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

Year 2011 *Paper* 287

Targeted Minimum Loss Based Estimation Based on Directly Solving the Efficient Influence Curve Equation

Paul Chaffee* Mark J. van der Laan†

*University of California, Berkeley, chafe66@gmail.com

†University of California - Berkeley, laan@berkeley.edu

Targeted Minimum Loss Based Estimation Based on Directly Solving the Efficient Influence Curve Equation

Paul Chaffee and Mark J. van der Laan

Abstract

Applying targeted maximum likelihood estimation to longitudinal data can be computationally intensive. As the number of time points and/or number of intermediate factors grows, the computation resources consumed by these algorithms likewise increases. Different TMLE algorithms have different computational speeds and implementation challenges; there may also be efficiency differences of the corresponding estimators. The algorithm we describe here proceeds by solving the empirical efficient influence curve equation directly using numerical computation methods, rather than indirectly (by solving a score equation), which is the usual route. We believe that this estimator is the simplest of the TMLE procedures to implement in the longitudinal data structure simulated here, which mimics a sequential randomized controlled trial with dynamic treatment rules. Our choice of numerical methods is the well-known secant method for finding the root of a function. The resulting estimation algorithm has computational speed approximately equal to one of the two existing TMLE algorithms for the data generating distributions considered here.

1 Introduction

1.1 Background

Experimental Setting

Sequentially Randomized Controlled Trials (SRCTs) are rapidly becoming essential tools in the search for optimized treatment regimes in ongoing treatment settings. Analyzing data for multiple time-point treatments with a view toward optimal treatment regimes is of interest in many types of afflictions.

A common setting for ongoing treatment therapy involves randomization to initial treatment (or randomization to initial treatment within subgroups of the population of interest), followed by later treatments which may also be randomized, or randomized to a certain subset of possible treatments given that certain intermediate outcomes occurred after the initial treatment. We give a very brief list of recent method papers for SRCTs in Chaffee and van der Laan (2011).

In particular we concern ourselves with parameters of the observed data that are indexed by dynamic treatment rules or dynamic treatment regimes. A generic example of such a rule is to randomize subjects of a study to an initial pair of treatments (A or B, say), and if a subject responds poorly to the initial treatment, then he or she is again randomized to A or B at the second treatment point. On the other hand, if the subject does well on the first treatment (as determined by some intermediate biomarker), then he or she is assigned the same treatment at the second time point as the first. If the intermediate biomarker in such SRCTs is affected by initial treatment, and in turn affects decisions at the second time-point treatment as well as the final outcome, then it is a so-called time-dependent confounder. Many of the methods developed for multiple time-point treatments are designed to remove bias due to this type of confounder.

Context of the New Estimator

In Chaffee and van der Laan (2011) we describe the implementation of two distinct targeted maximum likelihood estimation (TMLE) algorithms (corresponding to "one-step" and "iterative" procedures) for estimating specified counterfactual parameters of the underlying distribution corresponding to a

particular longitudinal data structure, indexed by dynamic treatment rules. We also compared their performance to that of some well-known existing estimators. The comparative advantages amongst TMLE's involve differences in computational resources needed, and in complexity of implementation. In this article we present a third algorithm, which we find conceptually easier to implement than either of the foregoing methods, and whose speed is comparable to that of the iterative method.

We emphasize that there are targeted maximum likelihood estimators (plural) of a given parameter, since TMLE is a class of estimation methods that utilizes *i)* a fluctuation submodel of an initial estimator and *ii)* a loss function or other empirical criterion for fitting the submodel.

The TMLEs presented in Chaffee and van der Laan (2011), as well as all other heretofore implemented TMLEs independent of data type, all solve a score equation as the means of constructing the estimator. Here we present a TMLE based on a different empirical criterion, namely, solving the empirical efficient influence curve equation directly. We have been moving toward the term "targeted minimum loss-based" rather than "targeted maximum likelihood" in describing this class of estimators, and the procedure we describe here motivates this terminological adjustment since maximizing the likelihood is not involved in the construction of the estimator.

The procedure bears a superficial similarity to that of estimating equation methodology, though our procedure solves the efficient influence curve (EIC) equation by adjusting the amount of fluctuation of a predesignated fluctuation submodel, and does not solve it in the parameter, ψ. Like all TMLE's (and unlike estimating equation-based estimators), this TMLE is a substitution estimator, and retains the associated benefits. Results from our simulations indicate that the procedure also exhibits all of the finite sample advantages of the existing TMLE procedures, which have been described in a variety of applications (van der Laan et al., 2009), and which were also seen in Chaffee and van der Laan (2011).

2 Data Structure and Likelihood

As in our earlier paper, we consider the longitudinal data structure $O = (L(0), A(0), L(1), A(1), Y = L(2)) \sim P_0$, where $L(0)$ is a vector of baseline

covariates, $A(0)$ is initial randomized treatment, $L(1)$ is a single intermediate outcome or other time-varying covariate, $A(1)$ is the second time point treatment, $Y = L(2)$ is the outcome of interest and P_0 is the joint distribution of O. The likelihood of the data described above can be factorized as

$$p(O) = \prod_{j=0}^{2} P[L(j) \mid \bar{L}(j-1), \bar{A}(j-1)] \prod_{j=0}^{1} P[A(j) \mid \bar{L}(j), \bar{A}(j-1)], \quad (1)$$

where $\bar{A}(j) = (A(0), A(1), ..., A(j))$ and $\bar{L}(j)$ is similarly defined. This particular factorization is the natural one given the time-ordering of the factors specified by the data structure.

