The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 21

Targeted Minimum Loss Based Estimation of a Causal Effect on an Outcome with Known Conditional Bounds

Susan Gruber, *Harvard University*
Mark J. van der Laan, *University of California, Berkeley*

Recommended Citation:
Gruber, Susan and van der Laan, Mark J. (2012) "Targeted Minimum Loss Based Estimation of a Causal Effect on an Outcome with Known Conditional Bounds," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 21.
DOI: 10.1515/1557-4679.1413

Targeted Minimum Loss Based Estimation of a Causal Effect on an Outcome with Known Conditional Bounds

Susan Gruber and Mark J. van der Laan

Abstract

This paper presents a targeted minimum loss based estimator (TMLE) that incorporates known conditional bounds on a continuous outcome. Subject matter knowledge regarding the bounds of a continuous outcome within strata defined by a subset of covariates, X, translates into statistical knowledge that constrains the model space of the true joint distribution of the data. In settings where there is low Fisher Information in the data for estimating the desired parameter, as is common when X is high dimensional relative to sample size, incorporating this domain knowledge can improve the fit of the targeted outcome regression, thereby improving bias and variance of the parameter estimate. We show that TMLE, a substitution estimator defined as a mapping from a density to a (possibly d-dimensional) real number, readily incorporates this global knowledge, resulting in improved finite sample performance.

KEYWORDS: TMLE, targeted maximum likelihood estimation, targeted minimum loss based estimation, boundedness, conditional bounds

Author Notes: This work was supported by the National Institutes of Health (grant no. 5R01AI74345-5) and the National Institutes of Health/National Heart, Lung, and Blood Institute (grant no. R01HL080644).

1 Introduction

We have previously described the improved performance of the targeted minimum loss based estimator (TMLE) when known bounds on the observed data distribution are incorporated into the estimation procedure (Gruber and van der Laan, 2010). This paper extends that result to known bounds that are conditional on measured covariates, and demonstrates there is potential for substantial gains in performance. Subject matter knowledge regarding the bounds of the value of a continuous outcome within strata defined by a subset of covariates, X, translates into statistical knowledge that constrains the model space of the true joint distribution of the data (for example, consider weight for children from birth to age 5, within strata defined by age and gender). When there is sparsity in the data for estimating the desired parameter (i.e., low Fisher Information), as commonly encountered when X is high dimensional relative to sample size, incorporating this domain knowledge may improve the fit of the outcome regression, benefitting precision and accuracy of the estimated causal effect. This effect can be defined as a mapping, Ψ, that maps a probability distribution into a (possibly d-dimensional) real number, $\Psi(P) \mapsto \mathbf{R}^d$. TMLE is a double robust estimator that provides consistent parameter estimates when an outcome regression model or censoring mechanism model is correctly specified (van der Laan and Rubin, 2006, van der Laan and Rose, 2011). Because TMLE is a substitution estimator designed to make a bias/variance tradeoff in the estimate the relevant portion of the true distribution that is favorable with respect to the parameter of interest, it can readily be made to exploit knowledge that constrains the model space, $\mathscr{M}$, of possible probability distributions.

The article is organized as follows. Section 2 provides an overview of TMLE and summarizes previous work on incorporating global bounds in the estimation procedure. Section 3 extends this work to present a TMLE that incorporates conditional bounds, and introduces two functions of treatment (or exposure) and covariates, a and b, that are known to bound the outcome. The use of conditional bounds that can vary among observational units provides an opportunity to improve accuracy and precision of parameter estimates beyond that achieved using global bounds. Section 4 presents Monte Carlo simulations that compare performance of two TMLEs in a point treatment setting, one relying on constant (global) bounds, and the other that makes use of conditional bounds. When the range of outcome values varies widely across strata defined by X, the use of conditional bounds is shown to improve bias and variance. This improvement is most important under misspecification of the outcome regression model, and has a marked effect on efficiency when there is sparsity in the data. Section 5 presents a data analysis where conditional bounds are not known, but are estimated using knowledge from external sources in combination with the data. TMLEs incorporating estimated conditional

and global bounds are applied to estimate the additive effect of smoking on forced expiratory volume (FEV) in children using data from a publicly available dataset (Rosner, 1999b). Incorporating conditional bounds improves the variance of the estimator in comparison with using global bounds. The extension to incorporating conditional bounds in the analysis of longitudinal data is described in the discussion section concluding this article. R code for implementing this TMLE for estimating a binary point treatment effect is provided in an appendix.

