Chapter 7
Bounded Continuous Outcomes

Susan Gruber, Mark J. van der Laan

This chapter presents a TMLE of the additive treatment effect on a bounded continuous outcome. A TMLE is based on a choice of loss function and a corresponding parametric submodel through an initial estimator, chosen so that the loss-function-specific score of this parametric submodel at zero fluctuation equals or spans the efficient influence curve of the target parameter. Two such TMLEs are considered: one based on the squared error loss function with a linear regression model, and one based on a quasi-log-likelihood loss function with a logistic regression submodel. The problem with the first TMLE is highlighted: the linear regression model is not a submodel and thus does not respect global constraints implied by the statistical model. It is theoretically and practically demonstrated that the TMLE with the logistic regression submodel is more robust than a TMLE based on least squares linear regression. Some parts of this chapter assume familiarity with the core concepts, as presented in Chap. 5. The less theoretically trained reader should aim to navigate through these parts and focus on the practical implementation and importance of the presented TMLE procedure. This chapter is adapted from Gruber and van der Laan (2010b).

7.1 Introduction

TMLE of a target parameter of the data-generating distribution, known to be an element of a semiparametric model, involves selecting a loss function (e.g., log-likelihood) and constructing a parametric *submodel* through an initial density estimator with parameter ϵ, so that the loss-function-specific "score" at $\epsilon = 0$ equals or spans the efficient influence curve (canonical gradient) at the initial estimator. This ϵ represents an amount of fluctuation of the initial density estimator. The latter "score" constraint can be satisfied by many loss functions and parametric submodels, since it represents only a local constraint of the submodels' behavior at zero fluctuation.

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_7,

However, it is very important that the fluctuations encoded by the parametric model stay within the semiparametric model for the observed data distribution (otherwise it is not a submodel!), even if the target parameter can be defined on fluctuations that fall outside the assumed observed data model.

In particular, in the context of sparse data, by which we mean situations where the generalized Cramer–Rao lower bound is high, a violation of this property can significantly affect the performance of the estimator. We demonstrate this in the context of estimation of a causal effect of a binary treatment on a continuous outcome that is bounded. It results in a TMLE that inherently respects known bounds and consequently is more robust in sparse data situations than a TMLE using a naive parametric fluctuation working model that is actually not a *submodel* of the assumed statistical model.

Sparsity is defined as low information in a data set for the purpose of learning the target parameter. Formally, the Fisher information I is defined as sample size n divided by the variance of the efficient influence curve: $I = n/\text{var}(D^*(O))$, where $D^*(O)$ is the efficient influence curve of the target parameter at the true data-generating distribution. The reciprocal of the variance of the efficient influence curve can be viewed as the information one observation contains for the purpose of learning the target parameter. Since the variance of the efficient influence curve divided by n is the asymptotic variance of an asymptotically efficient estimator, one can also think of the information I as the reciprocal of the variance of an efficient estimator of the target parameter. Thus, sparsity with respect to a particular target parameter corresponds with small sample size relative to the variance of the efficient influence curve for that target parameter.

The following section begins with background on the application of TMLE methodology in the context of sparsity and its power relative to other semiparametric efficient estimators since it is a substitution estimator respecting global constraints of the semiparametric model. Even though an estimator can be asymptotically efficient without utilizing global constraints, the global constraints are instrumental in the context of sparsity with respect to the target parameter, motivating the need for semiparametric efficient *substitution* estimators, and for a careful choice of fluctuation function for the targeting step that fully respects these global constraints. A rigorous demonstration of the proposed TMLE of the causal effect of a binary treatment on a bounded continuous outcome follows, and the TMLE using a linear fluctuation function (i.e., that does not represent a parametric submodel) is compared with the proposed TMLE using a logistic fluctuation function. In Sect. 7.3, we carry out simulation studies that compare the two TMLEs of the causal effect, with and without sparsity in the data. Results for other commonly applied estimators discussed in Chap. 6 (MLE according to a parametric statistical model, IPTW, and A-IPTW) are also presented.

7.2 TMLE for Causal Estimation on a Continuous Outcome

We first review general TMLE so that we can clarify the important role of the choice of parametric working model, and thereby the fluctuation function, that defines the targeting update step of the initial estimator. Subsequently, in order to be specific, we define TMLE of the additive causal effect of a binary treatment on a bounded continuous outcome, which fully respects the known global bounds. Finally, we discuss its robustness in finite samples in the context of sparsity.

