_0\left[\left|\frac{g_0(1|W)-g(1|W)}{g(1|W)}\left(\bar{Q}_0(1,W)-\bar{Q}(1,W)\right)\right|+\left|\frac{g_0(0|W)-g(0|W)}{g(0|W)}\left(\bar{Q}_0(0,W)-\bar{Q}(0,W)\right)\right|\right]\\ \leq&K\mathbb{E}_0\left|\frac{g_0(A|W)-g(A|W)}{g(A|W)g_0(A|W)}\left(\bar{Q}_0(A,W)-\bar{Q}(A,W)\right)\right|\\ \leq&K\|g_0(A|W)-g(A|W)\|_{L^2(P_0)}\|\bar{Q}_0(A,W)-\bar{Q}(A,W)\|_{L^2(P_0)}\end{aligned}$$

where the last inequality follows from cauchy-schwarz and the strict positivity assumption on g_0. The difficult term in (11) is $\mathbb{E}_0\left(b_0(W)-b(W)\right)^2$ but if we obtain the required L^2 rates it is obvious this term will be second order and we have proven the theorem. □

3.2.2 TML Estimator of Smoothed TE CDF

Let us assume we have computed a CV-TMLE, P^*_{n,B_n} from initial estimate, P^0_{n,B_n}, of the observed data generating distribution, P_0, as defined in section 1.4.1 or a TMLE, P^*_n, for initial estimate, P^0_n as in van der Laan and Rubin, 2006 or van der Laan and Gruber, 2016. To lighten the notation, we will just call $P^*_{n,B_n}, P^0_{n,B_n}, P^*_n$ or P^0_n, P, and then the estimated TE function $b(W)=\mathbb{E}_P(Y\mid 1,W]-\mathbb{E}_P(Y\mid 0,W]=\bar{Q}(1,W)-\bar{Q}(0,W)$ and the true TE function is $b_0(W)=\mathbb{E}_{P_0}(Y\mid 1,W]-\mathbb{E}_{P_0}(Y\mid 0,W]=\bar{Q}_0(1,W)-\bar{Q}_0(0,W)$.

Lemma 3.2.2. *Assume lipschitz* $F_0=1-S$*, where* $S(t)=\mathbb{E}\mathbb{I}(b_0(W)>t)$ *and assume WLOG the support of the kernel is* $[-1,1]$
then $P_0\mathbb{I}(b_0(W)>t+\delta,b(W)<t+\delta)=O(\|b_0-b\|_\infty$
proof:

$$\begin{aligned}&\int\mathbb{I}(b_0(w)>t+\delta,b(w)<t+\delta)dP_{W,0}(w)\\ &=\int\mathbb{I}(b_0(w)>t+\delta,b(w)<t+\delta)\mathbb{I}(b_0(w)-b(w)>b_0(w)-(t+\delta))dP_{W,0}(w)\\ &\leq\int\mathbb{I}(b_0(w)>t+\delta,b(w)<t+\delta)\mathbb{I}(\|b_0-b\|_\infty>b_0(w)-(t+\delta))dP_{W,0}(w)\\ &\leq Pr(t+\delta<b_0(W)<\|b_0-b\|_\infty+t+\delta)\\ \text{Lipschitz}\implies&\leq L\|b_0-b\|_\infty+O(\|b_0-b\|^2_\infty)\end{aligned}$$

QED

Theorem 3.2.3. *$R_2(P, P_0) = P_0 D^*(P) + \Psi_{\delta,t}(P) - \Psi_{\delta,t}(P_0)$ has the following order:*

1. $\frac{1}{\delta} O \left(\|g - g_0\|_{L^2_{P_0}} \|\bar{Q} - \bar{Q}_0\|_{L^2_{P_0}} \right) + \frac{1}{\delta} O \left(\|b - b_0\|^2_\infty \right)$ *and*
2. $\frac{1}{\delta} O \left(\|g - g_0\|_{L^2_{P_0}} \|\bar{Q} - \bar{Q}_0\|_{L^2_{P_0}} \right) + \frac{1}{\delta^2} O \left(\|b - b_0\|^2_{L^2_{P_0}} \right)$

and the remainder term is given by

$$\begin{aligned} &\frac{-1}{\delta} \int \left[k \left(\frac{b(w) - t}{\delta} \right) \frac{2a - 1}{g(a \mid w)} (g_0(a|w) - g(a|w)) \, (b_0(w) - b(1, w)) \right] p_{A,W}(a, w) dv(a, w) \qquad (3.44) \\ &+ \frac{1}{\delta} \int \left[\int_{b(w)}^{b_0(w)} k \left(\frac{x - t}{\delta} \right) dx + k \left(\frac{b(w) - t}{\delta} \right) (b(w) - b_0(w)) \right] p_W(w) dv(w) \end{aligned}$$

Proof.

$$\begin{aligned} &R_2(P_0 P) = P_0 D^*(P) + \Psi(P) - \Psi(P_0) \\ &= \int \left[\frac{-1}{\delta} k \left(\frac{b(w) - t}{\delta} \right) \frac{2a - 1}{g_n(a|a)} \left(y - \bar{Q}(a, w) \right) + \int \frac{1}{\delta} k \left(\frac{x - t}{\delta} \right) \mathbb{I} \left(b(w) > x \right) dx - \int_x \frac{1}{\delta} k \left(\frac{x - t}{\delta} \right) \mathbb{I}(b_0(w) > x) dx \right] p(o) dv(o) \\ &= \int \left[\frac{-1}{\delta} k \left(\frac{b(w) - t}{\delta} \right) \left(\frac{2a - 1}{g_n(a|w)} \left(y - \bar{Q}(a, w) \right) \right) + \int \frac{1}{\delta} k \left(\frac{x - t}{\delta} \right) \left(\mathbb{I} \left(b(w) > x \right) - \mathbb{I}(b_0(w) > x) \right) dx \right] p(o) dv(o) \\ &= \frac{-1}{\delta} \int \left[k \left(\frac{b(w) - t}{\delta} \right) \left(\left(\frac{g_0(1|w)}{g(1|w)} \right) (\bar{Q}_0(1, w) - \bar{Q}(1, w)) - \left(\frac{g_0(0|w)}{g(0|w)} \right) (\bar{Q}_0(0, w) - \bar{Q}(0, w)) \right) \right. \\ &+ \left. \int_{b(w)}^{b_0(w)} k \left(\frac{x - t}{\delta} \right) dx \right] p(o) dv(o) \\ &= \frac{-1}{\delta} \int \left[k \left(\frac{b(w) - t}{\delta} \right) \left(\left(\frac{g_0(1|w)}{g(1|w)} - 1 \right) (\bar{Q}_0(1, w) - \bar{Q}(1, w)) - \left(\frac{g_0(0|w)}{g(0|w)} - 1 \right) (\bar{Q}_0(0, w) - \bar{Q}(0, w)) \right) \right] p_W(w) dv(w) \qquad (3.45) \\ &+ \frac{1}{\delta} \int \left[\int_{b(w)}^{b_0(W)} k \left(\frac{x - t}{\delta} \right) dx + k \left(\frac{b(w) - t}{\delta} \right) (b(w) - b_0(w)) \right] p_W(w) dv(w) \qquad (3.46) \end{aligned}$$

(3.45) becomes (reftermrobust) and will disappear if g_0 is known. Otherwise the term is $\frac{1}{\delta} \|g - g_0\|_{L^2_{p_0}} \|\bar{Q} - \bar{Q}_0\|_{L^2_{P_0}}$ by cauchy-schwarz. We can now divide the W space into disjoint parts and integrate (3.46):
a) $t - \delta < b_0(w) < t + \delta$:
Assuming F_0 is lipschitz, we have as follows and k to have a bounded derivative,

$$\frac{1}{\delta} \int \mathbb{I}(t - \delta < b_0(w) \leq t + \delta) * \left[\int_{b_0(w)}^{b(w)} k \left(\frac{x - t}{\delta} \right) dx + k \left(\frac{b(w) - t}{\delta} \right) (b_0(w) - b(w)) \right] p_W(w) dv(w)$$

taylor expanding $k \left(\frac{x - t}{\delta} \right)$ about $\frac{b(w) - t}{\delta}$ we obtain

$$\frac{1}{\delta} \int \mathbb{I}(t - \delta < b_0(w) \leq t + \delta) * \int_{b_0(w)}^{b(w)} k' \left(\gamma \left(x, b(w), \delta \right) \right) \left(\frac{x - b(w)}{\delta} \right) p_W(w) dv(w)$$

where $\gamma \left(x, b(w), \delta \right)$ is an intermediary point

$$\leq C \int \mathbb{I}(t - \delta < b_0(w) \leq t + \delta) * \left(\frac{b_0(w) - b(w)}{\delta} \right)^2 p_W(w) dv(w) \text{ which proves 2.} \qquad (3.47)$$

or $\leq \frac{C}{\delta^2}\left(F_0(t+\delta) - F_0(t-\delta)\right) * \|b_0 - b\|_\infty^2$

employing the Lipschitz condition for F_0 we arrive at $\leq \frac{1}{\delta} O(\|b - b_0\|_\infty^2)$

b) $b_0(w) > t+\delta, b(w) \leq t+\delta$

$$\frac{1}{\delta}\int \mathbb{I}(b_0(w) > t+\delta, b(w) \leq t+\delta) * \left[\int_{b_0(w)}^{b(w)} k\left(\frac{x-t}{\delta}\right)dx + k\left(\frac{b(w)-t}{\delta}\right)(b_0(w) - b(w))\right] p_W(w)dv(w) \tag{3.48}$$

$$= \frac{1}{\delta}\int \left[\int_{b_0(w)}^{b(w)} \left(k\left(\frac{b(w)-t}{\delta}\right) + \left(\frac{x-t}{\delta}\right)k'\left(\gamma\left(b(w), x, t, h\right)\right)\right)dx + k\left(\frac{b(w)-t}{h}\right)(b_0(w) - b(w))\right] *$$
$$\mathbb{I}(b_0(w) > t+\delta, b(w) \leq t+\delta)p_W(w)dv(w)$$
$$\leq \frac{1}{\delta}\int \frac{\mathbb{I}(t+\delta < b_0(w) < t+\delta+\|b_0 - b\|_\infty)}{\delta} * C(b_0(w) - b(w))^2 p_W(w)dv(w) \text{ which proves 2.} \tag{3.49}$$

or $\frac{1}{\delta}O\|b_0 - b\|_\infty^2$which proves 1. if we apply lemma 3.2.2 from line 3.48

c) $b_0(w) \leq t-\delta, b(w) > t-\delta$ This region follows identically to b).
d) for the cases, $b(w)$ and $b_0(w) < t-\delta$ or $b(w)$ and $b_0(w) > t+\delta$, we can notice
$\left[\int_{b_0(w)}^{b(w)} k\left(\frac{x-t}{\delta}\right)dx + k\left(\frac{b(w)-t}{\delta}\right)(b_0(w) - b(w))\right] p_W(w)dv(w) = 0.$ □

Theorem 3.2.4. *The asymptotic variance of our CV-TMLE estimator is of order* $1/\delta$ *if we satisfy the CV-TMLE conditions of section 1.2.1.*

Proof. We will compute the variance of the efficient influence curve and show it is of order $1/\delta$, thus proving the point.

$$\begin{aligned}
&\mathbb{E}\frac{1}{\delta^2}k^2\left(\frac{b(W)-t}{\delta}\right) * \left[\frac{2A-1}{g(A|W)}(Y - \bar{Q}_n^*(A,W))\right]^2 \\
&\leq \mathbb{E}\frac{\mathbf{C}}{\boldsymbol{\delta}^{\mathbf{2}}}\mathbf{k}^{\mathbf{2}}\left(\frac{\mathbf{b}(\mathbf{W})-\mathbf{t}}{\boldsymbol{\delta}}\right) \\
&= \mathbb{E}\frac{C}{\delta^2}\frac{d}{db(W)}\int_{-\infty}^{\infty}\mathbb{I}(b(W) \leq x)k^2\left(\frac{x-t}{\delta}\right)dx \\
&= \mathbb{E}\frac{C}{\delta}\frac{d}{db(W)}\int_{-\infty}^{\infty}\mathbb{I}\left(\frac{b(W)-t}{\delta} \leq y\right)k^2\left(y\right)dy \\
&= O\left(1/\delta\right)
\end{aligned}$$