For simplicity, we introduce the notation $Q_{L(j)}$, $j = 0, 1, 2$ to denote the factors of (1) under the first product and $g_{A(j)}$, $j = 0, 1$ for those under the second, which is the treatment mechanism. In the simpler notation,

$$p = \prod_{j=0}^{2} Q_{L(j)} \prod_{j=0}^{1} g_{A(j)} = Qg.$$

We are interested in a treatment-specific mean for the multiple time point data structure, where here a particular treatment means a specific treatment course over time. Instead of a static treatment regime, we define a *treatment rule*, $d = (d_0, d_1)$ for the treatment points $(A(0), A(1))$ where $d_0 : L(0) \rightarrow d_0(L(0))$ and $d_1 : (A(0), \bar{L}(1)) \rightarrow d_1(A(0), \bar{L}(1))$. Since following the rule entails $A(0) = d_0(L(0))$ we can also write $d_1 : \bar{L}(1) \rightarrow d_1(\bar{L}(1))$ and hence the overall rule $d(\bar{L}(1)) = \big(d_0(L(0)), d_1(\bar{L}(1))\big)$. Under this definition we can easily express either static or dynamic treatment rules, or a combination of the two. Several examples of dynamic treatment rules are given in section 4.1.

We next define the G-formula to be the product across all nodes, excluding intervention nodes, of the conditional distribution of each node given its parent nodes, and with the values of the intervention nodes fixed according to the static or dynamic intervention of interest. This formula thus expresses the distribution of $\bar{L}$ under the dynamic intervention $\bar{A} \equiv (A(0), A(1)) = d(\bar{L})$:

$$P^{(d)}(\bar{L}) = \prod_{j=0}^{2} Q^{(d)}_{L(j)}(\bar{L}(j)), \quad (2)$$

where

$$Q^{(d)}_{L(j)}(\bar{L}(j)) \equiv P(L(j) \mid \bar{L}(j-1), \bar{A}(j-1) = d(\bar{L}(j-1))).$$

The superscript (d) here indicates that the conditional distribution of each node given its parent L nodes is also conditional on treatment being set according to the specified treatment rule. We reserve subscript d to refer to counterfactually-defined variables.

Under the so-called sequential randomization assumption (SRA) and positivity assumption, the G-computation formula equals the counterfactual distribution of the data had one carried out the specified intervention described by the causal graph, which graph is assumed to underlie data generation.

In Chaffee and van der Laan (2011), we present the causal model we assume, which is nothing but a set of causal dependencies implied by the time ordering $(L(0), A(0), ..., L(2))$. The model can be viewed as a set of structural equations that codify functional dependencies between the graph nodes, and which therefore allow us to define the counterfactual Y_d. This in turn enables the identification of a corresponding counterfactual parameter of interest, $\Psi^F = EY_d$, which is a well-defined mapping $\mathcal{M}^F \to \mathbb{R}$ under the full data distribution. Using (2),

$$\begin{aligned}\Psi^F &= EY_d \\ &= \sum_{l(0),l(1)} E\left(Y_d \mid L(0) = l(0), L_d(1) = l(1)\right) \prod_{j=0}^{1} Q_{L_d(j)}(\bar{l}(j)), \qquad (3)\end{aligned}$$

where $Q_{L_d(j)} \equiv P(L_d(j) \mid Pa(L_d(j)))$. In words, this parameter is the mean outcome under the causal model where all treatments are set according to the dynamic treatment intervention $\bar{A} = d(\bar{L})$.

For the parameter of interest here, EY_d, the sequential randomization assumption (SRA), $Y_d \perp A(j) \mid Pa(A(j))$ for $j = 0, 1$, is sufficient for identifiability of the causal parameter Ψ^F and the following parameter of the observed data distribution, $\Psi(P_0)$ (Robins, 1986).

$$
\begin{aligned}
\Psi^F &\equiv EY_d \\
&\overset{SRA}{=} \sum_{l(0),l(1)} E\left(Y \mid L(0)=l(0), L(1)=l(1), \bar{A}=d(\bar{L})\right) \times \\
&\qquad P\left(L(1)=l(1) \mid L(0)=l(0), A(0)=d_0\right) \times \\
&\qquad P(L(0)=l(0)) \\
&= \Psi(P_0). \qquad (4)
\end{aligned}
$$

Note that this parameter depends only on the Q part of the likelihood and we therefore also write $\Psi(P_0) = \Psi(Q_0)$. Note also that the first two factors in the summand are undefined if either $P\left(\bar{A} = d(\bar{L}) \mid L(0) = l(0), L(1) = l(1)\right)$ or $P\left(A(0) = d_0 \mid L(0) = l(0)\right)$ are 0 for any $(l(0), l(1))$, and so we require these two conditional probabilities to be positive—the so-called *positivity assumption*.

3 Method

3.1 Existing TML Estimators

In targeted minimum loss-based estimation (TMLE) we begin by obtaining an initial estimator of Q_0; we then update this estimator with a fluctuation function that is tailored specifically to remove bias in estimating the particular parameter of interest. Naturally, this means that the fluctuation function is a function of the parameter of interest. The initial estimator, Q^0 of Q_0 can be obtained in a number of ways, but we advocate a data-adaptive approach in all cases. In any case, the TMLE methods do not require any particular estimation method for Q^0, though there are clear gains if Q^0 is close to Q_0.

Upon obtaining an initial estimate Q^0, the next step in TMLE is to apply a fluctuation function to this initial estimator that is the least favorable parametric submodel through the initial estimate, Q^0, for the parameter Ψ (van der Laan and Rubin, 2006). We signify this fluctuated update $Q_n(\epsilon)$.