7.2.1 A Substitution Estimator Respecting the Statistical Model

A TMLE is a semiparametric efficient substitution estimator of a target parameter $\Psi(P_0)$ of a true distribution $P_0 \in \mathcal{M}$, known to be an element of a statistical model $\mathcal{M}$, based on sampling n i.i.d. $O_1, \ldots, O_n$ from P_0. Firstly, one notes that $\Psi(P_0) = \Psi(Q_0)$ only depends on P_0 through a relevant part $Q_0 = Q(P_0)$ of P_0. Secondly, one proposes a loss function $L(Q)$ such that

$$Q_0 = \arg\min_{Q \in \mathcal{Q}} E_0 L(Q)(O),$$

where $\mathcal{Q} = \{Q(P) : P \in \mathcal{M}\}$ is the set of possible values for Q_0. Thirdly, one uses minimum-loss-based learning, such as super learning, fully utilizing the power and optimality results for loss-based cross-validation to select among candidate estimators, to obtain an initial estimator Q_n^0 of Q_0. Fourthly, one proposes a parametric fluctuation $Q_{g_n,n}^0(\epsilon)$, possibly indexed by the estimator g_n of nuisance parameter $g_0 = g(P_0)$, such that

$$\frac{d}{d\epsilon} L(Q_{g_n,n}^0(\epsilon))(O)\Big|_{\epsilon=0} = D^*(Q_n^0, g_n)(O), \tag{7.1}$$

where $D^*(P) = D^*(Q(P), g(P))$ is the efficient influence curve of the pathwise derivative of the statistical target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ at $P \in \mathcal{M}$. If a multivariate ϵ is used, then the derivatives with respect to each of their components ϵ_j must span the efficient influence curve $D^*(Q_n^0, g_n)$. Fifthly, one computes the amount of fluctuation with minimum-loss-based estimation:

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} L(Q_{g_n,n}^0(\epsilon))(O_i).$$

This yields an update $Q_n^1 = Q_{g_n,n}^0(\epsilon_n)$. This updating of an initial estimator Q_n^0 into a next Q_n^1 is iterated until convergence, resulting in a final update Q_n^*. Since at the last step the amount of fluctuation $\epsilon_n \approxeq 0$, this final Q_n^* will solve the efficient influence curve estimating equation:

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

representing a fundamental ingredient for establishing the asymptotic efficiency of $\Psi(Q_n^*)$. Recall that an estimator is efficient if and only if it is asymptotically linear with an influence curve equal to the efficient influence curve $D^*(Q_0, g_0)$. Finally, the TMLE of ψ_0 is the substitution estimator $\Psi(Q_n^*)$.

Thus we see that TMLE involves constructing a parametric submodel $\{Q_n^0(\epsilon) : \epsilon\}$, and thereby its corresponding fluctuation function $\epsilon \rightarrow Q_n^0(\epsilon)$, through the initial estimator Q_n^0 with parameter ϵ, where the score of this parametric submodel at $\epsilon = 0$ equals the efficient influence curve at the initial estimator. The latter constraint can be satisfied by many parametric submodels, since it represents only a local constraint of its behavior at zero fluctuation. However, it is very important that the fluctuations stay within the statistical model for the observed data distribution, even if the target parameter Ψ can be defined on fluctuations of densities that fall outside the assumed observed data model. In particular, in the context of sparse data (i.e., data that will not allow for precise estimation of the target parameter), a violation of this property can significantly affect the performance of the estimator.

One important strength of the semiparametric efficient TMLE relative to the alternative semiparametric efficient estimating equation methodology is that it respects the global constraints of the observed data model. This is due to the fact that it is a substitution estimator $\Psi(Q_n^*)$ with Q_n^*, an estimator of a relevant part Q_0 of the true distribution of the data in the observed data model. The estimating equation methodology does not result in substitution estimators and consequently often ignores important global constraints of the observed data model, which comes at a price in the context of sparsity. Indeed, simulations have confirmed this gain of TMLE relative to the efficient estimating equation method in the context of sparsity (see Chap. 20 and also Stitelman and van der Laan 2010), which is demonstrated in this chapter. However, if TMLE violates the principle of being a substitution estimator by allowing Q_n^* to fall outside the assumed observed data model, this advantage is compromised. Therefore, it is crucial that TMLE use a fluctuation function that is guaranteed to map the fluctuated initial estimator into the statistical model.