Now, we assume k has finite support, WLOG, between $[-1, 1]$ and that $|k^2(x)| \leq M$

$$\begin{aligned}
\mathbb{E}\frac{\mathbf{C}}{\boldsymbol{\delta}^{\mathbf{2}}}\mathbf{k}^{\mathbf{2}}\left(\frac{\mathbf{b}(\mathbf{W})-\mathbf{t}}{\boldsymbol{\delta}}\right) \leq &\mathbb{E}\frac{\mathbf{C_1}}{\boldsymbol{\delta}^{\mathbf{2}}}\mathbb{I}\left(-\mathbf{1} \leq \frac{\mathbf{b}(\mathbf{W})-\mathbf{t}}{\boldsymbol{\delta}} \leq \mathbf{1}\right) \\
= &\mathbb{E}\frac{\mathbf{C_1}}{\boldsymbol{\delta}^{\mathbf{2}}}\mathbb{I}\left(-\boldsymbol{\delta}+\mathbf{t} \leq \mathbf{b}(\mathbf{W}) \leq \boldsymbol{\delta}+\mathbf{t}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{C_1}{\delta^2} \left[F\left(\delta + t\right) - F\left(-\delta + t\right) \right] \\
&\overset{Lipschitz}{\leq} \frac{C_2}{\delta}
\end{aligned}
$$

□

Theorem 3.2.5. *Let $\Psi_t(P) = \mathbb{E}_{P_W} I(b(W) \leq t)$, the unsmoothed TE CDF at TE level, t. The bias, $\Psi_t(P_0) - \Psi_{\delta,t_i}(P_0)$ is of order δ^J, where J is the order of the kernel (power of the kernel's first non-zero moment) and we assume the TE CDF to have J continuous derivatives. Without smoothness, a lipschitz condition on the TE CDF assures that the bias is of order δ.*

Proof.

$$
\begin{aligned}
&\mathbb{E}_{P_W} I(b(W) \leq t) - \mathbb{E}_{P_W} \int \frac{1}{\delta} k\left(\frac{x-t}{\delta}\right) I(b(W) \leq x) dx \\
&\overset{fubini}{=} F(t) - \int \frac{1}{\delta} k\left(\frac{x-t}{\delta}\right) F(x) dx \\
&= F(t) - \int k\left(y\right) F(y\delta + t) dy \\
&= \int k\left(y\right) \left[F(t) - F(y\delta + t)\right] dy \\
&= \int k\left(y\right) \left[\sum_{i=1}^{\infty} F^{(i)}(t)(y\delta)^i / i!\right] dy
\end{aligned}
$$

□

Generating Kernels

We generate a kernel of order $K+1$ as follows. by generating symmetric polynomial kernels of finite support, the integration can be obtained via an explicit formula and is thus much faster and more accurate than numerical integration. We form polynomials of the form $k(x) = \sum_{i=0}^{K+2} a_i x^{2i}$ where the support of the kernel is from $-R$ to R. The kernel $k(\cdot)$ is of course orthogonal to any odd power. To make it order $K+1$ for K an even positive number, we solve the following equations.

1. make sure the kernel is 0 at the end pts of the support:

$$
\sum_{i=0}^{K+2} a_i R^{2i} = 0
$$

2. make sure the kernel has derivative 0 at the end pts of the support in consideration of

the remainder term analysis:

$$\sum_{i=0}^{K+1} 2a_i R^{2i+1} = 0$$

3. To enforce the necessary orthogonality, we solve for $K > 0$ and each r in the $2, 4, ..., 2K$

$$2\sum_{i=0}^{K+2} a_i \frac{R^{2i+1+r}}{2i+1+r} = 0$$

3.2.3 TML Estimator of Transported SDE, SIE, Restricted and Unrestricted Models

Here we derive the remainder term for the TML estimator of the parameter for the unrestricted model introduced in Chapter 2.1, which has influence curve derived in section 3.1.2. We note, for the same argument used here, the same robustness conditions hold for the restricted model and corresponding TML estimator using the influence curve in 3.1.3. We remind the reader that we have a time ordering of variables from most recent as follows: Y, M, Z, A, W, S. $\bar{M} = (M, Z, A, W, S)$, i.e., the "bar" means the variable and its parents.

The remainder term is given by

$$R_2(P, P_0) = \Psi(P) - \Psi(P_0) + P_0 D^*(P)$$

where we consider P as and estimate of P_0, which is either the TMLE updated estimate of P_0 or, in the case of the EE estmator, the initial estimate of P_0.

Theorem 3.2.6. *For model $\mathcal{M}_{II}$, we have the following:*

$$\begin{aligned}
&R_2(P, P_0) \\
=&\mathbb{E}_{P_0}(\bar{Q}_Y - \bar{Q}_{Y,0})(\bar{M})\Big[h_1(O)(g_{M,0} - g_M)(M \mid \bar{Z}) + h_2(O)(p_{Z,0} - p_Z)(Z \mid \bar{A}) \\
&+ h_3(O)(p_{A,0} - p_A)(A \mid \bar{W})h_4(O)(p_{S|W,0} - p_{S|W})(S \mid W)\Big] \qquad (3.50) \\
&+ \mathbb{E}_{P_0} h_5(O)(\bar{Q}_{Z,0} - \bar{Q}_Z)(\bar{A})(p_{A,0} - p_A)(A \mid \bar{W}) \qquad (3.51) \\
\leq& k\sum_{i=1}^{4} \|\bar{Q}_Y - \bar{Q}_{Y,0}\|_{L^2(P_0)}\|f_{i,0} - f_i\|_{L^2(P_0)} + k\|\bar{Q}_{Z,0} - \bar{Q}_Z\|_{L^2(P_0)}\|f_{3,0} - f_3\|_{L^2(P_0)}
\end{aligned}$$

where we substituted the following: $f_{1,0}(o) = g_{M,0}(m \mid z, x, w, s)$, $f_{2,0} = p_{Z,0}(z \mid a, w, 1)$, $f_{3,0} = p_{A,0}(x = a \mid w, s)$, $f_{4,0}(o) = p_{S|W,0}(x \mid w, s)$. Dropping the subscript, 0, indicates the estimated counterpart. Also, h_i is a bounded function by the positivity assumption (see section 2.4) and thus the last inequality holds with a sufficiently large k.

Proof.

$$R_2 = \Psi(P) - \Psi(P_0) + \mathbb{E}_{P_0}\left\{(Y - \bar{Q_Y}(\bar{M})\,\frac{\hat{g}_{a^*,W,s}(M \mid W)p_Z(Z \mid A, W, S = 0)p_{S|W}(S = 0 \mid W)\mathbb{I}(S = 1, A = a)}{g_M(M \mid \bar{Z})p_Z(Z \mid \bar{A})p_A(A \mid \bar{W})p_{S|W}(S \mid W)P(S = 0)}\right\}$$

$$+ \mathbb{E}_{P_0}\left\{(\bar{Q}_M(\bar{Z}) - \bar{Q}_Z(\bar{A}))\,\frac{\mathbb{I}(S = 0, A = a)}{p_A(A \mid \bar{W})P(S = 0)}\right\} + \mathbb{E}_{P_0}\left\{(\bar{Q}_Z(a, W, S) - \Psi_n)\,\frac{\mathbb{I}(S = 0)}{P(S = 0)}\right\}$$

$$
\begin{aligned}
&\approx \mathbb{E}_{P_0}\left\{\big(Y-\bar{Q}(\bar{M})\big)\,\frac{\hat{g}_{a^*,W,s}(M\mid W)p_Z(Z\mid A,W,S=0)p_{S|W}(S=0\mid W)\mathbb{I}(S=1,A=a)}{g_M(M\mid \bar{Z})p_Z(Z\mid \bar{A})p_A(A\mid \bar{W})p_{S|W}(S\mid W)P(S=0)}\right\}\\
&+\mathbb{E}_{P_0}\left\{\big(\bar{Q}_M(\bar{Z})-\bar{Q}_Z(\bar{A})\big)\,\frac{\mathbb{I}(S=0,A=a)}{p_A(A\mid \bar{W})P(S=0)}\right\}+\mathbb{E}_{P_0}\left\{\big(\bar{Q}_Z(a,W,S)-\Psi_0\big)\,\frac{\mathbb{I}(S=0)}{P(S=0)}\right\}\\
&=\underbrace{\mathbb{E}_{P_0}\left\{(\bar{Q}_{Y,0}-\bar{Q}_Y)(\bar{M})\frac{\hat{g}_{a^*,W,s}(M\mid W)p_Z(Z\mid A,W,S=0)p_{S|W}(S=0\mid W)\mathbb{I}(S=1,A=a)}{g_M(M\mid \bar{Z})p_Z(Z\mid \bar{A})p_A(A\mid \bar{W})p_{S|W}(S\mid W)P(S=0)}\right\}}_{\text{term 3}}\\
&+\underbrace{\mathbb{E}_{P_0}\left\{\big(\bar{Q}_{M,0}(\bar{Z})-\bar{Q}_Z(\bar{A})\big)\,\frac{\mathbb{I}(S=0,A=a)}{p_A(A\mid \bar{W})P(S=0)}\right\}+P_0\left\{\big(\bar{Q}_Z(a,W,S)-\bar{Q}_{Z,0}(a,W,S)\big)\,\frac{\mathbb{I}(S=0)}{P(S=0)}\right\}}_{\text{term 4}}\\
&+\underbrace{\mathbb{E}_{P_0}\left\{(\bar{Q}_M-\bar{Q}_{M,0})(\bar{Z})\frac{\mathbb{I}(S=0,A=a)}{p_A(A\mid \bar{W})P(S=0)}\right\}}_{\text{term 5}}
\end{aligned}
$$

We can see term 3.56 in the proof statement is term 4 above, where the positivity assumption guarantees $h_5(o)\le k$ for some $k<\infty$. We will now rearrange term 3.55 and prove the result.