Since the Cramer-Rao lower bound corresponds with a standardized L_2 norm of $d\Psi(Q_n(\epsilon))/d\epsilon$ evaluated at $\epsilon = 0$, this is equivalent to selecting the parametric submodel for which this derivative is maximal w.r.t. this L_2 norm.

The above described fluctuated update $Q_n(\epsilon)$ also results in an asymptotically efficient estimator, because the score of our parametric submodel at zero fluctuation equals the EIC of the pathwise derivative of the target parameter, Ψ (also evaluated at $\epsilon = 0$). As we mentioned in the last section the TMLE of $\Psi(Q)$ essentially consists in i) selecting a submodel $Q_g(\epsilon)$ possibly indexed by nuisance parameter g, and ii) a valid loss function $L(Q, O) : (Q, O) \rightarrow L(Q, O) \in \mathbb{R}$. Given these two elements, TMLE solves

$$P_n \left\{ \frac{d}{d(\epsilon)} [L(Q_n^*(\epsilon))]_{\epsilon=0} \right\} = 0, \tag{5}$$

so if this "score" is equal to the EIC, $D^*(Q_n^*, g_n)$, then we have that Q_n^* solves $P_n D^*(Q_n^*, g_n) = 0$. Now a result from semi-parametric theory is that solving this efficient score for the target parameter yields, under regularity conditions (including the requirement that Q_n and g_n consistently estimate Q_0 and g_0, respectively), an asymptotically linear estimator with influence curve equal to $D^*(Q_0, g_0)$ (Bickel et al., 1997). The TMLE of the target parameter is therefore efficient. Moreover, the TMLE is double-robust in that it is a consistent estimator of $\Psi(Q_0)$ if either Q_n or g_n is consistent.

TMLE acquires this property by choosing the fluctuation function, $Q(\epsilon)$, such that it includes a term derived from the efficient influence curve of Ψ.

3.2 Numerical Solution TMLE

General Description

Above we mentioned that existing TMLE's solve (5). The method we present here involves solving instead

$$P_n D^*(Q_n(\epsilon), g) = 0 \tag{6}$$

in ϵ, or to the same effect, selecting ϵ_s such that

$$\epsilon_s = \underset{\epsilon}{argmin} \, |P_n D^*(Q_n(\epsilon), g)|, \tag{7}$$

where g is either the given, known treatment mechanism or an estimate of it, and $\epsilon \in [a, b] \subset \mathbb{R}$, which interval is assumed to contain the solution to (6). The general idea for this method was first suggested in van der Laan and Rubin (2006). $Q_n(\epsilon)$ takes the exact form as for the loss-based TMLE's, i.e., it uses the same parametric submodel through Q^0 (see below). What remains is to choose ϵ. If the empirical EIC is well-behaved on $\epsilon \in [a, b]$ and the solution is contained in that interval, then one should be able to find an ϵ_s such that $P_n D^*(Q_n(\epsilon_s), g_n)$ is arbitrarily close to 0, which means one has effectively found an estimator $Q_n(\epsilon_s)$ of Q_0 that solves (6).

Accordingly, let us define $Q_s^* \equiv Q_n(\epsilon_s)$, where ϵ_s is the solution to (6), or to (7) if a finite number of candidate solutions is considered. $\Psi(Q_s^*)$ is then the corresponding "numerical methods TMLE" of $\Psi(Q_0)$. Since this choice of Q_s^* solves $P_n D^*(Q_n(\epsilon), g) = 0$, it necessarily solves (5) with Q_s^* in place of $Q_n(\epsilon)$. However, since the solution ϵ_s was not arrived at via application of the loss function $L(Q, O)$ assumed in (5), we have no assurance that the likelihood for Q_s^* has increased relative to Q^0, the latter estimator being some initial estimate of Q_0 without fluctuation applied. That is, assuming the negative log likelihood as the loss function, we have no set of conditions that guarantees that $P_n L(Q_s^*, O) \leq P_n L(Q^0, O)$. Nevertheless, Q_s^* represents a movement along the hardest submodel from some initial Q^0, which does indeed result in an estimator $\Psi(Q_s^*)$ that is less biased than $\Psi(Q^0)$, even if in practice Q_s^* does not have a greater likelihood than Q^0, though it would be surprising if it failed to. It is nevertheless encouraging to see that $P_n L(Q_s^*) \leq P_n L(Q^0)$ in practice, where in our simulations Q^0 is the standard MLE, and this was indeed the case without fail in our simulation runs. In fact, the likelihood of the numeric solution estimator was strictly greater than that of Q^0 in all simulations.

Efficient Influence Curve and Parametric Submodel

In order to obtain a numerical solution to (6), one of course needs the explicit form of $D^*(Q, g)$ for the parameter being estimated. The EIC for parameter $\psi = EY_d$ when $L(1)$ and $Y \equiv L(2)$ are binary, and where d is a treatment rule is given in Theorem 1 of van der Laan (2010a), and the EIC for ψ when $L(1)$ is discrete-valued is given in Chaffee and van der Laan (2011).

The empirical estimate of D^* (say D_n^*) for a single observation substitutes the corresponding estimates $Q_{L(j),n}$ and $Q_{L(j),n}^{(d)}$ in place of $Q_{L(j)}$ and $Q_{L_d(j)}$.

It is instructive to represent D_n^* in terms of these estimates of the Q-components of the likelihood and, by implication, in terms of ϵ, which is to be selected.

Though the D^* we present here is for the case of binary $L(1)$, our simulations for this article are for discrete-valued $L(1)$ with four levels. The EIC for the parameters we identified above are more complex in the discrete $L(1)$ case than the binary case, but the method we present here is independent of the types of variables involved. We therefore develop the method for the binary $L(1)$ case to avoid unnecessary conceptual and notational complexity.