2 Boundedness of targeted minimum loss-based estimators

TMLE is an efficient semi-parametric substitution estimator that can be used to estimate any pathwise differentiable parameter at any density p in a class of semi-parametric statistical models, $\mathscr{M}$, given n i.i.d. observations $O_1,\ldots,O_n$ from an underlying distribution (P_0) belonging to $\mathscr{M}$. The target parameter often only depends on P_0 through a relevant part $Q_0 = Q(P_0)$ of P_0, thus $\Psi(P_0) = \Psi(Q_0)$. An estimator of P_0 optimized with respect to a global loss function may be suboptimal with respect to the bias/variance trade-off for the desired parameter. TMLE attempts to improve upon this trade-off by fluctuating the initial estimate of P_0. This fluctuation requires specifying a parametric submodel and loss function, $\mathscr{L}(Q)(O)$, that is minimized at the truth: $Q_0 = \arg\min_{Q\in\mathscr{Q}} E_0\mathscr{L}(Q)(O)$, where $\mathscr{Q} = \{Q(P) : P \in \mathscr{M}\}$. This parametric fluctuation $Q^0_{ng}(\varepsilon)$, possibly indexed by nuisance parameter $g_0 = g(P_0)$, so that

$$\frac{d}{d\varepsilon}\mathscr{L}(Q^0_{ng}(\varepsilon))(O)\bigg|_{\varepsilon=0} = D^*(Q^0_n, g)(O), \tag{1}$$

where $D^*(Q_0, g_0)$ is the canonical gradient/efficient influence curve of $\Psi : \mathscr{M} \to \mathbf{R}$ at P_0. Recall that an estimator is efficient if and only if it is asymptotically linear with influence curve equal to the efficient influence curve $D^*(Q_0, g_0)$ (Bickel, Klaassen, Ritov, and Wellner, 1997). The magnitude of the fluctuation is given by

$$\varepsilon_n = \arg\min_{\varepsilon} \sum_{i=1}^{n} \mathscr{L}(Q^0_{ng_n}(\varepsilon))(O_i),$$

where g_n is an estimator of the unknown nuisance parameter g_0. This yields an update $Q^1_n = Q^0_{ng_n}(\varepsilon_n)$. This updating of an initial estimator Q^0_n into a next Q^1_n is iterated until convergence resulting in a final targeted estimate Q^*_n. The magnitude of

the fluctuation at each iteration corresponds with the degree of residual confounding. Since at the last step the amount of fluctuation $\varepsilon_n \approx 0$, this final Q_n^* will solve the efficient influence curve estimating equation

$$0 = \sum_{i=1}^{n} D^*(Q_n^*, g_n)(O_i),$$

representing a fundamental ingredient for establishing asymptotic efficiency of $\Psi(Q_n^*)$. Finally, the targeted MLE of ψ_0 is the substitution estimator $\Psi(Q_n^*)$.

Thus we see that the targeted MLE involves constructing a parametric model $Q_n^0(\varepsilon)$ through the initial estimator Q_n^0 with parameter ε representing an amount of fluctuation of the initial estimator, where the score of this fluctuation model at $\varepsilon = 0$ equals the efficient influence curve. This local constraint on the behavior at zero fluctuation can be satisfied by many parametric models. However, it is very important that the fluctuations stay within the model for the observed data distribution, even if the parameter can be defined on fluctuations that fall outside the assumed observed data model. In particular, in the context of sparse data (i.e., little information in the data for identifying the target parameter), a violation of this property can heavily affect the performance of the estimator. We have previously shown that estimator performance suffers when the fluctuation model is not guaranteed to stay within the specified observed data model, and defined a TMLE procedure that incorporates constant global bounds on the outcome (Gruber and van der Laan, 2010). However, these bounds can at times be overly conservative, and as we demonstrate in the remainder of the paper, in finite samples incorporating less conservative bounds that are conditional on measured covariates can lead to improved bias and variance and robustify the estimator against model misspecification.

3 A targeted minimum loss-based estimator that incorporates conditional bounds

We observe n i.i.d. copies of $O = (W,A,Y)$, where W is a vector of baseline covariates, A is a binary treatment or exposure indicator taking on the value 1 when the subject is treated and 0 when the subject is untreated. Let P_0 be its probability distribution. The likelihood of the data factorizes as $P_0(Y,A,W) = P_0(Y \mid A,W)P_0(A \mid W)P_0(W)$. In our notation $Q_W(P_0)$ is the marginal distribution of W and $Q_Y(P_0)$ is the conditional distribution of Y given A and W. Let g_0 be the conditional probability distribution of A, given W. Let $X \equiv (W,A)$. We assume that $P_0(Y \in [a(X),b(X)] \mid X) = 1$ for known functions $a(X),b(X)$ with $a(X) \leq b(X)$. We also assume that $g_0 \in \mathscr{G}$ for a set of possible conditional distributions $\mathscr{G}$, where

$\mathscr{G}$ could be nonparametric in which case nothing is assumed about g_0. We make no assumptions about the marginal distribution of W. Beyond the support-constraint above, we make no further assumptions about the conditional distribution of Y, given A,W. This defines the statistical model $\mathscr{M}$, i.e., the set of possible probability distributions that is known to contain the true probability distribution P_0.

Our statistical target parameter $\Psi : \mathscr{M} \to \mathbb{R}$ is defined as

$$\psi_0 = \Psi(P) = E_P\{E_P(Y \mid A=1,W) - E_P(Y \mid A=0,W)\}.$$

Note that $\Psi(P)$ only depends on P through the marginal distribution $Q_W(P)$ of W and the conditional mean $\bar{Q}(P)$ of Y, given A,W. Therefore, with abuse of notation, we will also use the notation $\Psi(Q)$ where $Q = (Q_W, \bar{Q})$. This target parameter is pathwise differentiable and its canonical gradient is given by

$$D^*(P)(O) = H_g(A,W)(Y - \bar{Q}(A,W)) + \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(Q),$$

where $H_g(A,W) = (2A-1)/g_0(A \mid W)$. This canonical gradient is not affected by the choice of model $\mathscr{G}$ for g_0 and the global constraints defined by the functions a and b.