$$
\begin{aligned}
&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid w)\frac{p_{W,0}(w)}{p_S(s=0)}\Bigg[\frac{g_{M,0}(m\mid z,x,w,1)}{g_M(m\mid z,x,w,1))}\,\frac{p_{Z,0}(z\mid a,w,1)}{p_Z(z\mid a,w,1)}\times\\
&\frac{p_Z(z\mid a,w,0)p_{A,0}(a\mid w,1)p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_A(a\mid w,1)p_{S|W}(1\mid w)}-\frac{p_{Z,0}(z\mid a,w,0)p_{A,0}(a\mid w,0)p_{S|W,0}(0\mid w)}{p_A(a\mid w,0)}\Bigg]dv(o)\\
=&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid w)\frac{p_{W,0}(w)p_{Z,0}(z\mid a,w,0)}{p_S(s=0)}\Bigg[\frac{g_{M,0}(m\mid z,x,w,1)}{g_M(m\mid z,x,w,1))}\times\\
&\frac{p_{A,0}(a\mid w,1)p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_A(a\mid w,1)p_{S|W}(1\mid w)}-\frac{p_{A,0}(a\mid w,0)p_{S|W,0}(0\mid w)}{p_A(a\mid w,0)}\Bigg]dv(o)\\
&+\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_Z(z\mid a,w,0)}{p_S(s=0)}\Bigg[\frac{g_{M,0}(m\mid z,x,w,1)}{g_M(m\mid z,x,w,1))}\times\\
&\frac{(p_{Z,0}-p_Z)(z\mid a,w,s)}{p_Z(z\mid a,w,1)}\,\frac{p_{A,0}(a\mid w,1)}{p_A(a\mid w,1)}\,\frac{p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_{S|W}(1\mid w)}\\
=&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid w)\frac{p_{W,0}(w)p_{Z,0}(z\mid a,w,0)}{p_S(s=0)}\Bigg[\frac{g_{M,0}(m\mid z,x,w,1)}{g_M(m\mid z,x,w,1))}\times\\
&\frac{p_{A,0}(a\mid w,1)p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_A(a\mid w,1)p_{S|W}(1\mid w)}-\frac{p_{A,0}(a\mid w,0)p_{S|W,0}(0\mid w)}{p_A(a\mid w,0)}\Bigg]dv(o)\\
&+\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)(p_{Z,0}-p_Z)(z\mid a,w,s)h_2(o)p(o)dv(o)
\end{aligned}\tag{3.52}
$$

where by the positivity assumption, $h_2(o)$ is bounded. Continuing with term 3.52:

$$
\begin{aligned}
=&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid w)\frac{p_{W,0}(w)p_{Z,0}(z\mid a,w,0)}{p_S(s=0)}\Bigg[\frac{g_{M,0}(m\mid z,x,w,1)}{g_M(m\mid z,x,w,1))}\times\\
&\frac{p_{A,0}(a\mid w,1)p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_A(a\mid w,1)p_{S|W}(1\mid w)}-\frac{p_{A,0}(a\mid w,0)p_{S|W,0}(0\mid w)}{p_A(a\mid w,0)}\Bigg]dv(o)\\
=&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(z\mid a,w,0)}{p_S(s=0)}\,\frac{(g_{M,0}-g_M)(m\mid z,x,w,1)}{g_M(m\mid z,x,w,1)})\times\\
&\frac{p_{A,0}(a\mid w,1)}{p_A(a\mid w,1)}\,\frac{p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_{S|W}(1\mid w)}+\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(z\mid a,w,0)}{p_S(s=0)}\times\\
&\Bigg[\frac{p_{A,0}(a\mid w,1)}{p_A(a\mid w,1)}\,\frac{p_{S|W,0}(1\mid w)p_{S|W}(0\mid w)}{p_{S|W}(1\mid w)}-\frac{p_{A,0}(a\mid w,0)}{p_A(a\mid w,0)}p_{S|W,0}(0\mid w)\Bigg]dv(o)
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}_{P_0}(\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{M})(g_{M,0} - g_M)(M \mid \bar{Z})h_1(O) + \int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)} \\
&\left[\frac{p_{A,0}(a \mid w, 1)}{p_A(a \mid w, 1)}\frac{p_{S|W,0}(1 \mid w)p_{S|W}(0 \mid w)}{p_{S|W}(1 \mid w)} - \frac{p_{A,0}(a \mid w, 0)}{p_A(a \mid w, 0)}p_{S|W,0}(0 \mid w)\right]dv(o) \qquad (3.53)
\end{aligned}
$$

where by the positivity assumption, h_1 is bounded. Continuing with the integral in term 3.53:

$$
\begin{aligned}
&\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\frac{p_{A,0}(a \mid w, s)}{p_A(a \mid w, s)} \\
&\left[\frac{\mathbb{I}(s=1)p_{S|W,0}(s \mid w)p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)} - \mathbb{I}(s=0)p_{S|W,0}(s \mid w)\right]dv(o) \\
=&\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\frac{(p_{A,0} - p_A)(a \mid w, s)}{p_A(a \mid w, s)}\times \\
&\left[\frac{\mathbb{I}(s=1)p_{S|W,0}(s \mid w)p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)} - \mathbb{I}(s=0)p_{S|W,0}(s \mid w)\right]dv(o) \\
&+\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\times \\
&\left[\frac{\mathbb{I}(s=1)p_{S|W,0}(s \mid w)p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)} - \mathbb{I}(s=0)p_{S|W,0}(s \mid w)\right]dv(o) \\
=&\mathbb{E}_{P_0}(\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{M})(p_{A,0} - p_A)(a \mid \bar{W})h_3(O) \\
&+\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\times \\
&\left[\frac{\mathbb{I}(s=1)p_{S|W,0}(s \mid w)p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)} - \mathbb{I}(s=0)p_{S|W,0}(s \mid w)\right]dv(o) \qquad (3.54)
\end{aligned}
$$

where h_3 is bounded by the positivity assumption. Continuing with term 3.54:

$$
\begin{aligned}
&\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\times \\
&\left[\frac{\mathbb{I}(s=1)p_{S|W,0}(s \mid w)p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)} - \mathbb{I}(s=0)p_{S|W,0}(s \mid w)\right]dv(o) \\
=&\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{W,0}(w)p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\frac{\mathbb{I}(s=1)(p_{S|W,0}(s \mid w) - p_{S|W}(s \mid w))p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)}dv(o) \\
&+\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})\hat{g}_{a^*,W}(m \mid z)\frac{p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\mathbb{I}(s=0)\left[p_{S|W}(s \mid w) - p_{S|W,0}(s \mid w)\right]p_{W,0}(w)dv(o) \\
=&\int (\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{m})(p_{S|W} - p_{S|W,0})(s \mid w)\hat{g}_{a^*,W}(m \mid z)\frac{p_{Z,0}(z \mid a, w, 0)}{p_S(s=0)}\left(\frac{\mathbb{I}(s=1)p_{S|W}(0 \mid w)}{p_{S|W}(s \mid w)} - \mathbb{I}(s=0)\right)p_{W,0}(w)dv(o) \\
=&\mathbb{E}_{P_0}(\bar{Q}_{Y,0} - \bar{Q}_Y)(\bar{M})(p_{S|W} - p_{S|W,0})(S \mid W)h_4(O)
\end{aligned}
$$

where h_4 is bounded by the positivity assumption.

□

Corollary. *Assume:*

- *A1*

$$
\begin{aligned}
||\bar{Q}_{Y,0} - \bar{Q}_{Y,n}||_{L_0^2(P_0)}||p_{M,0} - p_{M,n}||_{L_0^2(P_0)} &= \\
||\bar{Q}_{Y,0} - \bar{Q}_{Y,n}||_{L_0^2(P_0)}||p_{Z,0} - p_{Z,n}||_{L_0^2(P_0)} &= \\
||\bar{Q}_{Y,0} - \bar{Q}_{Y,n}||_{L_0^2(P_0)}||p_{A,0} - p_{A,n}||_{L_0^2(P_0)} &=
\end{aligned}
$$

$$\|\bar{Q}_{Y,0} - \bar{Q}_{Y,n}\|_{L_0^2(P_0)} \|p_{S|W,0} - p_{S|W,n}\|_{L_0^2(P_0)} = o_P(1/\sqrt{n})$$

- *A2*

$$\|\bar{Q}_{Z,0} - \bar{Q}_{Z,n}\|_{L_0^2(P_0)} \|p_{A,0} - p_{A,n}\|_{L_0^2(P_0)} = o_P(1/\sqrt{n})$$

Then $\sqrt{n}R_2(P_n, P_0) \xrightarrow{p} 0$

The proof is immediate when applying the cauchy-schwarz inequality.

Remark. Such conditions are guaranteed asymptotically when using the highly adaptive lasso to fit the regressions if the true regressions are of finite sectional variation norm and are left-hand continuous with right-hand limits (van der Laan 2016).

Remark. If A1 and A2 are satisfied, the TMLE and EE estimators will be consistent. If A1 is satisfied and we know the treatment mechanism, as in an RCT, then the TMLE and EE estimators are consistent.

Robustness for restricted model

Theorem 3.2.7. *For model $\mathcal{M}_I$, we also have the following:*

$$\begin{aligned}
&R_2(P, P_0) \\
=&\mathbb{E}_{P_0}(\bar{Q}_Y - \bar{Q}_{Y,0})(\bar{M})\Big[h_1(O)(g_{M,0} - g_M)(M \mid \bar{Z}) + h_2(O)(p_{Z,0} - p_Z)(Z \mid \bar{A}) \\
&+ h_3(O)(p_{A,0} - p_A)(A \mid \bar{W})h_4(O)(p_{S|W,0} - p_{S|W})(S \mid W)\Big] && (3.55) \\
&+ \mathbb{E}_{P_0}h_5(O)(\bar{Q}_{Z,0} - \bar{Q}_Z)(\bar{A})(p_{A,0} - p_A)(A \mid \bar{W}) && (3.56) \\
\leq& k\sum_{i=1}^{4} \|\bar{Q}_Y - \bar{Q}_{Y,0}\|_{L^2(P_0)}\|f_{i,0} - f_i\|_{L^2(P_0)} + k\|\bar{Q}_{Z,0} - \bar{Q}_Z\|_{L^2(P_0)}\|f_{3,0} - f_3\|_{L^2(P_0)}
\end{aligned}$$

where we substituted the following: $f_{1,0}(o) = g_{M,0}(m \mid z, x, w, s)$, $f_{2,0} = p_{Z,0}(z \mid a, w, 1)$, $f_{3,0} = p_{A,0}(x = a \mid w, s)$, $f_{4,0}(o) = p_{S|W,0}(x \mid w, s)$. *Dropping the subscript, 0, indicates the estimated counterpart. Also,* h_i *is a bounded function by the positivity assumption (see section 2.4) and thus the last inequality holds with a sufficiently large* k.

Proof. We remind the reader that $p_{Z/A}(z \mid w, s) = \int p_Z(z \mid x, w, s)p_A(x \mid w, s)dv(x)$.

$$\begin{aligned}
R_2 = \Psi(P) - \Psi(P_0) &+ \mathbb{E}_{P_0}\left\{\left(y - \mathbb{E}\Big[y \mid m, z, w\Big]\right)\frac{\hat{g}_{M|a^*,W,s}(m \mid w)p_Z(z \mid a_0, w, s=0)p_{S|W}(s=0 \mid w)I(s=1)}{g_{M,r}(m \mid z, w, s)p_{Z/A}(z \mid w, s)p_{S|W}(s \mid w)p_S(s=0)}\right\} \\
&+ \mathbb{E}_{P_0}\left\{(\bar{Q}_M(\bar{Z}) - \bar{Q}_Z(\bar{A}))\,\frac{\mathbb{I}(S=0, A=a)}{p_A(A \mid \bar{W})P(S=0)}\right\} + \mathbb{E}_{P_0}\left\{(\bar{Q}_Z(a, W, S) - \Psi_n)\,\frac{\mathbb{I}(S=0)}{P(S=0)}\right\} \\
&= \mathbb{E}_{P_0}\left\{\left(y - \mathbb{E}\Big[y \mid m, z, w\Big]\right)\frac{\hat{g}_{M|a^*,W,s}(m \mid w)p_Z(z \mid a_0, w, s=0)p_{S|W}(s=0 \mid w)I(s=1)}{g_{M,r}(m \mid z, w, s)p_{Z/A}(z \mid w, s)p_{S|W}(s \mid w)p_S(s=0)}\right\} \\
&+ \mathbb{E}_{P_0}\left\{(\bar{Q}_M(\bar{Z}) - \bar{Q}_Z(\bar{A}))\,\frac{\mathbb{I}(S=0, A=a)}{p_A(A \mid \bar{W})P(S=0)}\right\} + \mathbb{E}_{P_0}\left\{(\bar{Q}_Z(a, W, S) - \Psi_0)\,\frac{\mathbb{I}(S=0)}{P(S=0)}\right\} \\
&= \underbrace{\mathbb{E}_{P_0}\left\{\left(y - \mathbb{E}\Big[y \mid m, z, w\Big]\right)\frac{\hat{g}_{M|a^*,W,s}(m \mid w)p_Z(z \mid a_0, w, s=0)p_{S|W}(s=0 \mid w)I(s=1)}{g_{M,r}(m \mid z, w, s)p_{Z/A}(z \mid w, s)p_{S|W}(s \mid w)p_S(s=0)}\right\}}_{\text{term 3}}
\end{aligned}$$

$$
\begin{aligned}
&\underbrace{+\,\mathbb{E}_{P_0}\left\{(\bar{Q}_{M,0}(\bar{Z})-\bar{Q}_Z(\bar{A}))\,\frac{\mathbb{I}(S=0,A=a)}{p_A(A\mid\bar{W})P(S=0)}\right\}+P_0\left\{(\bar{Q}_Z(a,W,S)-\bar{Q}_{Z,0}(a,W,S))\,\frac{\mathbb{I}(S=0)}{P(S=0)}\right\}}_{\text{term 4}}\\
&\underbrace{+\,\mathbb{E}_{P_0}\left\{(\bar{Q}_M-\bar{Q}_{M,0})(\bar{Z})\frac{\mathbb{I}(S=0,A=a)}{p_A(A\mid\bar{W})P(S=0)}\right\}}_{\text{term 5}}
\end{aligned}
$$