We have

$$D_n^*(O) = D_n^*(Q_n, g_n)(O) = \sum_{j=0}^{2} D_{j,n}^*(Q_n, g_n)(O)$$

where

$$D_{0,n}^* = \sum_{l(1)} \left\{ Q_{L(2),n}^{(d)}(y=1, L(0), l(1)) Q_{L(1),n}^{(d)}(L(0), l(1)) \right\} - \psi_n,$$

$$D_{1,n}^* = \frac{I[A(0) = d_0(L(0))]}{g[A(0) = d_0(L(0)) \mid X]} \left\{ Q_{L(2),n}^{(d)}(y=1, l(1)=1, L(0)) - Q_{L(2),n}^{(d)}(y=1, l(1)=0, L(0)) \right\} \times \left\{ L(1) - Q_{L(1),n}(l(1)=1, A(0), L(0)) \right\},$$

$$D_{2,n}^* = \frac{I[\bar{A} = d(\bar{L})]}{g[\bar{A} = d(\bar{L}) \mid X]} \left\{ L(2) - Q_{L(2),n}\left(y=1, \bar{L}(1), \bar{A}(1)\right) \right\}, \tag{8}$$

and

$$\psi_n = \widehat{EY_d} = \frac{1}{n} \sum_{i=1}^{n} \sum_{l(1)} Q_{L(2),n}^{(d)}(y=1, L(0)_i, l(1)) \prod_{j=0}^{1} Q_{L(j),n}^{(d)}(L(0)_i, l(1)).$$

and where X refers to the full data. For TMLE, the EIC gives us the form of the parametric submodel, $Q(\epsilon)$ for the conditional probability of each factor $L(j)$ that is to be estimated:

$$logit(Q_{L(j)}(\epsilon)) = logit(Q^0_{L(j)}) + \epsilon C_{L(j),n}, \tag{9}$$

where $Q^0_{L(j)}$ is some initial estimate of $Q_{L(j)}$ (e.g., the MLE),

$$C_{L(1),n} = \frac{I[A(0) = d_0(L(0))]}{g[A(0) = d_0(L(0)) \mid X]} \left\{ Q^{(d)}_{L(2),n}(y = 1, l(1) = 1, L(0)) - Q^{(d)}_{L(2),n}(y = 1, l(1) = 0, L(0)) \right\},$$

and

$$C_{L(2),n} = \frac{I[\bar{A} = d(\bar{L})]}{g[\bar{A} = d(\bar{L}) \mid X]}.$$

Using now fluctuation submodels $Q_{L(j)}(\epsilon)$ and $Q^{(d)}_{L(j)}(\epsilon)$ given in (9) for the elements $Q_{L(j),n}$ and $Q^{(d)}_{L(j),n}$, respectively, in the formula above, our method attempts to solve (6) or (7) with D^*_n in place of D^*.

3.3 Numerical Methods for Solving Empirical Efficient Influence Curve Equation

Though complex, (8) for our present purposes is nothing but a one dimensional function of ϵ. For notational convenience let us thus write $f(\epsilon) \equiv P_n D^*(Q_n(\epsilon))$. If $f(\epsilon)$ is continuous and has a unique root, then the well-known bisection and secant methods of numerical analysis (see, e.g., Faires and Burden, 2003) are promising techniques for finding the root. If, further, $f(\epsilon)$ is differentiable w.r.t ϵ on the interval over which it is being evaluated, then Newton's method is also a candidate. (Other well-known methods include the *method of false position* and *Müller's method*.)

The purpose of this article is primarily to present the solving of the empirical EIC equation—given a specified fluctuation submodel—via numerical techniques as a method of producing a TMLE. We thus omit technical and detailed comparisons of various numerical techniques for obtaining these solutions. For a suitably well behaved function f, the specific technique employed

to find ϵ_s, though central to the actual implementation of the estimator, is of secondary importance to the overall method described here. There are certainly pros and cons of each technique, which can be assessed a priori if one knows the exact form of $f(\epsilon)$ under all applicable data sets, but one generally does not have such knowledge. The advantages associated with these techniques have to do with whether or not the algorithm is guaranteed to converge, and if it does converge, how quickly. Basic texts on the subject (e.g., Faires and Burden, 2003) give an adequate treatment of these comparisons and we refer the reader there for more detail. To re-iterate: our research interest is in the performance of a TMLE that is produced by solving (6) or (7) in the manner explained in the previous section, and we assume that in all cases of interest there is a numerical technique adequate to the task.

Nevertheless, a brief comparison of the best known techniques for the present context is in order. We have in fact implemented both the bisection and secant methods, and have not attempted Newton's method. The appeal of Newton's method is its rate of convergence (in terms of number of iterations)—under most circumstances if it does converge it has the fastest convergence rate. This is not universally true however. Moreover, the method has the drawback that for each iteration both $f(\epsilon)$ and $f'(\epsilon)$ must be evaluated. For functions that are computationally intensive to evaluate, as in our case, this undercuts the advantage of requiring fewer iterations for a given tolerance compared to the secant method, and could even make Newton's method slower to converge in real time, even if in fewer iterations. The latter fact combined with the added complexity of implementation make the possible gains of Newton's method over the secant method negligible in our case. We thus pursued the secant method as our primary numerical method, having first implemented the bisection method.