We wish to construct a TMLE of ψ_0, which will yield a substitution estimator $\Psi(Q_n^*)$ with $Q_n^* = (Q_{W,n}, \bar{Q}_n^*)$ a targeted estimate of Q_0 that fully respects the statistical model. Firstly, we note that we can parameterize

$$E(Y \mid X) = (b(X) - a(X))E(Y^{\#} \mid X) + a(X), \text{ where } Y^{\#} = \tfrac{Y - a(X)}{b(X) - a(X)}.$$

The parameter space of $\bar{Q} = (b-a)\bar{Q}^{\#} + a$ can thus be represented as follows:

$$\{(b-a)\bar{Q}^{\#} + a : \bar{Q}^{\#} \text{maps into} (0,1)\},$$

where we are suppressing the dependence of a and b on (A,W). $Y^{\#}$ and $Q^{\#}$ are the result of shifting and scaling operations designed so that $Y^{\#}$ and $Q^{\#}$ are constrained to lie in (0,1).

The efficient influence curve can be represented accordingly as a function of $(Q_W, \bar{Q}^{\#}, g)$:

$$\begin{aligned} D^*(Q,g)(O) &= H_g(A,W)(b-a)(A,W)(Y^{\#} - \bar{Q}^{\#}(A,W)) \\ &\quad + (b-a)(1,W)\bar{Q}^{\#}(1,W) - (b-a)(0,W)\bar{Q}^{\#}(0,W) - \Psi(Q) \\ &\equiv D_Y(g,\bar{Q}^{\#}) + D_W(Q), \end{aligned}$$

where $D_Y(g,\bar{Q}^{\#}) = H_g(A,W)(b-a)(Y^{\#} - \bar{Q}^{\#}(A,W))$ is a score of the conditional distribution of Y, given A,W, and $D_W(Q)$ is a score of marginal distribution of W.

We define the following loss function for $\bar{Q}^{\#} = E(Y^{\#} \mid X)$:

$$-\mathscr{L}(\bar{Q}^{\#})(O) = Y^{\#} \log \bar{Q}^{\#}(X) + (1 - Y^{\#}) \log(1 - \bar{Q}^{\#}(X)).$$

Indeed, this is a valid loss function for $\bar{Q}_0^{\#}$:

$$\bar{Q}_0^{\#} = \arg\min_{\bar{Q}^{\#}} P_0 \mathscr{L}(\bar{Q}^{\#}),$$

where we used the notation $P_0 f \equiv \int f(o) dP_0(o)$. We can use $\mathscr{L}_W(Q_W) = -\log Q_W$ as loss-function for Q_W, which gives us the sum loss function $\mathscr{L}(Q) = \mathscr{L}(\bar{Q}^{\#}) + \mathscr{L}_W(Q_W)$ for $Q = (Q_W, \bar{Q})$. As the least favorable fluctuation of $\bar{Q}^{\#}$ we use

$$\text{Logit}\bar{Q}^{\#}(\varepsilon) = \text{Logit}\bar{Q}^{\#} + \varepsilon H_g(b - a).$$

In our notation we suppressed the dependence of these fluctuation models on g. As the least favorable fluctuation of Q_W we can use $Q_W(\varepsilon) = (1 + \varepsilon D_W(Q)) Q_W$. Indeed,

$$\begin{aligned}
\frac{d}{d\varepsilon} \mathscr{L}(\bar{Q}^{\#}(\varepsilon)) \Big|_{\varepsilon=0} &= D_Y(Q, g) \\
\frac{d}{d\varepsilon} \mathscr{L}_W(Q_W(\varepsilon)) \Big|_{\varepsilon=0} &= D_W(Q), \\
\frac{d}{d\varepsilon} \mathscr{L}(Q(\varepsilon)) \Big|_{\varepsilon=0} &= D^*(Q, g).
\end{aligned}$$

The TMLE can now be defined as follows.

Let $\bar{Q}_n^{\#,0}$ be an initial estimator of $\bar{Q}_0^{\#}$, which could be a loss-based learner based on loss function $\mathscr{L}(\bar{Q}^{\#})$. Let $Q_{W,n}$ be the empirical distribution of $W_1, \ldots, W_n$. Let g_n be an estimator of g_0. This defines the initial estimator Q_n^0 of Q_0 and g_n of g_0. Define

$$\begin{aligned}
\varepsilon_{1n} &= \arg\min_{\varepsilon} P_n \mathscr{L}(\bar{Q}_n^{\#,0}(\varepsilon)) \\
\varepsilon_{2n} &= \arg\min_{\varepsilon} P_n \mathscr{L}_W(Q_{W,n}(\varepsilon)),
\end{aligned}$$

or equivalently, define the two dimensional $\varepsilon_n = \arg\min_{\varepsilon} P_n \mathscr{L}(Q_n^0(\varepsilon))$.