7.2.2 Procedure

To demonstrate the important consideration of selecting a fluctuation function in the construction of TMLE that corresponds with a parametric *sub*model, we consider the problem of estimating the additive causal effect of a binary treatment A on a continuous outcome Y, based on observing n i.i.d. copies of $O = (W, A, Y) \sim P_0$, where W is the set of confounders. Consider the following SCM: $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(W, A, U_Y)$ with the functions f_W, f_A, and f_Y unspecified, representing a set of assumptions about how O is generated. We assume that U_A is independent of U_Y such that the randomization assumption $(A \perp Y_a \mid W)$ holds

with respect to the counterfactuals $Y_a = f_Y(W, a, U_Y)$ as defined by this SCM. In this SCM for the data-generating distribution of the observed data O, the additive causal effect $E_0(Y_1 - Y_0)$ can be identified from the observed data distribution through the statistical parameter of P_0:

$$\Psi(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)].$$

Suppose that it is known that $Y \in [a, b]$ for some $a < b$. Alternatively, one might have truncated the original data to fall in such an interval and focus on the causal effect of treatment on this truncated outcome, motivated by the fact that estimating the conditional means of unbounded, or very heavy tailed, outcomes requires very large data sets. The SCM implies no assumptions about the statistical model $\mathcal{M}$ so that the statistical model is nonparametric. The target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ and the estimand $\psi_0 = \Psi(P_0)$ are now defined. The statistical estimation problem is to estimate ψ_0 based on observing n i.i.d. copies $O_1, \ldots, O_n$.

Let $Y^* = (Y - a)/(b - a)$ be the linearly transformed outcome within $[0, 1]$, and we define the statistical parameter

$$\Psi^*(P_0) = E_0[E_0(Y^* \mid A = 1, W) - E_0(Y^* \mid A = 0, W)],$$

which can be interpreted as the causal effect of treatment on the bounded outcome Y^* in the postulated SCM. We note the following relation between the causal effect on the original outcome Y and the causal effect on the transformed outcome Y^*:

$$\Psi(P_0) = (b - a)\Psi^*(P_0).$$

An estimate, normal limit distribution, and confidence interval for $\Psi^*(P_0)$ is now immediately mapped into an estimate, normal limit distribution, and confidence interval for $\Psi(P_0)$ by simple multiplication. Suppose $\sqrt{n}(\psi_n - \Psi^*(P_0)) \xrightarrow{d} N(0, \sigma^{2*})$, then $\sqrt{n}((b-a)\psi_n - \Psi(P_0)) \xrightarrow{d} N(0, \sigma^2)$, with $\sigma^2 = (b-a)^2\sigma^{2*}$. Upper and lower bounds on the confidence interval for $\Psi^*(P_0)$, given as (c^*_{lb}, c^*_{ub}), are multiplied by $(b - a)$ to obtain upper and lower bounds on $\Psi(P_0)$, $c_{lb} = (b-a)c^*_{lb}$, and $c_{ub} = (b-a)c^*_{ub}$. As a consequence, for notational convenience, without loss of generality, we can assume $a = 0$ and $b = 1$ so that $Y \in [0, 1]$.

To determine a loss function and corresponding fluctuation function, and thereby the definition of the TMLE, we need to know the efficient influence curve. The efficient influence curve of the statistical parameter $\Psi : \mathcal{M} \to \mathbb{R}$, defined on a nonparametric statistical model $\mathcal{M}$ for P_0 at the true distribution P_0, is given by

$$D^*(P_0) = \frac{2A - 1}{g_0(A \mid W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(Q_0), \qquad (7.2)$$

where $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ and $Q_0 = (Q_{W,0}, \bar{Q}_0)$ denotes both this conditional mean $\bar{Q}_0$ and the marginal distribution $Q_{W,0}$ of W. Note that indeed $\Psi(P_0)$ only depends on P_0 through the conditional mean $\bar{Q}_0$ and the marginal distribution of W. We will use the notation $\Psi(P_0)$ and $\Psi(Q_0)$ interchangeably. Note also that the

efficient influence curve only depends on P_0 through Q_0, g_0, so that we will also denote the efficient influence curve $D^*(P_0)$ with $D^*(Q_0, g_0)$. In order to stress that $D^*(P_0)$ can also be represented as an estimating function in ψ, we also now and then denote it by $D^*(Q_0, g_0, \psi_0)$.

We are ready to define a TMLE of $\Psi(Q_0)$, completely analogous to the TMLE presented in Chaps. 4 and 5 for a binary outcome. Let $\bar{Q}_n^0$ be an initial estimate of $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ with predicted values in $(0, 1)$. This could be a loss-based super learner based on the squared error loss function or the quasi-log-likelihood loss function presented below. In addition, we estimate $Q_{W,0}$ with the empirical distribution of $W_1, \ldots, W_n$. Let Q_n^0 denote the resulting initial estimate of Q_0. The targeting step will also require an estimate g_n of $g_0 = P_{A|W}$. As we will see, only the estimate $\bar{Q}_n^0$ of the conditional mean $\bar{Q}_0$ will be modified by the TMLE procedure defined below: this makes sense since the empirical distribution of W is already a nonparametric maximum likelihood estimator so that no bias gain with respect to the target parameter will be obtained by modifying it.