Most worthy of mention in comparing these latter two numerical techniques is that 1) the bisection method is guaranteed to converge if the function of interest has a root on the initially specified interval, and the secant method is not (though this is not problematic in our context—see below) and 2) the bisection method is much slower to converge than the secant method in general. In our context the latter factor drives the choice between these two numerical techniques. (Recall that we seek estimators that have computational advantages in the longitudinal setting, which setting is generalizable to any number of time points, and multiple intermediate outcomes per time

point.) Though we have implemented the bisection method, we found that in all cases tested, the secant method converged in far fewer iterations for a given tolerance (usually chosen to be $\sim 10^{-6}$). The difference in the values of the solution ϵ_s produced by the two methods can be made arbitrarily small by performing enough iterations. Since the secant method is superior in every way applicable to our function (except guaranteed convergence) we focus entirely on it as the chosen technique. As mentioned above, lack of guaranteed convergence is not a concern here, which we address in detail in the Discussion section. We therefore give a brief description of how to apply the secant method for our function of interest, $f(\epsilon) \equiv P_n D^*(Q_n(\epsilon))$.

Secant Method

The secant method is based on a sequence of approximations to the root of a function, generated by drawing secant lines through, in our case, the points $(\epsilon_k, f(\epsilon_k))$ and $(\epsilon_{k+1}, f(\epsilon_{k+1}))$, $k = 0, 1, ..., K$. The zero of each such line is computed and this defines the position of the next approximation, ϵ_{k+2}. The initial values (ϵ_0, ϵ_1) need not bracket the solution though the closer they are to it, the more rapidly the algorithm will converge. Starting with initial approximations (ϵ_0, ϵ_1), the first iteration produces a new approximation

$$\epsilon_2 = \epsilon_1 - \frac{(\epsilon_1 - \epsilon_0) f(\epsilon_1)}{f(\epsilon_1) - f(\epsilon_0)}.$$

This result follows from a straight-forward application of point-slope algebra. The next iteration uses (ϵ_1, ϵ_2) as starting values and the process is iterated until $|f(\epsilon_k)| \leq T$ where T is the tolerance deemed sufficient. In our case, the difference in successive estimates $|\psi_k - \psi_{k+1}|$ was typically on the order of $|P_n D^*(\epsilon_k)|$. Thus, an ϵ_k that yields a $|P_n D^*(\epsilon_k)| \leq 10^{-3}$ is quite sufficient, though for our simulations we used $T = 10^{-6}$.

4 Simulations

We simulated data corresponding to the data structure described in section 2 for discrete-valued $L(1)$ under correct and incorrect model specification, and at various sample sizes. Incorrect model simulations were done to illustrate the double-robustness property of the TMLE's. $A(0)$ was assigned randomly

but $A(1)$ was assigned in response to an individual's $L(1)$; the latter corresponding to an individual's intermediate response to treatment $A(0)$. We give the specification of these dynamic regimes in the following section.

For each simulated data set, we computed the estimate of our target parameter $\Psi(P_0) \equiv EY_d$ for the following estimators: 1) Secant TMLE; 2) Iterative TMLE; 3) One-step TMLE; 4) Inverse Probability of Treatment Weighting (IPTW); 5) Efficient Influence Curve Estimating Equation Methodology (EE); 6) Maximum Likelihood Estimation using the G-computation formula. In the *Results* subsection we give bias, variance and relative MSE estimates. A brief description of each of the comparison estimators is given in Chaffee and van der Laan (2011).

4.1 Specific Treatment Rules

As in our earlier work, we considered several treatment rules.

- *Rule 1*. $A(0) = 1$, $A(1) = A(0) * I(L(1) > 1) + (1 - A(0)) * I(L(1) \leq 1)$. In words, set treatment at $A(0)$ to treatment 1, and if the patient does well on that treatment as defined by $L(1) > 1$, continue with same treatment at $A(1)$. Otherwise, switch at $A(1)$ to treatment 0.
- *Rule 2*. $A(0) = 0$, $A(1) = A(0) * I(L(1) > 1) + (1 - A(0)) * I(L(1) \leq 1)$. Identical in principle to Rule 1 except that patients start on treatment 0 instead of treatment 1.
- *Rule 3*. $A(0)$ either 0 or 1, $A(1) = A(0) * I(L(1) > 1) + (1 - A(0)) * I(L(1) \leq 1)$. In words, set treatment at $A(1)$ to be the same as A(0) if the patient is doing well, and switch treatments otherwise.

The parameters we estimate are indexed by these three treatment rules, which we signify by EY_d, $d = 1, 2, 3$.

4.2 Data Generation

Please see our earlier article (Chaffee and van der Laan, 2011) for a full description of the data-generation process for discrete (4-valued) $L(1)$.

4.3 Simulation Results

Estimates of bias, variance and relative mean squared error (Rel MSE) for all three parameters specified above are presented for the TMLE's and several comparison estimators in tables 1 and 2. We define estimated relative MSE for each estimator as the ratio of its estimated MSE to that of an efficient, unbiased estimator. The efficiency bound here is the variance of the efficient influence curve. Thus for each estimator ψ_n of ψ_0,

$$\text{Rel MSE} \equiv \frac{(\hat{E}(\psi_n) - \psi_0)^2 + \widehat{var}(\psi_n)}{var(D^*(Q,g))/n},$$

where D^* is the efficient influence curve for the relevant parameter of the full data distribution, Ψ^F. In fact, the value used in these computations for $var\,(D^*)$ is itself an estimate computed from taking the variance of $D^*(Q_0, g_0)(O)$ from a large number of observations generated from P_0.

The estimates of bias in all cases are not accurate to much less than 10^{-3}; we indicate estimates that appeared to be less than this with an asterisk.

Qm, gc denotes simulations where g (the treatment mechanism) was correctly specified, but $Q^0_{L(2)}$ was purposely misspecified. Qc, gc are simulations for which both Q and g are correctly specified. Note that the IPTW estimator is not affected by the form of Q^0 since this estimator does not estimate Q_0. Differences in IPTW performance between the sets of runs where Q is correctly specified and those where it is misspecified are thus the result of randomness in the simulations.