Identically to theorem 3.2.6, term 4 is handled. Now we can regard terms 3 and 5.

$$
\begin{aligned}
&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(Z\mid a,w,0)}{p_S(s=0)}\left[\frac{p_{Z,0}(Z\mid w,1)}{p_Z(Z\mid w,1)}\times\right.\\
&\left.\frac{p_{S\mid W,0}(1\mid w)p_{S\mid W}(0\mid w)}{p_{S\mid W}(1\mid w)}-\frac{p_{A,0}(a\mid w,0)}{p_A(a\mid w,0)}p_{S\mid W,0}(0\mid w)\right]dv(o)\\
=&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(Z\mid a,w,0)}{p_S(s=0)}\,\frac{p_{Z,0}(Z\mid w,1)-p_Z(Z\mid w,1)}{p_Z(Z\mid w,1)}\times\\
&\frac{p_{S\mid W,0}(1\mid w)p_{S\mid W}(0\mid w)}{p_{S\mid W}(1\mid w)}dv(o)\\
&+\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(Z\mid a,w,0)}{p_S(s=0)}\left[\frac{p_{S\mid W,0}(1\mid w)p_{S\mid W}(0\mid w)}{p_{S\mid W}(1\mid w)}\right.\\
&\left.-\frac{p_{A,0}(a\mid w,0)}{p_A(a\mid w,0)}p_{S\mid W,0}(0\mid w)\right]dv(o)\\
=&\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(Z\mid a,w,0)}{p_S(s=0)}\times\\
&\frac{p_{Z,0}(Z\mid x,w,1)(p_{A,0}(x\mid w,1)-p_A(x\mid w,1))+p_A(x\mid w,1)(p_{Z,0}(z\mid x,w,1)-p_Z(z\mid x,w,1))}{p_Z(Z\mid w,1)}\times\\
&\frac{p_{S\mid W,0}(1\mid w)p_{S\mid W}(0\mid w)}{p_{S\mid W}(1\mid w)}dv(o)\\
&+\int(\bar{Q}_{Y,0}-\bar{Q}_Y)(m,z,x,w)\hat{g}_{a^*,W}(m\mid z)\frac{p_{W,0}(w)p_{Z,0}(Z\mid a,w,0)}{p_S(s=0)}\left[\frac{p_{S\mid W,0}(1\mid w)p_{S\mid W}(0\mid w)}{p_{S\mid W}(1\mid w)}\right.\\
&\left.-\frac{p_{A,0}(a\mid w,0)}{p_A(a\mid w,0)}p_{S\mid W,0}(0\mid w)\right]dv(o)
\end{aligned}
$$

From here, a very similar argument to theorem 3.2.6 proves the result.

□

3.3 Logistic Regression Plug-in Estimator Inference

3.3.1 defining the parameter of interest

Let us consider the nonparametric model, $\mathcal{M}$. We will employ basic statistics to obtain the influence curve for a logistic regression plug-in estimator for both the mean and variance of the TE function, $b(W)=\mathbb{E}[Y\mid A=1,W]-\mathbb{E}[Y\mid A=0,W]$. We define the plug-in estimator as plugging in the outcome conditional density given by the MLE for β where

$$\beta=\underset{\gamma}{argmin}[-\mathbb{E}_{P_{A,W}}\mathbb{E}_{P_{\gamma,Y\mid A,W}}log(p_\gamma(Y\mid A,W)p_{A,W}(A,W))]$$

Y and A are binary and $P_\gamma(Y \mid A, W)$ is defined as a conditional density,

$$p_\gamma(Y \mid A, W) = expit\left(m(A, W|\gamma)\right)^Y \left(1 - expit\left(m(A, W|\gamma)\right)\right)^{1-Y}$$

for a fixed function $m(\cdot|\cdot)$.
$\Psi_1(P) = \mathbb{E}_W b_\beta(W)$ and
$\Psi_2(P) = \mathbb{E}_W(b_\beta(W) - \Psi_1(P))^2$
and we may consider the two dimensional parameter: $\Psi(P) = (\Psi_1(P), \Psi_2(P))$.
We note to the reader, that $b_\beta(W) = expit\left(m(1, W|\beta)\right) - expit\left(m(0, W|\beta)\right)$ is the conditional average treatment effect under the parametric model for strata, W.

3.3.2 Finding the MLE for the coefficients, β

Now if we take n iid draws of O we get that the likelihood of drawing $\{O_i\}_{i=1}^n$ is

$$\prod_{i=1}^n expit\left(m(A_i, W_i|\beta\right)^{Y_i} \left(1 - expit\left(m(A_i, W_i|\beta)\right)\right)^{1-Y_i} p_A(A_i|W_i) p_W(W_i)$$

We thus have:

$$\nabla_\beta log \prod_{i=1}^n expit\left(m(A_i, W_i|\beta\right)^{Y_i} \left(1 - expit\left(m(A_i, W_i|\beta)\right)\right)^{1-Y_i} p_A(A_i|W_i) p_W(W_i) =$$

$$\nabla_\beta \sum_{i=1}^n \left[Y_i log\left(expit\left(m(A_i, W_i|\beta)\right) + (1 - Y_i) log\left(1 - expit\left(m(A_i, W_i|\beta)\right)\right)\right] = \tag{3.57}$$

$$\sum_{i=1}^n \left(\frac{\partial m}{\partial \beta_{0,n}}, \frac{\partial m}{\partial \beta_{1,n}}, ..., \frac{\partial m}{\partial \beta_{d,n}}\right)^T \left(Y_i - expit\left(m(A_i, W_i|\beta_n)\right)\right) = 0 \tag{3.58}$$

We can now derive the multidimensional influence function via the use of a taylor series about $\mathbb{E}_{P_\beta} S_\beta(O)$ where

$$S_\beta(O) = \left(\frac{\partial m}{\partial \beta_0}, \frac{\partial m}{\partial \beta_1}, ..., \frac{\partial m}{\partial \beta_d}\right)^T \left(Y - expit\left(m(A, W|\beta)\right)\right)$$

Also note that the derivative of the log-likelihood or score, $S_\beta(O)$, has mean 0 by assumption.
So $P_\beta S_\beta(O) = \mathbb{E}_{P_\beta} S_\beta(O) = 0$.
Thus we get by virtue of β_n being the MLE:

$$\begin{aligned}
P_n S_{\beta_n} - P_\beta S_\beta &= 0 \\
P_n S_{\beta_n} - P_\beta S_\beta + P_\beta S_{\beta_n} - P_\beta S_{\beta_n} &= 0 \\
(P_n - P_\beta) S_{\beta_n} &= P_\beta \left(S_\beta - S_{\beta_n}\right) \\
\sqrt{n}(P_n - P) S_{\beta_n} &= -\sqrt{n} P_\beta \left(\nabla_\beta S_\beta\right)(\beta_n - \beta) + \sqrt{n} O_p \|\beta_n - \beta\|^2 && (3.59) \\
\implies \sqrt{n}(\beta_n - \beta) &\overset{D}{\Longrightarrow} \sqrt{n}(P_n - P_\beta)\left(-P_\beta \left(\nabla_\beta S_\beta\right)^{-1} S_\beta\right) && (3.60)
\end{aligned}$$

since we can consider the $\|\beta_n - \beta\|^2$ term as second order. Therefore

$$IC_{\beta_n}(O) = \left(-P_\beta \left(\nabla_\beta S_\beta\right)^{-1} S_\beta(O)\right)$$

is the influence curve for the maximum likelihood estimator of the truth, β. In the case of logistic regression we have $m(A, W|\beta) = X(A, W)^T\beta$ where we might have the main terms linear case, $X(A, W)^T = (1, A, W)^T$ and we consider β as a column vector of coefficients, including the intercept, β_0. However, $X(A, W)^T$ might be any combination of columns of the covariates, as in any of variables, containing interactions and so forth of the main terms in the right-hand side of our regression formula. For now we will drop the arguments in X(A,W) and just use X unless it is necessary.

$$S_\beta(O) = X\left(Y - expit(\beta^T X)\right)$$

$$\nabla_\beta S_\beta(O) = \begin{bmatrix} \nabla_\beta^T S_{\beta,0} \\ . \\ \nabla_\beta^T S_{\beta,d} \end{bmatrix} = expit(\beta^T X)(1 - expit(\beta^T X))XX^T$$

3.3.3 Influence curve for MLE estimate of $b_\beta(W_0)$

Consider the parameter $\Psi_{a,w}(P) = expit\left(m(a, w|\beta)\right)$ for fixed (a, w). $expit\left(m(a, w|\beta)\right.$ is a continuously differentiable function of β, which means we can apply the ordinary delta method as follows to find the plug-in estimator influence curve estimating $\Psi_{a,w}(P)$.

$$\begin{aligned}\bar{Q}_{\beta_n}(a, w) - \bar{Q}_\beta(a, w) &= \nabla_\beta^T \bar{Q}_\beta(a, w) \sum_{i=1}^{n} IC_{\beta_n}(O_i) + R_2(P_\beta, P_{\beta_n}) \\ &= expit(\beta^T x)(1 - expit(\beta^T x))x^T \sum_{i=1}^{n} IC_{\beta_n}(O_i) + R_2(P_\beta, P_{\beta_n})\end{aligned}$$

where $x = X(a, w)$ and $R_2 = o_p(n^{-.5})$.

We thus have the influence curve for the logistic regression MLE plug-in estimator for $b_\beta(W_0)(O)$, notated as $IC_{b_{\beta_n}(W_0)}(O)$:

$$\begin{aligned}&IC_{b_{\beta_n}(W_0)}(O) \\ =&\Big[expit\Big(\beta^T X(1, W_0)\Big)\Big(1 - expit\Big(\beta^T X(1, W_0)\Big)\Big)X(1, W_0)^T - \\ &expit\Big(\beta^T X(0, W_0)\Big)\Big(1 - expit\Big(\beta^T X(0, W_0)\Big)\Big)X(0, W_0)^T\Big]IC_{\beta_n}(O) \\ =&f_\beta(W_0)IC_{\beta_n}(O)\end{aligned}$$

3.3.4 Influence curve for MLE estimate of $b_\beta(W_0)^2$

by the delta method we easily get

$$IC_{b_{\beta_n}(W_0)^2} = 2b_\beta(W_0) IC_{b_\beta(W_0)}(O)$$

3.3.5 Telescoping to find the 2-d influence curve for 2-d parameter $\Psi(P)$

Note, that $\Psi_n = (\Psi_{1,n}, \Psi_{2,n})$ is the MLE plug-in estimate for $\Psi(P) = (\Psi_1, \Psi_2)$.

$$\begin{aligned}
\Psi_{1,n} - \Psi_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(b_{\beta_n}(W_i) - \Psi_1\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(b_{\beta_n}(W_i) - b_\beta(W_i)\right) + \underbrace{\frac{1}{n}\sum_{i=1}^{n} b_\beta(W_i) - \Psi_1}_{\text{set aside}}
\end{aligned}$$