Note that $\varepsilon_{2n} = 0$, since $Q_{W,n}$ is an NPMLE. The TMLE of Q_0 is now defined by $Q_n^* = (Q_{W,n}, \bar{Q}_n^{\#,*} = \bar{Q}_n^{\#,0}(\varepsilon_{1n}))$, and, accordingly,

$$\bar{Q}_n^*(A,W) = (b-a)(A,W)\bar{Q}_n^{\#,*}(A,W) + a(A,W).$$

The TMLE of $\Psi(Q_0)$ is thus given by

$$\begin{aligned} \Psi(Q_n^*) &= \frac{1}{n}\sum_{i=1}^{n}\left\{(b-a)(1,W_i)\bar{Q}_n^{\#,*}(1,W_i) + a(1,W_i)\right\} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\left\{(b-a)(0,W_i)\bar{Q}_n^{\#,*}(0,W_i) + a(0,W_i)\right\}. \end{aligned} \tag{2}$$

If $a(X)$ only depends on X through W, then $a(1,W_i) - a(0,W_i) = 0$ cancels out. By definition of the loss and least favorable submodel, the TMLE solves $P_n D_W(Q_n^*) = P_n D_Y(\bar{Q}_n^*, g_n) = 0$, and thereby the efficient influence curve equation:

$$0 = P_n D^*(Q_n^*, g_n).$$

This TMLE is robust since $\bar{Q}_n^*(X) \in [a(X), b(X)]$ a.e.

Even if there are practical violations of the positivity assumption that $P(A = \mathsf{a}|W) > 0, \forall \mathsf{a} \in A$, so that $g_n(1 \mid W)$ is close to zero or 1, the TMLE update $\bar{Q}_n^*$ is guaranteed to fall between these two known functions a and b. Previously, we already observed the enormous importance of constructing a TMLE that respects fixed bounds a, b (Gruber and van der Laan, 2010), and thereby we can expect that if these known bounds vary as a function of X, further gains in robustness and finite sample efficiency can be achieved. For example, consider the case where $X = (A,W)$, and a and b are increasing functions in W with steep slopes. Under near violations of the positivity assumption, a TMLE that uses the actual functions (a,b) would be more stable than a TMLE that used constant bounds $a_m = \min_x a(x)$ and $b_m = \max_x b(x)$. In addition, consider the case where $a(X) \approx b(X)$ while $b_m - a_m$ is large. In this setting X is very predictive of the range of Y, and exploiting this knowledge should lead to a more precise estimator of $\bar{Q}_0$ in finite samples that should improve efficiency and bias, even though the asymptotic efficiency is not affected. Simulation studies presented in the next section demonstrate estimator performance under these two type of scenarios.

4 Simulation studies

Monte Carlo simulations were carried out to compare the performance of two TMLEs that differ only in the two methods used to enforce known constraints on

the model space. The target parameter is the marginal average treatment effect (ATE) discussed above, $\psi_0 = E_{P_0}\{E_{P_0}(Y \mid A = 1, W) - E_{P_0}(Y \mid A = 0, W)\}$. Data for the two studies are generated so that the measured continuous outcome falls between bounds $(a(W), b(W))$, independent of treatment assignment. Estimates were obtained for a TMLE that incorporates fixed bounds, (a_m, b_m), into the estimation procedure, and a second TMLE that incorporates conditional bounds $(a(W), b(W))$, with a and b known functions of W. Recall that TMLE requires initial estimates $(\bar{Q}_n^0(A,W), g_n(1,W))$, and is consistent when either one of these is correctly specified. A correctly specified $\bar{Q}_n^0(A,W)$ implies that bounds are inherently respected, so we would not expect to see much difference in the performance of these two TMLEs in this case. However, when $\bar{Q}_n^0(A,W)$ is misspecified but $g_n(1,W)$ is correct, bounding should make a difference in finite sample performance, particularly when there is sparsity in the data. We use the term sparsity to refer to a lack of information in the data to identify the target parameter. Sparsity is signaled by large variance in the empirical influence curve, and will occur when g_0 is close to 0 or 1. Under dual misspecification, enforcing the constraints imposed by the bounds also places limits on an estimator's bias and variance, but the effect this has on the bias and variance of the parameter estimate will depend on the particular data generating mechanism.

4.1 Simulation study 1

In the first example a and b are functions that have steep slopes such that $(\bar{Q}_0(\mathsf{a}, W_i) - \bar{Q}_0(\mathsf{a}, W_j))$ is large relative to the dissimilarity in (W_i, W_j). In this context we would expect that bounds set at values (a_m, b_m) are overly broad, and that tighter bounds conforming more closely to the underlying distribution of Y given W will yield estimates with lower mean squared error when at least one of $\bar{Q}_0$ and g_0 is correctly specified.