We generated 4000 independent simulations for each model specification/sample size combination (six sets of simulations in all).

5 Discussion

Please see our earlier article (Chaffee and van der Laan, 2011) for discussion of the results as they pertain to the maximum likelihood-based TMLEs and the rest of the comparison estimators. We mention a few highlights here but focus on results from the secant-based TMLE.

In terms of the performance measures given in the tables, the differences between the three TMLEs implemented are insignificant. The one-step algorithm appears to hold a very slight bias advantage at the small sample size

Qc, gc

n = 30

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.028	-0.028	-0.023	-0.009	-0.041	-0.032
	Var	0.024	0.023	0.024	0.069	0.022	0.022
	Rel MSE	1.4	1.4	1.5	4.1	1.4	1.4
EY_2	Bias	-0.018	-0.018	-0.017	-0.004	-0.024	-0.019
	Var	0.027	0.027	0.028	0.062	0.027	0.027
	Rel MSE	1.4	1.4	1.4	3.0	1.3	1.3
EY_3	Bias	-0.017	-0.017	-0.017	-0.007	-0.024	-0.017
	Var	0.012	0.012	0.012	0.024	0.012	0.012
	Rel MSE	1.2	1.2	1.2	2.5	1.2	1.2

n = 100

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0019	-0.0019	-0.0020	*	-0.0017	-0.0021
	Var	0.0054	0.0054	0.0054	0.0212	0.0048	0.0054
	Rel MSE	1.1	1.1	1.1	4.2	1.0	1.1
EY_2	Bias	*	*	*	0.0016	*	*
	Var	0.0068	0.0068	0.0068	0.0197	0.0064	0.0067
	Rel MSE	1.1	1.1	1.1	3.2	1.0	1.1
EY_3	Bias	*	*	*	*	*	*
	Var	0.0032	0.0032	0.0032	0.0075	0.0031	0.0032
	Rel MSE	1.1	1.1	1.1	2.5	1.0	1.1

n = 200

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	*	*	*	*	*	*
	Var	0.0026	0.0026	0.0026	0.0103	0.0023	0.0026
	Rel MSE	1.0	1.0	1.0	4.1	0.9	1.0
EY_2	Bias	0.0015	0.0015	0.0016	0.0028	0.0010	0.0015
	Var	0.0032	0.0032	0.0032	0.0094	0.0029	0.0031
	Rel MSE	1.0	1.0	1.0	3.1	1.0	1.0
EY_3	Bias	*	*	*	0.0016	*	*
	Var	0.0014	0.0014	0.0014	0.0034	0.0014	0.0014
	Rel MSE	1.0	1.0	1.0	2.3	0.9	1.0

Table 1: *Qc, gc.* Estimator performance for various sample sizes with Q and g correctly specified, for each of three estimated parameters. The estimates for the iterative TMLE were from the 4th iteration. (*) indicates an estimated bias $< 10^{-3}$. (Based on 4000 simulations at each sample size.)

Qm, gc

n = 30

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0049	-0.0050	-0.0042	-0.0071	-0.2975	-0.0093
	Var	0.021	0.020	0.020	0.068	0.071	0.028
	Rel MSE	1.2	1.2	1.2	4.1	9.5	1.7
EY_2	Bias	-0.0010	*	*	0.0022	-0.1677	0.0090
	Var	0.024	0.024	0.024	0.064	0.075	0.029
	Rel MSE	1.2	1.1	1.2	3.1	5.0	1.4
EY_3	Bias	-0.0025	-0.0026	-0.0026	-0.0025	-0.2326	*
	Var	0.011	0.011	0.011	0.025	0.072	0.013
	Rel MSE	1.1	1.1	1.1	2.5	12.8	1.3

n = 100

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0040	-0.0040	-0.0044	0.0014	-0.3159	-0.0034
	Var	0.0056	0.0056	0.0056	0.0203	0.0326	0.0078
	Rel MSE	1.1	1.1	1.1	4.0	26.3	1.6
EY_2	Bias	0.0021	0.0021	0.0026	-0.0024	-0.1855	0.0026
	Var	0.0063	0.0063	0.0064	0.0187	0.0351	0.0080
	Rel MSE	1.0	1.0	1.0	3.0	11.3	1.3
EY_3	Bias	*	*	*	*	-0.251	*
	Var	0.0030	0.0030	0.0030	0.0072	0.0338	0.0036
	Rel MSE	1.0	1.0	1.0	2.4	32.6	1.2

n = 200

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0038	-0.0038	-0.0042	-0.0016	-0.3276	-0.0029
	Var	0.0028	0.0028	0.0028	0.0104	0.0187	0.0039
	Rel MSE	1.1	1.1	1.1	4.1	50.0	1.5
EY_2	Bias	0.0017	0.0017	0.0020	0.0012	-0.1962	0.0019
	Var	0.0033	0.0033	0.0034	0.0095	0.0200	0.0040
	Rel MSE	1.1	1.1	1.1	3.1	19.0	1.3
EY_3	Bias	-0.0010	-0.0010	-0.0011	*	-0.2620	*
	Var	0.0015	0.0015	0.0015	0.0034	0.0193	0.0017
	Rel MSE	1.0	1.0	1.0	2.3	59.4	1.1

Table 2: *Qm, gc.* Estimator performance for various sample sizes with Q misspecified and g correctly specified, for each of three estimated parameters. Estimates for the iterative TMLE were from the 4th iteration. (*) indicates an estimated bias $< 10^{-3}$. (Based on 4000 simulations at each sample size.)

of 30 in estimating EY_1, but the relative MSE's are nearly the same. The overall performance of the secant TMLE could be improved slightly by intervening on simulation runs to ensure the algorithm converges. This makes the differences in bias that we report partly an artifact of the process of bias estimation by simulation, and not a true bias difference, assuming that in actual practice one can examine the empirical EIC for any given data set, which is the case.