Now we can ignore the terms that are set aside as they are part of the influence curve in the tangent space of mean 0 functions of W.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(b_{\beta_n}(W_i) - b_\beta(W_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} IC_{b_{\beta_n}(W_i)}(O_j)\right) + o_p(n^{-0.5}) \\
&= \frac{1}{n}\sum_{i=1}^{n} f_\beta(W_i)\left(\frac{1}{n}\sum_{j=1}^{n} IC_{\beta_n}(O_j)\right) + o_p(n^{-0.5}) \\
&= P f_\beta\left(\frac{1}{n}\sum_{j=1}^{n} IC_{\beta_n}(O_j)\right) + \underbrace{\left(P_n - P\right) f_\beta\left(\frac{1}{n}\sum_{j=1}^{n} IC_{\beta_n}(O_j)\right)}_{o_p(n^{-0.5})} + o_p(n^{-0.5}) \\
&= \mathbb{E} f_\beta(W)\left(\frac{1}{n}\sum_{j=1}^{n} IC_{\beta_n}(O_j)\right) + o_p(n^{-0.5})
\end{aligned}$$

Therefore, the influence for the MLE-based estimate of $\Psi_1(P)$,denoted by $IC_{\Psi_{1,n}}$, is

$$\mathbf{IC_{\Psi_{1,n}}(O) = \mathbb{E} f_\beta(W) IC_{\beta_n}(O)}$$

$$\Psi_{2,n} - \Psi_2 = \frac{1}{n}\sum_{i=1}^{n}\left(b_{\beta_n}(W_i) - \Psi_{1,n}\right)^2 - \Psi_2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(b_{\beta_n}(W_i)-\Psi_{1,n})^2-(b_\beta(W_i)-\Psi_1)^2\right]+\underbrace{\frac{1}{n}\sum_{i=1}^{n}(b_\beta(W_i)-\Psi_1)^2-\Psi_2}_{\text{set aside}}$$

regarding the terms not set aside:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[(b_{\beta_n}(W_i)-\Psi_{1,n})^2-(b_\beta(W_i)-\Psi_1)^2\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}b_{\beta_n}(W_i)^2-b_\beta(W_i)^2-\left(\frac{1}{n}\sum_{i=1}^{n}b_{\beta_n}(W_i)\right)^2+\Psi_1^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}IC_{b_\beta(W_i)}(O_j)-\frac{2}{n}\Psi_1\sum_{i=1}^{n}IC_{\Psi_1}(O_i)+o_p(n^{-0.5})\\
&=\frac{1}{n}\sum_{i=1}^{n}2b_\beta(W_i)f_\beta(W_i)\frac{1}{n}\sum_{j=1}^{n}IC_{\beta_n}(O_j)-\frac{2}{n}\Psi_1\sum_{i=1}^{n}IC_{\Psi_1}(O_i)+o_p(n^{-0.5})\\
&=P\left(b_\beta f_\beta\right)\frac{2}{n}\sum_{j=1}^{n}IC_{\beta_n}(O_j)-\frac{2}{n}\Psi_1\sum_{i=1}^{n}IC_{\Psi_1}(O_i)+o_p(n^{-0.5})+\\
&\underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}2b_\beta(W_i)f_\beta(W_i)-P\left(b_\beta(W)f_\beta(W)\right)\right]}_{o_p(n^{-0.5})}\frac{2}{n}\sum_{j=1}^{n}IC_{\beta_n}(O_j)\\
&=P\left(b_\beta f_\beta\right)\frac{2}{n}\sum_{j=1}^{n}IC_{\beta_n}(O_i)-\frac{2}{n}\Psi_1\sum_{i=1}^{n}IC_{\Psi_1}(O_i)+o_p(n^{-0.5})
\end{aligned}$$

$$\begin{aligned}
IC_{\Psi_{2,n}}(O)&=2\mathbb{E}\left[b_\beta(W)f_\beta(W)\right]IC_{\beta_n}(O)-2\Psi_1 IC_{\Psi_1}(O)\\
&=2\mathbb{E}\left[\left(b_\beta(W)-\Psi_1\right)f_\beta(W)\right]IC_{\beta_n}(O)
\end{aligned}$$

Thus we arrive at the following influence curve for the plug-in estimator of the two dimensional parameter $(\Psi_1(P),\Psi_2(P))$. Note, this IC below is written as a sum of two 2-dimensional vectors, one for the components of the IC in T_Y and the other for the components in T_W.

$$IC_{\Psi_n}(O)=\mathbb{E}\left(f_\beta(W),2(b_\beta(W)-\Psi_1(P))f_\beta(W)\right)IC_{\beta_n}(O)+\left(b_\beta(W)-\Psi_1(P),(b_\beta(W)-\Psi_1(P))^2-\Psi_2(P)\right)$$

And we are finished deriving this influence curve for the plug-in logistic regression estimator of ATE and VTE.

3.4 Canonical Least Favorable Submodels

Introduction

We offer a new way to construct a targeted maximum likelihood estimator for multidimensional parameters via defining the canonical least favorable submodel (clfm). TMLE is a plug-in estimator so it follows that we might prefer to use the same model estimate for all dimensions of a parameter of interest. The obvious example of such is a survival curve, in order to insure monotonicity of the estimates in time. The clfm leads naturally to the construction of the one-step TMLE (van der Laan and Gruber 2016). The resulting TMLE algorithm can be seen as an iterative version of the one-step TMLE in that both TMLE's use a single dimensional submodel in their construction.

The TMLE defined here-in can converge much faster than its one-step recursive counterpart when evaluating the efficient influence curve has a cost. This is due to relatively few logistic regression fits as compared to very small recursions. The procedure also enables placing the denominator of the clever covariate as an inverse weight in an offset intercept model, shown to stabilize large weights caused by near positivity violations. In addition, like the one-step TMLE, the TMLE based on a clfm involves the use of a one-dimensional submodel, which avoids high dimensional regressions to perform the targeting step in the algorithm.

3.4.1 Mapping P_n^0 to $P_n^\star$: The Targeting Step

The preceding section sketched the framework by which TMLE provides asymptotically efficient estimators for nonparametric models. Here we will explain how TMLE maps an initial estimate P_n^0 to $P_n^\star$, otherwise known as the targeting step. P_n^0 is considered to be the initial estimate for the true distribution, P_0.

Definition 3.4.1. We can define a canonical 1-dimensional locally least favorable submodel (clfm) of an estimate, P_n^0, of the true distribution as

$$\{P_{n,\epsilon}^0 \text{ s.t } \frac{d}{d\epsilon} P_n L(P_{n,\epsilon}^0)\bigg|_{\epsilon=0} = \|P_n D^\star(P_n^0)\|_2, \epsilon \in [-\delta, \delta]\} \tag{3.61}$$

where $P_{n,\epsilon}^0 = P_n^0$ and $\|\cdot\|_2$ is the euclidean norm. We consider a $d-dimensional$ parameter mapping $\Psi : \mathcal{M} \longrightarrow \mathbb{R}^d$.

This definition only slightly differs slightly from the locally least favorable submodel (lfm) defined by Mark van der Laan (van der Laan and Gruber 2016) in that we can define a clfm with only a single epsilon and where as an lfm is defined so the score with respect to the loss spans the efficient influence curve and thus will employ an epsilon of dimension at least that of the parameter of interest.

Definition 3.4.2. A Universal Least Favorable Submodel (ulfm) of P_n^0 satisfies

$$\frac{d}{d\epsilon} P_n L(P_n^\epsilon) = \|P_n D^\star(P_n^\epsilon)\|_2 \ \forall \epsilon \in (-\delta, \delta)$$

and naturally, $P_n^{\epsilon=0} = P_n^0$.

We can construct the universal least favorable submodel (ulfm) in terms of the clfm if we use the difference equation $P_n(L(P_{n,dt}^0) - L(P_n^0)) \approx \|P_n D^\star(P_n^0)\|_2 dt$, where $P_n^{dt} = P_{n,dt}^0$ is an element of the clfm of P_n^0. More generally, we can map any partition $t = m \times dt$ for an arbitrarily small, dt, to an equation $P_n(L(P_n^{t+dt}) - L(P_n^t)) \approx \|P_n D^\star(P_n^t)\|_2 dt$, where P_n^{t+dt} is an element of the clfm of P_n^t. We therefore can recursively define the integral equation: $P_n(L(P_n^\epsilon) - L(P_n^0)) = \int_0^\epsilon \|P_n D^\star(P_n^t)\|_2 dt$ and P_n^ϵ will thusly be an element of the ulfm of P_n^0. For log likelihood loss, which is valid for both continuous outcome scaled between 0 and 1 as well as binary outcomes, an analytic formula for a ulfm of distribution with density, p, is therefore defined by the density $p_\epsilon = p \times exp(\int_0^\epsilon \|D^*(P^t)\|_2 dt)$ (van der Laan and Gruber 2016) where P^{t+dt} is an element of the clfm of P^t.

In applying the one-step TMLE, when the empirical loss is minimized at a given ϵ, we will have solved, $\|P_n D^\star(P_n^\epsilon)\|_2 = 0$. Therefore, the loss is decreased and all influence curve equations are solved simultaneously with a single ϵ in one step. Specifically, $P_n D_j^\star(P_n^\star) = 0$ for all j. Thus $P_n^\star = P_n^\epsilon$ and we have defined the required TMLE mapping.

3.4.2 The Iterative Approach Offered in This Paper

With an iterative approach, we first find $P_{n,\epsilon_0}^0 = P_n^1$, that is an element of the clfm of P_n^0 such that

$$\frac{d}{d\epsilon} P_n L(P_{n,\epsilon}^0)\bigg|_{\epsilon=\epsilon_0} = 0 \tag{3.62}$$

This initializes an iterative process where by

$$\frac{d}{d\epsilon} P_n L(P_{n,\epsilon}^{j-1})\bigg|_{\epsilon=\epsilon_j} = 0. \tag{3.63}$$

where $P_{n,\epsilon}^j$ is an element of the clfm of P_n^{j-1}. When $\epsilon_j = 0$, we stop the process and our TMLE is $P_n^\star = P_n^{j-1}$.

3.4.3 CLFM Construction for Generalized Scenario

Assume we have a parameter mapping as defined in the previous section, where the data is of the form $O = (W, A, Y) \sim P_0$ where Y and A are binary and W is a vector of confounders. We consider the likelihood factored according to $p_0(w, a, y) = \bar{Q}_0(a, w)^Y (1 - \bar{Q}_0(a, w))^{1-Y} g_0(a \mid w) q_{W,0}(w)$. We also assume we have efficient influence curve for the jth component of the parameter of the form:

$$D_j^*(P_0)(O) = H_{1,j}(p_0)(A, W)(Y - \bar{Q}_0(A, W)) + H_{2,j}(p_0)(A, W)(A - g_0(A, W)) + H_{3,j}(A, W)(f(P_0)_j(A, W) - \Psi(P_0))$$

where $\Psi(P_0) = E_0[H_{2,j}(O_i(f(P_0)_j(O)]$ and $E_0[H_{j,2}(O_i) = 1$ for fixed function $H_{j,2}$. Also note the dependence of the function $H_{1,j}(p_0)$ and $H_{2,j}(p_0)$ on the distribution. Now assume we have an initial estimate of P_n^0, of P_0, via an estimate, p_n^0, of the density p_0. We define p_n^0

by estimates of factors of the likelihood. That is, $\bar{Q}_n^0 \approx \bar{Q}_0$, $g_n \approx g_0$, and $q_{W,n}$ the empirical density of W, is used to approximate $q_{W,0}$. A clfm of P_n^0 is defined by leaving $q_{W,n}$ fixed and defining

$$\bar{Q}^0_{n,\epsilon}(A,W) = expit\left(logit(\bar{Q}_n^0(A,W)) + \epsilon \left\langle H_1(P_n^0)(A,W), \frac{P_n D^*(P_n^0)}{\|P_n D^*(P_n^0)\|_2} \right\rangle_2 \right)$$

and

$$g^0_{n,\epsilon}(A \mid W) = expit\left(logit(g_n^0(A \mid W)) + \epsilon \left\langle H_2(P_n^0)(A,W), \frac{P_n D^*(P_n^0)}{\|P_n D^*(P_n^0)\|_2} \right\rangle_2 \right)$$

where $\| \cdot \|_2$ is the euclidean norm induced by dot product, $\langle \cdot, \cdot \rangle$. In the usual case we have $P_n H_{2,j}(f(P_n^0)_j - \Psi(P_n^0)) = 0$ and therefore $p^0_{n,\epsilon}$ defines an element, $P^0_{n,\epsilon}$, of a clfm of P_n^0.