4.1.1 Data generation

Baseline covariates (W_1, W_2) were generated for 1000 datasets of size $n = 1000$. A binary treatment indicator A was generated according to treatment mechanism g_0 specified below. A two-step procedure was used to generate the observed outcome. First a continuous variable, Y^* bounded by (0,1), was generated conditional on $(A, W_1, W2)$. Next, the observed outcome Y was calculated as $Y = Y^*(b_{steep}(W) - a_{steep}(W)) + a_{steep}(W)$. This data generating procedure was carried out for two choices of g_0, to generate one collection of 1000 datasets where there is no sparsity in the data, $O_{nosp} = (Y_{nosp}, A_{nosp}, W_1, W_2)$, and a corresponding collection of 1000

datasets where there is sparsity in the data, $O_{sp} = (Y_{sp}, A_{sp}, W_1, W_2)$. Specifically, we used the following data generating distributions:

$$\begin{aligned}
W_1 &\in U(1,2) \\
W_2 &\in Bern(0.5) \\
a_{steep}(W) &= 3W_1 \\
b_{steep}(W) &= 5W_1 + \sqrt{W_1} \\
g_{0,1} = P(A_{nosp} = 1 \mid W) &= expit(-0.5 + 0.4W_1 + 1.75W_2) \\
Y^*_{nosp} &= (A_{nosp} - 0.7W_1 + W_2 + \varepsilon_1 + 4.6)/9 \\
Y_{nosp} &= Y^*_{nosp}(b_{steep}(W_1) - a_{steep}(W_1)) + a_{steep}(W_1)
\end{aligned}$$

$$\begin{aligned}
g_{0,2} = P(A_{sp} = 1 \mid W) &= expit(1.5 + 0.4W_1 + 1.75W_2) \\
Y^*_{sp} &= (A_{sp} - 0.7W_1 + W_2 + \varepsilon_1 + 4.6)/9 \\
Y_{sp} &= Y^*_{sp}(b_{steep}(W_1) - a_{steep}(W_1)) + a_{steep}(W_1)
\end{aligned}$$

with ε_1 and $\varepsilon_2 \sim_{i.i.d.} N(0,1)$. A_{nosp} refers to the treatment indicator vector generated according to $g_{0,1}$, and there are no positivity violations. A_{sp} was generated according to $g_{0,2}$, leads to sparsity. Outlying values of Y^*_{nosp} and Y^*_{sp} that fell outside (0,1) were truncated. The true parameter value is $\psi_{0,1} = 0.4686$.

4.1.2 Results

When Q_0 is correctly specified both estimators perform well regardless of whether g_0 is correctly specified or misspecified as the intercept model (Table 1). As anticipated, when there is sparsity and Q_0 is misspecified as the unadjusted regression of Y on A, the variance is much less when the conditional bounds $(a(W), b(W))$ are used instead of constant bounds (a_m, b_m). The respecting of the conditional bounds also reduced bias under dual misspecification.

4.2 Simulation study 2

The data generated for this study incorporates functions a and b that are steep, but also approximately equal. Although $(b(W) - a(W))$ is small, these functions were defined specifically to ensure that $b_m - a_m$ is large. In this case $b_m - a_m = 20$.

Table 1: Simulation 1 Empirical bias, variance, MSE of 1000 estimates for TMLEs incorporating global or conditional bounds, $n = 1000$, $\psi_{0,1} = 0.4686$.

	Global Bds			**Cond. Bds**		
	bias	var	MSE	bias	var	MSE
No sparsity						
Q correct						
g cor	0.0004	0.0015	0.0015	0.0004	0.0015	0.0015
g mis	−0.0043	0.0012	0.0012	0.0001	0.0012	0.0012
Q misspecified						
g cor	0.0013	0.0026	0.0026	0.0003	0.0015	0.0015
g mis	0.2949	0.0064	0.0933	0.1769	0.0013	0.0326
Sparsity						
Q correct						
g cor	−0.0013	0.0074	0.0074	−0.0011	0.0075	0.0075
g mis	−0.0113	0.0038	0.0039	−0.0008	0.0038	0.0038
Q misspecified						
g cor	0.0025	0.0245	0.0245	−0.0035	0.0081	0.0081
g mis	0.2919	0.0252	0.1104	0.1600	0.0043	0.0299

4.2.1 Data generation

The data generation scheme from the first study was modified slightly by re-defining $a(W) = 20W_1$ and $b(W) = 1 + 20W_1$. The other equations remain unchanged, and estimates are again obtained for two collections of 1000 datasets, first where there is no sparsity in the data, and second where there is sparsity in the data. Because the observed outcome is on a different scale than in the first study, the marginal additive treatment effect is also different, $\psi_{0,2} = 0.0611$.

4.2.2 Results

Both TMLEs perform equally well when Q_0 is correctly specified, The use of conditional bounds $(a(W), b(W))$ greatly improves bias and variance when Q_0 is misspecified as the unadjusted regression of Y on A, with and without sparsity in the data (Table 2). Bias and variance are also greatly reduced when both Q_0 and g_0 are misspecified (where the misspecified g is the intercept model) .

Table 2: Simulation 2, Empirical bias, variance, MSE of 1000 estimates for TMLEs incorporating global or conditional bounds, $n = 1000$, $\psi_{0,2} = 0.0611$.