In almost every simulation the difference in estimates produced by the iterative and secant approaches was on the order of 10^{-4} or less, even at $n = 30$. The occasions in which the difference was significant were those in which one or the other algorithm failed to converge (in terms of yielding an estimate of Q that solved the empirical EIC) in the allotted number of steps.

The variance of the TMLE, EE and MLE estimators are already very close to the efficiency bound at $n = 100$ under Qc. In our earlier study we found this to be the case for sample sizes of 250 and greater rather than at 100.

The performance of the TMLE's at the small sample size of 30 is remarkable, particularly under model misspecification. Indeed, bias and variance of all three estimators are *better* when Q^0 is misspecified. The bias of the estimating equation estimator is also smaller under model misspecification. The advantage of the TMLEs' being substitution estimators also becomes apparent in these small sample results: at $n = 30$, many times the estimating equation and IPTW estimators gave estimates outside the range $[0, 1]$ even though the outcome is binary.

Misspecification of Q in all cases meant misspecifying $Q^0_{L(2)}$ but correctly specifying $Q^0_{L(1)}$. Thus under Qm, gc the MLE will be biased but the TMLE and EE estimators are double-robust and therefore still asymptotically unbiased under correct specification of g. Under the scenarios simulated here g is expected to be known and we therefore omitted simulations in which g is misspecified; the latter will of course result in bias of the IPTW estimator.

5.1 Convergence of the Secant Algorithm

In practice, we examined several plots of $f(\epsilon)$ vs ϵ to get a rough idea of its shape, and variability of shape, in order to select starting points for the secant algorithm in the simulations. Using these examples we selected fixed initial points for a given set of simulations, and never intervened on particular

runs to ensure convergence of the algorithm. (The algorithm never failed to converge for $n \geq 100$.) The shape of most curves examined was made to order for the secant method, assuming well-chosen starting values (see below). In general practice, one would not need to specify starting points without first examining such a plot, which allows one simply to select starting points that will clearly lead to convergence. In effect, this just means selecting starting points that are "close enough" to the root. Unfortunately there is no generally agreed upon (or even proposed) notion of "close-enough" in the literature, but there are clear cases of it. For example, if the curve is roughly linear near the root, then starting points in the linear region will suffice.

There are also clear cases which can be problematic for finding the root in a reasonable number of iterations using the secant method. We have discovered two such general cases, both of which were observed only at the small sample size of 30. The first is when the curve has a point approaching zero slope between the two starting values (see figure 1). That is, assuming $f(\epsilon)$ is differentiable, then there is an $\epsilon' \in [\epsilon_0, \epsilon_1]$ such that

$$f'(\epsilon)\Big|_{\epsilon=\epsilon'} \approx 0.$$

(The empirical EIC is in fact differentiable for our parameters of interest. More generally, if $P_n D^*(Q, g)$ were merely continuous and not differentiable at all points in the domain, then the situation above approximately corresponds to the existence of an ϵ' such that $\epsilon^* < \epsilon' < \epsilon_1$ or $\epsilon^* > \epsilon' > \epsilon_1$ implies $0 < |f(\epsilon')| \geq |f(\epsilon_1)|$, where ϵ^* is the root.) In fact the algorithm performs much worse if the position of zero slope is between ϵ_1 and ϵ^*, rather than between ϵ_0 and ϵ^*. Since this is true of the two starting points, (ϵ_0, ϵ_1), it is also true for the points $(\epsilon_k, \epsilon_{k+1})$ corresponding to the k^{th} iteration of the algorithm.

The second difficulty arises when $f(\epsilon)$ approaches zero slowly near the root (see figure 1). In this case the secant method is known to have trouble converging in a reasonable number of iterations even if the starting values yield a value of $f(\epsilon)$ that is relatively close to zero. Several of our simulations at $n = 30$ confirm this. Interestingly, these tend also to be cases in which all the TMLE's give a parameter estimate of either 1 or $1 - \delta$ with $\delta < 0.05$. In these cases, the TMLE is trying to force the estimate to 1. Regardless of the reason for this, the situation is reflected in the empirical EIC, which reveals that the solution ϵ_s, is relatively far from 0. Since ϵ_s is the coefficient in