3.4.4 General TMLE Algorithm using the clfm for Point Treatment Parameters

Initialization
We start the iterative process with our initial estimate p_n^0 as defined in the previous subsection.

$$\begin{aligned} P_n L(P_n^0) &= -\frac{1}{n}\sum_{i=1}^{n} \left[Y_i \text{log} \bar{Q}_n^0(A_i, W_i) + (1 - Y_i)\text{log}(1 - \bar{Q}_n^0(A_i, W_i)) \right] \\ &- \frac{1}{n}\sum_{i=1}^{n} \left[A_i \text{log} g_n^0(A_i \mid W_i) + (1 - A_i)\text{log}(1 - g_n^0(A_i \mid W_i)) \right] \\ &= L_0 \text{ our starting loss} \end{aligned}$$

The Targeting Step

Starting with $m = 0$
step 2:
Compute $H_1(P_n^m)(A,W)$, $H_2(P_n^m)(A,W)$ and $H_2(A,W)$ over the data and then check the following: If $|P_n D^\star_j(P_n^m)| < \hat{\sigma}(D^\star_j(P_n^m))/n$ for all j then $P_n^\star = P_n^m$ and go to step 4. This insures that we stop the process once the bias is second order. Note, $\hat{\sigma}(\cdot)$ refers to the sample standard deviation operator. Otherwise set $m = m + 1$ and go to step 3.
step 3 We perform a pooled logistic regression with Y as the outcome,
offset $=$ $logit(\bar{Q}_n^{m-1})(A,W)$ and so-called clever covariate,

$$\left\langle (H_1(P_n^{m-1})(A,W), \frac{P_n D(P_n^{m-1})}{\|P_n D(P_n^{m-1})\|_2} \right\rangle_2 .$$

and A as the outcome, offset $logit(g_n^{m-1})(A \mid W)$ and so-called clever covariate,

$$\left\langle (H_2(P_n^{m-1})(A,W), \frac{P_n D(P_n^{m-1})}{\|P_n D(P_n^{m-1})\|_2} \right\rangle_2.$$

Assume ϵ_j is the coefficient computed from the above pooled regression. We then update the models as per below, using euclidean inner product notation, $\langle \cdot, \cdot \rangle_2$:

$$\bar{Q}_n^m = expit\left(logit(\bar{Q}_n^{m-1}) - \epsilon_j \left\langle (H_1(P_n^{m-1})(A,W), \frac{P_n D(P_n^{m-1})}{\|P_n D(P_n^{m-1})\|_2} \right\rangle_2 \right) \tag{3.64}$$

and

$$g_n^m(A \mid W) = expit\left(logit(g^{m-1}(A \mid W)) - \epsilon_j \left\langle (H_2(P_n^{m-1})(A,W), \frac{P_n D(P_n^{m-1})}{\|P_n D(P_n^{m-1})\|_2} \right\rangle_2 \right) \tag{3.65}$$

Possible alternative targeting step to ameliorate near positivity violations

We can alternatively perform a pooled logistic regression as follows. For all observations we use Y as the outcome, offset $= logit(\bar{Q}_n^{m-1})(A,W)$. We denote the denominator of $H_{1,j}(P_n^{m-1})$ as $g_j(P_n^{m-1})$, which, in some cases is a fixed propensity score, $g(P_n^{m-1})$. We can use its inverse as a weight in a logistic regression model with covariate

$$g(P_n^{m-1})(A \mid W)^{-1}\left\langle (H_1(P_n^{m-1})(A,W), \frac{P_n D(P_n^{m-1})}{\|P_n D(P_n^{m-1})\|_2} \right\rangle_2.$$

We then stack all observations using A as the outcome, offset, $logit(g_n^{m-1})(A \mid W)$ and so-called clever covariate,

$$\left\langle (H_2(P_n^{m-1})(A,W), \frac{P_n D(P_n^{m-1})}{\|P_n D(P_n^{m-1})\|_2} \right\rangle_2.$$

Thus we use a weight of 1 for when A is the outcome because $H_2(P_n^{m-1})(A,W)$ generally does not have large values. We then update the models similarly as before upon solving for the coefficient ϵ_j. With either regression scheme we solve the same score equation so either are appropriate for the targeting step.

Once we are done with the targeting step we define the distribution, P_n^m, via its density:

$$p_n^m(W,A,Y) = \bar{Q}_n^m(A,W)^Y(1-\bar{Q}_n^m(A,W))^{1-Y}g_n^m(A|W)q_n(W)$$

where $q_{W,n}$ is the empirical density. Return to step 2.

step 4

Our estimate is $\hat{\Psi}(P_n) = \Psi(P_n^\star)$ which is really only dependent on $\bar{Q}_n^\star$ and the empirical distribution.

R Software employing the clfm

Currently there are three packages which employ the iterative TMLE as presented in this paper for parameters with influence curves of the form as in this paper. Note to the reader, we have yet to implement the weighted intercept targeting scheme as discussed in step 3 of the algorithm in section 4.

- tmle3, https://github.com/tlverse/tmle3 (Coyle, Malenica, et al. 2018c)

  There are various parameters for which one can perform a TMLE estimator, including variable importance measure for continuous variables (Chambaz, Neuvial, and Laan Mark J 2012), treatment effect among the treated, causal risk difference, treatment specific mean and more.

- gentmle2, https://github.com/jeremyrcoyle/gentmle2 (Coyle and Levy 2018) The reader may note this clfm is what is employed in this R package when specifying the approach as ”line”. An lfm with epsilon the same dimension as the parameter is employed with the ”full” option. Other than causal risk difference and treatment specific mean, there is also the variance of treatment effect (**catesurvival**) as well as the mean under stochastic intervention (Diaz Muñoz and van der Laan 2012).

- cateSurvival, https://github.com/jlstiles/cateSurvival (**cateSurvival**)

  This package implements a TMLE estimator for $\Psi_{k,t}(P) = \int k\left(\frac{x-t}{h}\right)\mathbb{E}_P\mathbb{I}(B(W) > x)dx$ which is kernel-smoothed version of the non-pathwise differentiable parameter, $\mathbb{E}_P\mathbb{I}(B(W) > t)$, where $B(W)$ is the treatment effect function or TE function, defined by $\mathbb{E}_P[Y \mid A = 1, W] - \mathbb{E}_P[Y \mid A = 0, W]$. The non-pathwise differentiable parameter gives the probability a subject selected at random will have treatment effect beyond the level t. It can be thought of as a ”survival” of the treatment effect function because it is monotonically decreasing. It is also more familiarly, 1 - CDF of the random variable that gives the treatment effect for a subject drawn at random. The user can select the kernel according to its support and its order.

3.5 An Easy Implementation of CV-TMLE

Introduction

The original formulation and theoretical results of cross-validated targeted maximum likelihood estimators, CV-TMLE (Zheng and van der Laan 2010), leads to an algorithm for the CV-TMLE that generally requires 10 targeting steps for each of 10 validation folds for each iteration in an iterative targeted maximum likelihood estimators or TMLE (van der Laan and Daniel Rubin 2006). Such can be done in one regression, which solves the efficient influence curve equation averaged over the validation folds. However, in this pooled regression, we must keep track of the means used in each fold, making the process different than a regular TMLE, once the initial predictions have been formed. The formulation of the CV-TMLE here-in leads to a simpler implementation of the targeting step in that the targeting step can be applied identically as for a regular TMLE once the initial estimates for each validation fold have been computed. The CV-TMLE as discussed here is currently implemented in the R software package of tlverse (Coyle, Malenica, et al. 2018a).

3.5.1 CV-TMLE Definition for General Estimation Problem

We refer the reader to the following sources (van der Laan 2016; van der Laan and Gruber 2016; van der Laan and Daniel Rubin 2006; van der Laan and Rose 2011) for a more detailed look at the theory of TMLE and Zheng and van der Laan, 2010 for theory regarding CV-TMLE. We consider iid data of the form $O \sim P \in \mathcal{M}$, nonparametric or semiparametric model and parameter mapping

$$\Psi(Q(\cdot)) : \mathcal{M} \longrightarrow \mathbb{R}^d$$

Where $Q(P)$ is a model upon which the parameter depends. If we consider $O = (W, A, Y)$ with outcome, Y, and treatment and covariates, A and W, then the outcome model $\bar{Q}(A, W) = E_P[Y \mid A, W]$ and distribution of W, Q_W, would define $Q(P)$. We consider the canonical least favorable submodel (reference myself) of model estimate $\hat{Q}(P_n)$ defined with one-dimensional ϵ:

$$\frac{d}{d\epsilon} L\left(\hat{Q}(P_n)(\epsilon)\right)\Bigg|_{\epsilon=0} = \| D^*\left(\hat{Q}(P_n), \hat{g}(P_n)\right) \|_2$$

This definition coincides with the least favorable submodel if the $d = 1$ because in that case we will have

$$\langle \frac{d}{d\epsilon} L\left(\hat{Q}(P_n)(\epsilon)\right)\Bigg|_{\epsilon=0} \rangle \supset \langle D^*\left(\hat{Q}(P_n), \hat{g}(P_n)\right)\rangle$$

where the above $\| \cdot \|_2$ is the euclidean norm. We then define a mapping $B_n \in 0, 1^n$ to be a random split of $1, .., n$. The training set is defined as $\mathcal{T} = \{i : B_n(i) = 0\}$ and the validation set, $\mathcal{V} = \{i : B_n(i) = 1\}$. As in Zheng 2010, P^0_{n,B_n} and P^1_{n,B_n} and the empirical distributions over $\mathcal{T}$ and $\mathcal{V}$ respectively.

The CV-TMLE estimator as in Zheng and van der Laan, 2010 is defined as

$$\Psi^{k_n}(P_n) = E_{B_n} \Psi\left(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon}_n^{\,k_n})\right)$$

where $\Psi\left(\hat{Q}(P^0_{n,B_n})(\vec{\epsilon}_n^{\,k_n})\right)$ is the plug-in estimator (usually an average of the plugged-in model over the validation set). $\vec{\epsilon}_n^{\,k_n}$ denotes the kth iteration of fluctuation parameters, where k could always be 1 if we use the one-step TMLE (van der Laan and Gruber 2016).

3.5.2 Illustrative Example, VTE

We will now go through the CV-TMLE algorithm for the VTE, variance of treatment effect. Here, we notice that we never target the distribution of W, but rather use the unbiased estimator, the empirical distribution. This is discussed in Zheng and van der Laan, 2010 so refer the reader there for more detail as to why this is often the case. In short, the component of the efficient influence curve in the tangent space of mean 0 functions of W (van der Vaart 2000) is given by $D^*_W(P) = (b(P)(W) - E_P b(P)(W))^2$ where $b(P)(W) = E_P[Y \mid A = 1, W] - E_P[Y \mid A = 0, W]$. For any approximation to this function, its empirical mean will automatically be zero. We denote the following to avoid heavy notation:

$$\bar{Q}^k_{B_n} = \hat{Q}(P^0_{n,B_n})(\overrightarrow{\epsilon_n}^k)$$

is the approximation of the outcome model at the kth iteration. This fit is entirely dependent on the training set P^0_{n,B_n} observations and the fluctuations to the model, performed on the corresponding validation set.

$$\bar{Q}^k_{1,B_n} = \hat{Q}(P_n)(\overrightarrow{\epsilon_n}^k)$$

is the approximation of the outcome model at the kth iteration. We will see it actually depends on P^0_{n,B_n} and $P^1_{n,B_n}\hat{B}(P^0_{n,B_n})(\overrightarrow{\epsilon_n}^k)$, and hence the entire empirical draw of the data.

$$\begin{aligned}
\hat{b}^k_{B_n}(W) &= \bar{Q}^k_{B_n}(1, W) - \bar{Q}^k_{B_n}(0, W) \\
\hat{b}^k_{1,B_n}(W) &= \bar{Q}^k_{1,B_n}(1, W) - \bar{Q}^k_{1,B_n}(0, W) \\
\hat{g}_{B_n}(A \mid W) &= \hat{g}(P^0_{n,B_n})(A \mid W)
\end{aligned}$$

- STEP 1: Initial estimates

  For each split, B_n as in standard 10-fold cross-validation, we use an ensemble learning package such as sl3 (Coyle, Malenica, et al. 2018a) or SuperLearner (Polley et al. 2017) to fit a model on the training set, denoting the model as P^0_{n,B_n}. In this case we will fit relevant factors of the likelihood, such as the propensity score and outcome model, but not the distribution of covariates, W. For those, we use the empirical distribution as an unbiased estimator and will not target it. The initial fit of the $E_P[Y \mid A, W]$, for split, B_n is denoted $\bar{Q}^0_{B_n}$ and the initial fit of the $E_P[A \mid W]$, for split, B_n is denoted g_{B_n}. For both procedures the initial fits are all the same.