We use as fluctuation function for the empirical distribution $Q_{W,n}$, $Q_{W,n}(\epsilon_1) = (1 + \epsilon_1 D_2^*(Q_n^0))Q_{W,n}$, where $D_2^*(Q_n^0) = \bar{Q}_n^0(1, W) - \bar{Q}_n^0(0, W) - \Psi(Q_n^0)$ is the second component of the efficient influence curve $D^*(Q_n^0, g_n)$. We use the log-likelihood loss function, $-\log Q_W$, as loss function for the marginal distribution of W. It follows that

$$\frac{d}{d\epsilon} \log Q_{W,n}(\epsilon_1)\Big|_{\epsilon_1=0} = D_2^*(Q_n^0),$$

showing that this fluctuation function and log-likelihood loss function for the marginal distribution of W indeed generates the wished score at zero fluctuation.

We can represent the estimate $\bar{Q}_n^0$ as

$$\bar{Q}_n^0 = \frac{1}{1 + \exp(-f_n^0)},$$

with $f_n^0 = \log(\bar{Q}_n^0/(1 - \bar{Q}_n^0))$. Consider now the following fluctuation function:

$$\bar{Q}_n^0(\epsilon_2) = \frac{1}{1 + \exp(-\{f_n^0 + \epsilon_2 H_{g_n}^*\})},$$

which maps a fluctuation parameter value ϵ_2 into a modification $\bar{Q}_n^0(\epsilon_2)$ of the initial estimate. This fluctuation function is indexed by a function

$$H_{g_n}^*(A, W) = \frac{2A - 1}{g_n(A \mid W)}.$$

Equivalently, we can write this fluctuation function in terms of fluctuations of the logit of $\bar{Q}_n^0$: $\text{logit}\bar{Q}_n^0(\epsilon_2) = \text{logit}\bar{Q}_n^0 + \epsilon_2 H^*(g_n)$.

Consider now the following quasi-log-likelihood loss function for the conditional mean $\bar{Q}_0$:

$$-L(\bar{Q})(O) = Y \log \bar{Q}(A, W) + (1 - Y) \log(1 - \bar{Q}(A, W)).$$

Note that this is the log-likelihood of the conditional distribution of a binary outcome Y, but now extended to continuous outcomes in $[0, 1]$. It is thus known that this loss function is a valid loss function for the conditional distribution of a binary Y, but we need it to be a valid loss function for a conditional mean of a continuous $Y \in [0, 1]$. It is indeed a valid loss function for the conditional mean of a continuous outcome in $[0, 1]$, as has been previously noted. See Wedderburn (1974) and McCullagh (1983) for earlier uses of logistic regression for continuous outcomes in $[0, 1]$. We formally prove this result in Lemma 7.1 at the end of this chapter. The proposed fluctuation function $\bar{Q}_n^0(\epsilon_2)$ and the quasi-log-likelihood loss function satisfy

$$\frac{d}{d\epsilon_2} L(\bar{Q}_n^0(\epsilon_2))\Big|_{\epsilon_2=0} = H^*(A, W)(Y - \bar{Q}_n^0(A, W)),$$

giving us the desired first component $D_1^*(\bar{Q}_n^0, g_n)$ of the efficient influence curve $D^* = D_1^* + D_2^*$, where $D_2^*(Q_0) = \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(Q_0)$.

Our combined loss function is given by $L(Q) = -\log Q_W + L(\bar{Q})$, and, for $\epsilon = (\epsilon_1, \epsilon_2)$, our parametric fluctuation function for the combined Q is given by $Q(\epsilon) = (Q_W(\epsilon_1), \bar{Q}(\epsilon_2))$. With these choices of loss function $L(Q)$ for Q_0 and fluctuation function $Q(\epsilon)$ of Q, we indeed now have that

$$\frac{d}{d\epsilon_j} L(Q(\epsilon))\Big|_{\epsilon=0} = D_j^*(Q, g), \ j = 1, 2.$$

This shows that we succeeded in defining a loss function for $Q_0 = (Q_{W,0}, \bar{Q}_0)$ and fluctuation function such that the derivatives as defined in (7.1) span the efficient influence curve. The TMLE is now defined!