	Global Bds			**Cond. Bds**		
	bias	var	MSE	bias	var	MSE
No sparsity						
Q correct						
g cor	0.0501	0.0001	0.0026	0.0501	0.0001	0.0026
g mis	0.0500	0.0001	0.0026	0.0500	0.0001	0.0026
Q misspecified						
g cor	0.0548	0.0261	0.0291	0.0501	0.0001	0.0026
g mis	0.6636	0.1397	0.5801	0.0919	0.0001	0.0085
Sparsity						
Q correct						
g cor	0.0496	0.0004	0.0029	0.0496	0.0004	0.0029
g mis	0.0498	0.0002	0.0027	0.0498	0.0002	0.0027
Q misspecified						
g cor	0.0686	0.4561	0.4608	0.0495	0.0004	0.0029
g mis	0.7423	0.5920	1.1429	0.0880	0.0002	0.0080

5 FEV data analysis

TMLE was applied to assess the marginal additive effect of smoking on forced expiratory volume (FEV) using data originally introduced in Rosner (1999b) and discussed in Kahn (2005). The data consists of 654 observations with five variables recorded for each subject: *age* (years), *fev* (liters), *ht* (height in inches, converted to centimeters for these analyses), *sex* (0=female, 1=male), *smoke* (0=non smoker, 1=smoker) (Rosner, 1999a). FEV is a measure of pulmonary function that is related to body size and lung capacity. Thus, the relationship between smoking and FEV is likely to be confounded by age and sex, both of which influence FEV and are associated with smoking status. Though height does not have an obvious link to smoking behavior, accounting for covariates predictive of the outcome can improve efficiency even if they are not confounders (van der Laan and Robins, 2003). The data are from an observational study of children 3 -19 years old, but because no children younger than nine years old smoked cigarettes we restrict the analysis to the subset of data containing observations on subjects ages 9 - 19 ($n = 439$).

Hankinson, Odencrantz, and Fedan (1999) constructed parametric regression models for predicting mean FEV values in children as a function of age, sex, and height (ht). Their regression model is $fev = \beta_0 + \beta_1 age + \beta_2 ht + \beta_3 ht^2$, with coefficients fit separately among boys and girls. The coefficient values they report are:

$$\begin{aligned} girls: &\quad \beta_0 = -0.8710, \quad \beta_1 = \quad 0.06537, \quad \beta_2 = 0, \qquad\quad \beta_3 = 0.00011496, \\ boys: &\quad \beta_0 = -0.74453, \quad \beta_1 = -0.04106, \quad \beta_2 = 0.004477, \quad \beta_3 = 0.00014098. \end{aligned}$$

These formulas allowed us to estimate subject-specific conditional bounds $(a(W), b(W))$ as 3 standard deviations (SD) above and below the predicted mean value for each subject. The SD was estimated as the square root of the MSE of the residuals from the linear regression, $\widehat{SD} = 0.443$. When the conditional bounds were set to $E(fev \mid age, sex, ht) \pm 3SD$ no measured FEV values in the dataset fell outside these bounds, however the procedure does not ensure that this will be the case. It is important to construct estimated bounds that do not contradict the data. For example, because FEV must be a positive number, any value for $a(W)$ less than 0 can be reset to 0, or perhaps to the smallest calculated non-zero value of $a(W)$. Figure 1 shows the recorded FEV values for each subject in the dataset, sorted by FEV value. The red dots are the upper and lower conditional bounds, and the blue dotted lines show the global bounds.

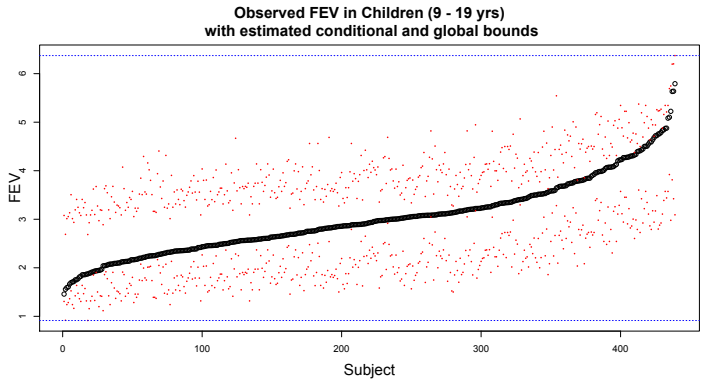

Figure 1: Measured FEV values for each subject. Red dots are subject-specific estimated upper and lower conditional bounds. Blue dotted lines show the estimated global bounds.

We obtained three separate estimates of the target parameter: 1) the untargeted G-computation estimate based on the initial fit Q_n^0 from the specified parametric regression model (Robins, 1986) , 2) a TMLE incorporating constant global

bounds that were set to the minimum and maximum values of the conditional bounds, $(a = a_m, b = b_m)$, and 3) a second TMLE that incorporates conditional bounds $a(X)$ and $b(X)$. The variance of these estimates was estimated as the variance of 1000 bootstrap samples.