- STEP 2: Check Tolerance

  For each fold evaluate the so-called clever covariate:
  $H^k_{B_n}(A, W) = 2(\hat{b}^k_{B_n}(W) - P^1_{n,B_n}\hat{b}^k_{B_n})\frac{2A-1}{\hat{g}_{B_n}(A|W)}$
  and the influence curve approximation

  $$D^*_{k,B_n}(O) = H^k_{B_n}(A, W)(Y - \bar{Q}^k_{B_n}(A, W))$$

  Our proposed procedure would do
  $H^k_{1,B_n}(A, W) = 2(\hat{b}^k_{1,B_n}(W) - E_{B_n}P^1_{n,B_n}\hat{b}^k_{1,B_n})\frac{2A-1}{\hat{g}_{B_n}(A|W)}$
  and the alternate influence curve approximation

$$D_{k,B_n}(O) = H^k_{1,B_n}(A,W)(Y - \bar{Q}^k_{1,B_n}(A,W))$$

Thus, in our procedure we need not keep track of the folds since the average within the clever covariate is merely taken over the entire sample. Thus the process is identical to a TMLE once the initial estimates are made. We just stack them on top of each other and act is if it is all one initial fit as with the regular TMLE.

We then compute the influence curve approximation for each fold and take the sample mean. Since the T_W component, as stated above always has empirical average 0, we only need to take the mean of the component of the influence curve approximation in the tangent space, $T_Y =$ mean 0 functions of $Y \mid A, W$, which have finite variance (van der Vaart 2000). We then check if the mean of the influence curve is below the tolerance level, $\hat{\sigma}/n$ where $\hat{\sigma}$ is the sample standard deviation of the above influence curve computations. This assures we stop the process when the bias is second order as any more fluctuations beyond that point are not helpful. If we are below the tolerance we go to step 4. Otherwise we continue onward.

- STEP 3: Targeting Step: Run a pooled logistic regression over all the folds with model:

$$Y = expit(logit\left(\bar{Q}^k_{B_n}(A,W) + \epsilon^k_n H(\bar{Q}^k_{B_n}(A \mid W)\right)$$

  That is, a model which suppresses the intercept and uses and the initial predictions as the offset. This is identical to our method, except we would use the slightly different clever covariate as stated above.

  Update all the predictions to form $\bar{Q}^{k+1}_{B_n}(A,W)$ or, as with our method $\bar{Q}^{k+1}_{1,B_n}(A,W)$.

- STEP 4: Compute the estimate and CI:

$$\Psi^{k_n}(P_n) = E_{B_n}\Psi\left(\hat{Q}(P^0_{n,B_n})(\overrightarrow{\epsilon_n}^{k_n})\right)$$

  and estimate the standard error via the standard deviation of the influence curve in step 3 divided by root n, which we will just call $\hat{\sigma}/\sqrt{n}$ and form the confidence bands

$$\Psi^{k_n}(P_n) \pm z_\alpha \hat{\sigma}/\sqrt{n}$$

  where z_α is the $1 - \alpha/2$ normal quantile. This entails computing the parameter separately per validation set before averaging the 10 estimates, i.e., compute the sample variance over the validation set for $\hat{b}^k_{B_n}$, getting 10 estimates and then average them. In our procedure we just have a list of n values of $\hat{b}^k_{1,B_n}$ and compute the sample variance over the entire sample.

Thus we can see our procedure simplifies the targeting and, like the original formulation, solves the efficient influence curve equation, i.e. $E_{B_n}P^1_{n,B_n}\hat{b}^k_{B_n}D^*_k(O)$ and $E_{B_n}P^1_{n,B_n}\hat{b}^k_{1,B_n}D_k(O) \approx 0$, leading to the second order expansion as given in section 2.2.

3.5.3 Donsker Condition

In the original formulation of the CV-TMLE, we view the estimator as 10 plug-in estimators. To compute each of the 10 estimators, the targeting step is performed on the validation set. Since we can therefore condition on the training set from which the initial estimate is formed, we essentially have a fixed functions $\bar{Q}^0_{B_n}$ and $\hat{g}_{B_n}$, which we are fluctuating on the validation set with a one-dimensional parametric submodel. Thus the entropy is very low for the class of functions containing $\bar{Q}^k_{B_n}$ in our above algorithm. With our procedure the entropy is a little bigger in that the function, $\bar{Q}^k_{1,B_n}$, can be viewed as fixed, yet depending on an average over all validation sets (therefore very slightly inbred before the targeting step) as well as the fluctuation parameter, ϵ, determined by the validation set. The influence curve approximation, D_{k,B_n}, defined above, will thus have similarly low entropy as if we allowed another parameter in the parametric submodel.

Consider the following, which we pull out of Zheng and van der Laan, 2010, for the convenience of the reader.

Definition 3.5.1. For a class of function, $\mathcal{F}$, whose elements are functions, f, that map observed data, O, to a real number, we define the entropy integral:

$$Entro(\mathcal{F}) = \int_0^\infty \sqrt{\log \sup_Q N\left(\epsilon, \|F\|_{Q,2}, \mathcal{F}, L^2(Q)\right) d\epsilon}$$

where $N\left(\epsilon, \mathcal{F}, L^2(Q)\right)$ is the covering number for $\mathcal{F}$, defined by the minimum number of balls of radius ϵ under the $L^2(Q)$ norm to cover $\mathcal{F}$. F is defined as the envelope of $\mathcal{F}$ or a function such that $|f| \leq F$ for all $f \in \mathcal{F}$.

Consider the following lemma (lemma 2.14.1 in ref van der Vaart and Wellner, 1996) (van der Vaart and Wellner 1996)

Lemma 3.5.1. *Let $\mathcal{F}$ denote a class of measurable functions of O. Let $G_n = \sqrt{n}(P_n - P_0)$. Then*

$$E(sup_{f\in\mathcal{F}} G_n f) \leq Entro(\mathcal{F})\sqrt{P_0 F^2}$$

This lemma then yields the following results in Zheng and van der Laan, 2010. Consider $\overset{\rightarrow k_n}{\epsilon_n}$, a sequence of $\epsilon^1_n, ..., \epsilon^{k_0}_n$ that are the fluctuation parameters dependent on the draw from the data. In the lemma below we assume the k_0 steps of a parametric fluctuation parameters converge in probability to a sequence of length k_0, a very weak assumption, the same as the estimated parameters of a parametric model converging to the truth in probability. NOTE: for the one-step TMLE (van der Laan and Gruber 2016) $k_n = k_0 = 1$ so the notation simplifies a bit.

Lemma 3.5.2. *Suppose $\|\overset{\rightarrow k_n}{\epsilon} - \overset{\rightarrow k_0}{\epsilon}\| \overset{P}{\rightarrow} 0$. For each sample split of B_n, we consider a class of measurable functions of O:*

$$\mathcal{F}\left(P^0_{n,B_n}\right) = \left\{ f_{\vec{\epsilon}}\left(P^0_{n,B_n}\right) = f\left(\vec{\epsilon}, P^0_{n,B_n}\right) - f\left(\vec{\epsilon}_0 P^0_{n,B_n}\right) : \vec{\epsilon} \right\}$$

where the index set contains ϵ_n with probability tending to 1. For a deterministic sequence $\delta_n \to 0$, define subclasses

$$\mathcal{F}_{\delta_n}\left(P^0_{n,B_n}\right) = \left\{f_{\vec{\epsilon}} \in \mathcal{F}\left(P^0_{n,B_n}\right) : \|\vec{\epsilon} - \vec{\epsilon_0}\| < \delta_n\right\}$$

If for deterministic sequence $\delta_n \to 0$ we have

$$E\left\{Entro(\mathcal{F}_{\delta_n}\left(P^0_{n,B_n}\right))\sqrt{P_0 F(\delta_n, P^0_{n,B_n})^2}\right\} \to 0 \text{ as } n \to 0$$

where $F(\delta_n, P^0_{n,B_n})$ is the envelope of $\mathcal{F}_{\delta_n}\left(P^0_{n,B_n}\right)$, then

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left\{f(\vec{\epsilon_n}, P^0_{n,B_n}) - f(\vec{\epsilon_0}, P_0)\right\} = o_P(1)$$

We note to the reader that we keep lemma 3.2 identical to what was in Zheng and van der Laan, 2010, except we do not condition solely on P^0_{n,B_n} when defining $\mathcal{F}\left(P^0_{n,B_n}\right)$. Such does not at all affect the truth of the lemma.

Remainder Term

Our estimate minus the truth is, using notation in Zheng and van der Laan, 2010, where $\vec{k_n}$, indicates the k_n iteration, we have

$$\Psi^{k_n}(P_n) = E_{B_n}\hat{\Psi}_{B_n}(P_n)$$

The second order remainder, $R_2(\cdot)$, can be written:

$$\begin{aligned}\Psi^{k_n}(P_n) - \Psi(P_0) &= E_{B_n}(P^1_{n,B_n} - P_0)D_{\vec{k_n},B_n} + R_2(P_n, P_0)\\ &= -E_{B_n}P_0 D_{\vec{k_n},B_n} + R_2(P_n, P_0)\end{aligned}$$

Assuming the remainder is $o_P(1/\sqrt{n})$, we then get that

$$\Psi^{k_n}(P_n) - \Psi(P_0) = E_{B_n}(P^1_{n,B_n} - P_0)D_{\vec{k_n},B_n} + o_P(1\sqrt{n}) = -E_{B_n}P_0 D_{\vec{k_n},B_n} + R_2(P_n, P_0)$$

since our procedure solves $E_{B_n}P^1_{n,B_n}D_{\vec{k_n},B_n} = 0$. As discussed, we can quite easily satisfy lemma 3.2 for the function class containing $D_{\vec{k_n},B_n}$. Again, assuming the remainder is $o_P(1/\sqrt{n})$ our estimator is asymptotically efficient if $D_{\vec{k_n},B_n}$ converges to the true influence curve in $L^2(P_0)$ (van der Laan and Daniel Rubin 2006). For TE variance the remainder term conditions are no more strict than for the original formulation of the CV-TMLE.

3.5.4 Conclusion

This slight adjustment to the CV-TMLE algorithm is easier to implement and retains the same theoretical properties, as shown in our example here. It remains to be more formally generalized to include a class of TMLE's for which it is valid but the example used here-in gives the reader sufficient intuition to understand when such can be done. For one, it is obvious if any polynomial factor of a mean (assuming the mean converges) appears as a factor in the clever covariate, then the entropy will be similarly small, so this procedure covers many examples one might find in practice. The procedure overlaps exactly with the originally formulated CV-TMLE with many common parameters where the clever covariates contain no empirical means. It is a subject for future research whether this procedure has any advantages in finite samples, such as in the case of simultaneously estimating the ATE, which is then used as the centering in the VTE computation. Such appears to be perhaps more sensible but simulations have shown no appreciable difference in performance for VTE.