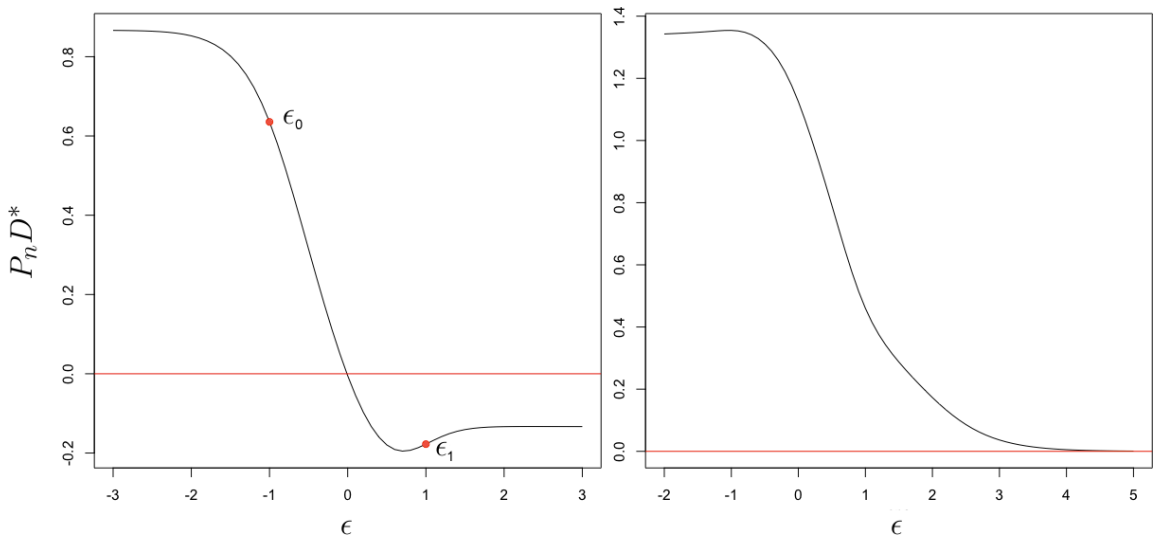

Figure 1: Two examples of $P_n D^*(\epsilon)$ for which the secant method failed to converge at $n = 30$. **Left:** Point of zero slope in starting interval. No convergence with the two indicated starting values (-1,1) in 10 steps or less. Starting values $(\epsilon_0, \epsilon_1) = (0.25, 0.5)$ did yield converence. **Right:** Curve approaches 0 slowly near the root. The true ϵ_s in this case was ≈ 20.81. Starting values (-0.5, 0.5) failed to converge but alternate starting points did yield convergence.

front of the clever covariate term, a large (absolute) value of ϵ_s (assuming a non-negligible clever covariate) will result in a large term in the exponential expression in the denominator of $Q_n(\epsilon)$, and drive $\Psi(Q_n(\epsilon))$ toward 0 or 1. Nevertheless, even in these cases the secant-based TMLE tends to agree with the maximum likelihood-based TMLE's—they all produce estimates very close to 1. These are cases in which the TMLE methods are breaking down due to sparsity. (They are not cases of positivity violation, since g_0 was given.)

Despite these potentially problematic types of curves, the fact that the secant method is not guaranteed to converge in general appears to be no drawback at all in our situation. One can always examine $f(\epsilon)$ in the neighborhood of the root and pick initial estimates in an informed way—i.e., close enough to the root to avoid the potential problems described above. We were able to do this whenever the initial starting values did not result in convergence to a solution in ten iterations of the algorithm or less. It may be that there

are cases in which the empirical EIC behaves so poorly in the neighborhood of the root that this technique fails when there is in fact a solution, but we observed no such cases.

There are also so-called "safeguarded" algorithms which force each iteration to bracket the solution by ensuring that the two current estimates are of opposite sign. The method of false position is one such algorithm (Faires and Burden, 2003). Such methods can be used to guard against divergence of the method, and can be used in place of the secant method if for some reason an a priori guarantee of convergence is required.

5.2 Comparison of the TMLE Algorithms

All targeted minimum loss-based estimators—including the "numerical methods" TMLE—are double-robust, and are efficient under correct model specification.

The advantage of the the numerical methods approach (secant, bisection, Newton, etc.) is that it is the easiest overall to implement, given $K \geq 2$ (where K is the number of time-points at which data is measured). Next in terms of implementation complexity is the one-step algorithm, and finally the iterative approach.

Also noteworthy is that the one-step TMLE requires estimation of two ϵ's in the binary $L(1)$ case and four ϵ's when $L(1)$ has four levels (three for L(1) and one for L(2)). For the general data structure $(L(0), A(0), ...L(K), A(K), L(K+1))$ where intermediate factor $L(j)$ has t_j levels, the number of ϵ's the one-step estimator must fit is $\sum_{j=1}^{K+1}(t_j - 1)$. In contrast, the iterative and numerical solution TMLE's perform a fitting of a single ϵ. It would therefore not be surprising to see at least a small efficiency advantage for the iterative and numerical methods as K and/or t_j increase, though we have not observed any such advantage in the present simulations.

The three methods also differ slightly in terms of computational resources required. For the data simulated here, at $n = 2000$, the order in terms of computational speed was 1) one-step, 2) iterative and 3) secant. However, this result was based on running four iterations of the iterative procedure and imposing a tolerance $|P_n D^*(Q_n(\epsilon))| \leq 10^{-6}$ on the secant algorithm, both of which criteria are overkill. Since typically for the k^{th} iteration of the secant method, $|\psi_k - \psi_{k+1}| \approx |P_n D^*(\epsilon_k)|$, a reasonable tolerance is, say,

$|P_n D^*(\epsilon_k)| \leq var(\psi_k)/10$. Such a tolerance will make the speed of the secant procedure comparable to the iterative procedure.

It is possible that under some conditions the empirical EIC has multiple solutions, or no solution, though we observed no such cases. If multiple solutions, one could select the ϵ_s that yielded the highest likelihood. If the EIC has no solution for the single ϵ approach then the one-step procedure would be favored.

References

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, New York, 1997.

P. Chaffee and M.J. van der Laan. Targeted maximum likelihood estimation for dynamic treatment regimes in sequential randomized controlled trials. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 277.*, 2011.

J. D. Faires and R. Burden. *Numerical Methods*. Thomson Brooks/Cole, Pacific Grove, CA, 3rd edition, 2003.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

M.J. van der Laan. Targeted maximum likelihood based causal inference: Part i. *The International Journal of Biostatistics*, 6(2), 2010a.

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

M.J. van der Laan, S. Rose, and S. Gruber. Readings in targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Available at: http://works.bepress.com/sgruber/6, 2009.