The first step of the TMLE procedure that incorporates global bounds was to obtain an initial fit $\bar{Q}_n^{\#0} = E(Y^{\#} \mid A, W)$ by performing a linear regression of $Y^{\#} = (Y - a_m)/(b_m - a_m)$ on A and W. An initial estimate of g_0 was used to construct the covariate $H_g = (2A - 1)/g(A|W)$ to fluctuate this initial estimate on the logit scale:

$$logit(\bar{Q}_n^{\#*}) = logit(\bar{Q}_n^{\#0}) + \varepsilon H_g(b_m - a_m).$$

ε was estimated by fitting a logistic regression of $Y^{\#}$ on $H_g(b_m - a_m)$ with offset $logit(\bar{Q}_n^{\#0})$. The ATE parameter is evaluated as in Eq.(2) above. The procedure for incorporating conditional bounds is identical, except for substituting $a(X)$ for a_m and $b(X)$ for b_m.

We ran these analyses using two different regression models for estimating $\bar{Q}_0$. The first adds the treatment indicator to the model constructed by Hankinson et. al., a regression of *fev* on binary smoking indicator $A = smoke$, age, sex age^2, and ht^2. The second is an unadjusted regression of *fev* on A, presumably a biased estimator due to confounding. Bias reduction depends upon consistent estimation of g_0. We used the DSA algorithm to select a regression model for estimating $P(A = 1 \mid sex, ht, age)$. DSA is a deletion-substitution-addition algorithm for model selection over a space of polynomials (Sinisi and van der Laan, 2004, Neugebauer and Bullard, 2010). For this analysis DSA was used to search over a space of polynomials of order 3, over models that included up to 10 terms. The selected model included the covariates *sex, age, ht,* ht^2, ht^3, age^3.

5.1 Results

Table 3 lists the additive treatment effect estimates and bootstrapped variance estimates for each estimator. Results in the row labelled $\bar{Q}_H$ were based on the regression model of Hankinson, et. al. Though the truth is not known, the agreement between the point estimates suggests that smoking decreases FEV. Targeting the initial estimate had a small effect on bias, and greatly increased the variance, however the variance of the TMLE that used conditional bounds is much smaller than that of the TMLE incorporating global bounds.

Next consider the results when the outcome regression model is severely misspecified as the unadjusted regression of *fev* on A. The unadjusted estimate of 0.3 is quite large, and presumably in the wrong direction, yet the variance is small. TMLE using global bounds greatly reduced the bias at a cost of increased variance,

Table 3: Targeted and untargeted estimates of the marginal additive effect of smoking on FEV. Results labelled $\bar{Q}_H$ were based on the regression model of Hankinson et. al., those labelled $\bar{Q}_{unadj}$ are based on the unadjusted regression of *fev* on treatment. Variance estimates were obtained from 1000 bootstrap samples.

	Untargeted		**Global Bds**		**Cond. Bds**	
	Est	Var	Est	Var	Est	Var
$\bar{Q}_H$	-0.156	0.007	-0.157	0.102	-0.123	0.024
$\bar{Q}_{unadj}$	0.304	0.011	0.030	0.049	-0.105	0.024

and the point estimate is still positive. TMLE using conditional bounds achieved greater bias reduction while paying a smaller price in variance.

These results highlight the problems that can arise from model misspecification, and illustrate the benefits of double-robust estimation in practice in finite samples. Smoking is rare in the data. Only 15% of subjects smoke, and among children 9-10 years old (40% of subjects), the smoking rate is only 3.6%. This lack of experimentation in the data is the kind of sparsity that poses a challenging estimation problem. When there is little direct evidence in the data, the untargeted estimator relies heavily on extrapolation, and its low variance masks the underlying lack of information. The quality of the fit of the initial estimate of $\bar{Q}_0$ had little effect on the variance of the untargeted parametric model-based estimator, however the nominal 95% confidence interval centered around the biased estimate based on $\bar{Q}^0_{n,unadj}$ is (0.10, 0.51), and almost certainly fails to include the true parameter value. The higher variance of the TMLEs reflects the true uncertainty inherent in the data. The use of conditional bounds was shown to improve performance over the use of global bounds.

6 Discussion

Ensuring that an estimated conditional mean outcome remains within the parameter space implicitly bounds the estimate of any parameter that is a function of conditional means, such as the average additive treatment effect, odds ratio, and risk ratio. We demonstrated that a TMLE that respects conditional bounds on the outcome is easily constructed by incorporating these bounds into the initial estimator and in the clever covariate used to fluctuate the initial estimate of the conditional mean outcome (re-scaled to lie between 0 and 1). The simulation studies used a binary

point treatment example to illustrate how this TMLE can be applied in practice, and demonstrated the potential for performance gains in finite samples. If the bounds are unknown, but are known to only depend on a discrete covariate, then one could estimate the bounds empirically. Since estimates of a minimum and maximum of a support converge at a rate faster than $1/\sqrt{n}$, the resulting estimator would asymptotically behave as if the bounds were known. Estimating unknown functions of continuous covariates is a harder problem, and might contribute to uncertainty. In our earlier article (Gruber and van der Laan, 2010) we demonstrated that the TMLE using estimated constant bounds indeed performed well in practice. Though we don't know the truth in the FEV data analysis, results indicate that when $\bar{Q}_0$ is estimated well little residual bias remains, and all estimators perform well. When the initial $\bar{Q}_n^0$ is misspecified, as is generally the case when analyzing finite samples, the double-robustness of TMLE provides the opportunity to reduce the bias, and depending upon the rate at which $\bar{Q}_n^0$ converges to the truth, may reduce variance as well.