Bibliography

Annane, Djillali et al. (2002). "Effect of Treatment With Low Doses of Hydrocortisone and Fludrocortisone on Mortality in Patients With Septic Shock." In: *JAMA* 288.7, pp. 862–871.

Athey, Susan and Guido Imbens (2015). "Machine Learning Methods for Estimating Heterogeneous Causal Effects". In: *arxiv.org*.

— (2016). "Recursive Partitioning for Heterogeneous Effects". In: *Proceedings of the National Academy of Sciences of the USA* 113(27), pp. 7353–7360.

Bareinboim, Elias and Judea Pearl (n.d.). "A general algorithm for deciding transportability of experimental results". In: *Journal of causal Inference* 1.1, pp. 107–134.

Benkeser, David, Marco Carone, et al. (2017). "Doubly robust nonparametric inference on the average treatment effect". In: *Biometrika* 104.4, pp. 63–880.

Benkeser, David, Chris Kennedy, and Oleg Sofrygin (2016). *halplus*. URL: `https://github.com/benkeser/halplus`.

Benkeser, David and Mark van der Laan (2016). "The Highly Adaptive Lasso Estimator". In: *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics*, pp. 689–696.

Bitler, Marianne, Jonah B. Gelbach, and Hilary Williamson Hoynes (2014). "Can Variation is Subgroup's Average Treatment Effects Explain Treatment Effect Heterogeneity?" In: *NBER Working Paper* w20142. URL: `https://ssrn.com/abstract=2438563`.

Cambanis, Stamatis, Gordon Simons, and William Stout (1976). "Inequalities for E(k(X, Y) When the Marginals Are Fixed". In: *Zeitschrift Fur Wahrscheinlichskeitstheorie* 36, pp. 285–294.

Chambaz, Antoine A, P Neuvial, and van der Laan Mark J (2012). "Estimation of a nonparametric variable importance measure of a continuous exposure". In: *Electron J Statistics*, pp. 1059–1099.

Chen, Tianqi et al. (2017). *xgboost: Extreme Gradient Boosting*. R package version 0.6-4. URL: `https://CRAN.R-project.org/package=xgboost`.

Cole, Stephen R and Elizabeth A Stuart (2010). "Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial". In: *American journal of epidemiology* 172.1, pp. 107–115.

Cox, D. R. (1958). *Planning of Experiments*. 5th. John Wiley and Sons, Inc.

Coyle, Jeremy and Jonathan Levy (2018). *gentmle2*. URL: `https://github.com/jeremyrcoyle/gentmle2`.

Coyle, Jeremy, Ivana Malenica, et al. (2018a). *sl3*. URL: `https://github.com/tlverse/sl3`.

— (2018b). *tlverse*. URL: `https://github.com/tlverse`.

Coyle, Jeremy, Ivana Malenica, et al. (2018c). *tmle3*. URL: `https://github.com/tlverse/tmle3`.

Diaz Muñoz, Ivan and Mark van der Laan (2012). "Population Intervention Causal Effects Based on Stochastic Interventions". In: *Biometrics* 68(2).541-549.

Dunn, Olive Jean (1961). "Multiple Comparisons Among Means". In: *Journal of the American Statistical Association* 56.293, pp. 52–64.

Fisher, Ronald A. (1951). *The Design of Experiments*. sixth. Hafner Publishing Company.

Folland, Gerald B. (1999). *Real Analysis, Modern Techniques and There Applications*. 2nd. Wiley.

Frangakis, Constantine (2009). "The calibration of treatment effects from clinical trials to target populations". In: *Clinical trials (London, England)* 6.2, p. 136.

Frechet, Maurice (1951). "Sur Les Tableaux de Correlation Dont Les Marges Sont Donnees". In: *Annals University Lyon* A.14, pp. 53–77.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33(1), pp. 1–22. URL: `http://www.jstatsoft.org/v33/i01/`.

Greenland, Sander and James Robins (1986). "Identifiability, Exchangeability, and Epidemiological Confounding". In: *International Journal of Epidemiology* 15.3.

Gruber, Susan and Mark van der Laan (2010). "A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome". In: *The International Journal of Biostatistics* 6. URL: `http://www.bepress.com/ijb/vol6/iss1/26`.

Hastie, Trevor (2017). *gam: Generalized Additive Models*. R package version 1.14-4. URL: `https://CRAN.R-project.org/package=gam`.

Heckman, James J. and Jeffrey Smith (1997). "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts". In: *Review of Economic Studies* 64.4, pp. 487–535.

Heckman, James J. and Jeffrey A. Smith (1998). "Evaluating the Welfare State". In: *National Bureau of Economic Research* 6542.

Kale, B.K. (1985). ""A note on the super efficient estimator". 12: 259–263". In: *Journal of Statistical Planning and Inference* 12, pp. 259–263.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007). "Experimental analysis of neighborhood effects". In: *Econometrica* 75.1, pp. 83–119.

van der Laan, Mark (2016). "A Generally Efficient Targeted Minimum Loss Based Estimator". In: *U.C. Berkeley Division of Biostatistics Working Paper Series* 343. URL: `http://biostats.bepress.com/ucbbiostat/paper343`.

— (2017). "Finite Sample Inference for Targeted Learning". In: *ArXiv e-prints*. arXiv: `1708.09502 [math.ST]`.

van der Laan, Mark and Susan Gruber (2016). "One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels". In: *The International Journal of Biostatistics* 12(1), pp. 351–378.

van der Laan, Mark, Eric C. Polley, and Alan Hubbard (2007). "Super Learner". In: *U.C. Berkeley Division of Biostatistics Working Paper Series* 6(1).222.

van der Laan, Mark and Sherri Rose (2011). *Targeted Learning*. New York: Springer.

— (2018). *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies (2018)*. Springer International Publishing AG.

van der Laan, Mark and Daniel Rubin (2006). “Targeted Maximum Likelihood Learning”. In: *U.C. Berkeley Division of Biostatistics Working Paper Series* 213. URL: `http://biostats.bepress.com/ucbbiostat/paper213`.

LeDell, Erin (2017). *h2oEnsemble: H2O Ensemble Learning.* R package version 0.2.1. URL: `https://github.com/h2oai/h2o-3/tree/master/h2o-r/ensemble/h2oEnsemble-package`.

Levy, Jonathan (2018a). “An Easy Implementation of CV-TMLE”. In: *arXiv:1811.04573 [stat.ME].* URL: `arxiv.org/abs/1811.04573`.

— (2018b). *blip CDF.* URL: `https://github.com/jlstiles/blipCDF`.

— (2018c). “Canonical Least Favorable Submodels: A New TMLE Procedure for Multidimensional Parameters”. In: *arXiv:1811.01261 [stat.ME].* URL: `https://arxiv.org/abs/1811.01261`.

— (2018d). *TECDFsim.* URL: `https://github.com/jlstiles/TECDFsim`.

Levy, Jonathan et al. (2018). “A Fundamental Measure of Treatment Effect Heterogeneity”. In: *arXiv:1811.03745 [stat.ME].* URL: `https://arxiv.org/abs/1811.03745`.

Luedtke, Alex and Mark van der Laan (2016). “Super-Learning of an Optimal Dynamic Treatment Rule”. In: *International Journal of Biostatistics* 12(1).305-332.

Luedtke, Alexander R, Marco Carone, and Mark J van der Laan (2015). “An Omnibus Nonparametric Test of Equality in Distribution for Unknown Functions”. In: *arXiv preprint arXiv:1510.04195.*

McCullagh, P. (1983). “Quasi-likelihood functions. Annals of Statistics”. In: 11.1, pp. 59–67.

Miettinen, Olli S (1972). “Standardization of risk ratios”. In: *American Journal of Epidemiology* 96.6, pp. 383–388.

Milborrow, Stephen (2017). *earth: Multivariate Adaptive Regression Splines.* Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller’s Fortran utilities with Thomas Lumley’s leaps wrapper. R package version 4.5.0. URL: `https://CRAN.R-project.org/package=earth`.

Neyman, Jerzy (1923). “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9”. In: *Statistical Sciences* 5.4. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original which appeared in Roczniki Nauk Rolniczych Tom X, 1923, pp. 465–480.

Pearl, Judea (1995). “Causal diagrams for empirical research Biometrika”. In: *Biometrika* 82.4, pp. 669–709.

— (2000). *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press, p. 484.

Pearl, Judea and Elias Bareinboim (2014). “External validity: From do-calculus to transportability across populations”. In: *Statistical Science*, pp. 579–595.

Petersen, Maya et al. (2012). “Diagnosing and Responding to Violations in the Positivity Assumption”. In: *Statistical Methods in Medical Research* 21.1, pp. 31–54.

Polley, Eric C. et al. (2017). *SuperLearner: Super Learner Prediction.* R package version 2.0-23-9000. URL: `https://github.com/ecpolley/SuperLearner`.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. URL: `https://www.R-project.org/`.

Riesz, Frgyes (1909). “Sur les opérations fonctionnelles linéaires”. In: *C.R. Academy of Sciences Paris* 149, pp. 974–977.

Robins, James (1986). “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period”. In: *Journal of Mathematical Modeling* 7, pp. 1393–512.

Rubin, Donald (1974). “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”. In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Rudolph, Kara E and Mark van der Laan (2017). “Robust estimation of encouragement-design intervention effects transported across sites”. In: *Journal of the Royal Statistical Society Series B Statistical Methodology* 79.5, pp. 1509–1525.

Rudolph, Kara E, Nicole M Schmidt, et al. (2018). “Composition or context: using transportability to understand drivers of site differences in a large-scale housing experiment”. In: *Epidemiology* 29.2, pp. 199–206.

Rudolph, Kara E, Oleg Sofrygin, Nicole M Schmidt, et al. (2017). “Mediation of neighborhood effects on adolescent substance use by the school and peer environments in a large-scale housing voucher experiment”. In: *Epidemiology*, In Press.

Rudolph, Kara E, Oleg Sofrygin, Wenjing Zheng, et al. (2017). “Robust and flexible estimation of data-dependent stochastic mediation effects: a proposed method and example in a randomized trial setting”. In: *Epidemiologic Methods*, In Press.

Strotz, RH and HO Wold (1960). “Recursive vs. nonrecursive systems: an attempt at synthesis (part I of a triptych on causal chain systems)”. In: *Econometrica* 28.2, pp. 417–427.

Stuart, Elizabeth A et al. (2011). “The use of propensity scores to assess the generalizability of results from randomized trials”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174.2, pp. 369–386.

Therneau, Terry, Beth Atkinson, and Brian Ripley (2017). *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-11. URL: `https://CRAN.R-project.org/package=rpart`.

Tian, Lu et al. (2014). “A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates”. In: *Journal of the American Statistical Association* 109(508), pp. 1517–1532.

van der Vaart, Aad and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

van der Vaart, Aad (2000). *Asymptotic Statistics.* Vol. Chapter 25. Cambridge, UK: Cambridge University Press.

VanderWeele, Tyler J and Eric J Tchetgen Tchetgen (2017). “Mediation analysis with time varying exposures and mediators”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3, pp. 917–938.

Venbles, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S.* Fourth. ISBN 0-387-95457-0. New York: Springer. URL: `http://www.stats.ox.ac.uk/pub/MASS4`.

Wedderburn, R.W.M. (1974). “Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method”. In: *Biometrika* 61.

Wolpert, David (1992). “Stacked Generalization”. In: *Neural Networks* 5.2, pp. 241–259.

Wright, Marvin N. and Andreas Ziegler (2017). *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.*

Wright, Sewall (1921). “Correlation and Causation”. In: *Journal of Agricultural Research* 20.7, pp. 557–585.

Zheng, Wenjing and Mark van der Laan (2017). “Longitudinal Mediation Analysis with Time-varying Mediators and Exposures, with Application to Survival Outcomes”. In: *Journal of Causal Inference*.

Zheng, Wenjing and Mark van der Laan (2010). “Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation”. In: *U.C. Berkeley Division of Biostatistics Working Paper Series* 273.