We expect this feature will prove valuable in estimating causal effects in a longitudinal setting, where sparsity is often an issue. Consider a longitudinal data structure $O = (L(0), A(0), \ldots, L(K), A(K), Y = L(K+1))$, where $L(0)$ are baseline covariates, $A(t)$ denotes an exposure or treatment at time t, $L(t)$ denotes covariates measured between two subsequent treatments $A(t-1)$ and $A(t)$, and Y is the final outcome measured after the final treatment. Estimating the effect of a treatment regime over time requires that there be experimentation within strata defined by the entire time-dependent and baseline covariate history, a requirement that becomes increasingly unlikely as a rich covariate history accumulates over time. Stitelman, Gruttola, and van der Laan (2011) present a TMLE for estimating the causal effect of treatment on survival in longitudinal data. van der Laan and Gruber (2012) describe an alternate longitudinal TMLE that builds upon an estimator described by Bang and Robins (2005) that iteratively estimates a sequence of conditional expectations in order to estimate the mean outcome under a particular treatment regime $a(0), \ldots a(K)$. If it is known that, conditional on $\bar{L}(k-1), \bar{A}(k-1)$, $L(k)$ falls between two known values (i.e., functions of $\bar{L}(k-1), \bar{A}(k-1)$) with probability one, then this knowledge can be incorporated in the initial estimators and fluctuations of these TMLEs, analoguously to the method presented here. One may expect that incorporating known bounds is even more important for estimation with longitudinal data, so that significant gains might be achieved.

Appendix

The R function *tmle_bd* implements a TMLE for estimating the marginal additive treatment effect that incorporates either conditional or global bounds on a continuous outcome. Required arguments are Y, a vector of continuous outcomes, A, a binary treatment vector, W, a matrix of baseline covariates, a, a constant or a vector of conditional lower bound values, b, a constant or a vector of conditional upper bound values, $gform$, a regression formula used to estimate the regression of treatment A on covariates W, and $Qform$, a regression formula used to estimate the initial regression of $Y^{\#}$ on A and W.

```
tmle_bd <- function(Y, A, W, a, b, gform, Qform, family = "gaussian"){
    Y.hash <- (Y - a)/(b - a)
    m <- glm(Qform, data = data.frame(Y = Y.hash, A, W), family = family)
    Qinit <- cbind(QAW = predict(m, type = "response"),
        Q1W = predict(m, newdata = data.frame(A = 1, W), type = "response"),
        Q0W = predict(m, newdata = data.frame(A = 0, W), type = "response"))
    logitQ <- qlogis(bound(Qinit, c(0.001, 0.999)))
    g <- glm(gform, data = data.frame(A, W), family = binomial)
    g1W <- predict(g, type = "response")
    h <- (b - a) * (A/g1W - (1 - A)/(1 - g1W))
    eps <- coef(glm(Y.hash ~ -1 + offset(logitQ[ ,"QAW"]) + h,
        family = "quasibinomial"))
    Q.hash <- plogis(logitQ + eps*cbind(h, (b - a)/g1W, -(b - a)/(1 - g1W)))
    return(mean((b-a) * (Q.hash[ ,"Q1W"] - Q.hash[ ,"Q0W"])))
}

bound <- function(x, bounds){
    x[x < min(bounds)] <- min(bounds)
    x[x > max(bounds)] <- max(bounds)
    return(x)
}
```

References

H. Bang and J.M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

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

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6:article 26, www.bepress.com/ijb/vol6/iss1/26, 2010.

J.L. Hankinson, J. R. Odencrantz, and K. B. Fedan. Spirometric reference values from a sample of the general U. S. population. *American Journal of Respiratory and Critical Care Medicine*, 159:179 – 87, 1999.

M. Kahn. An exhalent problem for teaching statistics. *The Journal of Statistical Education*, 13(2), 2005.

R. Neugebauer and J. Bullard. *DSA: Deletion/Substitution/Addition Algorithm*, 2010. URL `http://www.stat.berkeley.edu/ laan/Software/`. R package version 3.1.4.

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. *Math Mod*, 7:1393–1512, 1986.

B. Rosner. Fev dataset, 1999a. Submitted by M.J. Kahn, Wheaton College, Norton, MA.

B. Rosner. *Fundamentals of Biostatistics, 5th Ed.* Duxbury Press, Pacific Grove, CA, 1999b.

S. Sinisi and M.J. van der Laan. The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Journal of Statistical Methods in Molecular Biology*, 3(1), 2004.

O. M. Stitelman, V. De Gruttola, and M. J. van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. Technical report, Division of Biostatistics, University of California, Berkeley, April 2011.

M.J. van der Laan and S. Gruber. Targeted maximum loss based estimation of an intervention specific mean. *The International Journal of Biostatistics*, (in press), 2012.

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

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

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