In this first targeting step, the maximum likelihood estimator of ϵ_1 equals zero, so that the update of $Q_{W,n}$ equals $Q_{W,n}$ itself. As a consequence of $\epsilon_{1,n}^0 = 0$ being the maximum likelihood estimator, the empirical mean of the component $D_2^*(Q_n^*) = \bar{Q}_n^*(1, W) - \bar{Q}_n^*(0, W) - \Psi(Q_n^*)$ of the efficient influence curve at the final TMLE equals zero: of course, this is trivially verified.

The maximum likelihood estimator of ϵ_2 for fluctuating $\bar{Q}_n^0$ is given by

$$\epsilon_{2n}^0 = \underset{\epsilon_2}{\operatorname{argmin}} \, P_n L(\bar{Q}_n^0(\epsilon_2)),$$

where we used the notation $P_n f = 1/n \sum_i f(O_i)$. This "maximum likelihood" estimator of ϵ_2 can be computed with generalized linear regression using the binomial link, i.e., the logistic regression maximum likelihood estimation procedure, simply ignoring that the outcome is not binary, which also corresponds with iterative reweighted least squares estimation using iteratively updated estimated weights of the form $1/(\bar{Q}_n(1 - \bar{Q}_n))$.

This provides us with the targeted update $Q_n^1 = Q_n^0(\epsilon_n^0)$, where the empirical distribution of W was not updated, but $\bar{Q}_n^0$ did get updated to $\bar{Q}_n^0(\epsilon_n^0)$. Iterating this procedure now defines the TMLE Q_n^*, but, as in the binary outcome case, we have that $\bar{Q}_n^2 = \bar{Q}_n^1(\epsilon_n^1) = \bar{Q}_n^1$ since the next maximum likelihood estimator $\epsilon_n^1 = 0$, and, of

course, the maximum likelihood estimator of ϵ_1 remains 0. Thus convergence occurs in one step, so that $Q_n^* = Q_n^1$. The TMLE of ψ_0 is thus given by $\Psi(Q_n^*) = \Psi(Q_n^1)$. As a consequence of the definition of the TMLE, we have that the TMLE Q_n^* solves the efficient influence curve estimating equation $P_n D^*(Q_n^*, g_n, \Psi(Q_n^*)) = 0$.

7.2.3 Robustness of TMLE in the Context of Sparsity

We note that, even if there is strong confounding causing some large values of $H_{g_n}^*$, the resulting TMLE $\bar{Q}_n^*$ remains bounded in $(0, 1)$, so that the TMLE $\Psi(Q_n^*)$, which just averages values of $\bar{Q}_n^*$, fully respects the global constraints of the observed data model. An inspection of the efficient influence curve (7.2), $D^*(P_0)$, reveals that there are two potential sources of sparsity. Small values for $g_0(A \mid W)$ and large outlying values of Y inflate the variance. Enforcing (e.g., known) bounds on Y and g_0 in the estimation procedure provides a means for controlling these sources of variance. We note that, even if there is strong confounding causing some large values of $h_{g_n^0}$, the resulting TMLE $\bar{Q}_n^*$ remains bounded in $(0, 1)$, so that the TMLE $\Psi(Q_n^*)$ fully respects the global constraints of the observed data model. On the other hand, the A-IPTW estimator obtained by solving the efficient influence curve estimating equation, $P_n D^*(Q_n^0, g_n, \psi) = 0$, in ψ yields the estimator

$$\psi_n = \frac{1}{n} \sum_{i=1}^{n} H_{g_n}^*(A_i, W_i)(Y_i - \bar{Q}_n^0(A_i, W_i)) + \bar{Q}_n^0(1, W) - \bar{Q}_n^0(0, W).$$

This estimator can easily fall outside $[0, 1]$ if for some observations $g_n(1 \mid W_i)$ is close to 1 or 0. This represents the price of not being a substitution estimator.

It is also important to contrast this TMLE with the TMLE using the linear fluctuation function. The latter TMLE would use the $L(\bar{Q}) = (Y - \bar{Q}(A, W))^2$ loss function, and fluctuation function $\bar{Q}_n^0(\epsilon) = \bar{Q}_n^0 + \epsilon H^*(g_n)$, so that (7.1) is still satisfied. The TMLE is defined as above, and again converges in one step. One estimates the fluctuation ϵ with univariate least squares linear regression, using $\bar{Q}_n^0$ as offset. In this case, large values of $H^*(g_n)$ will result in predicted values of $\bar{Q}_n^0(\epsilon_n)$ that are outside the bounds $[a, b]$. Therefore, this version of TMLE does not repect the global constraints of the model, i.e., the knowledge that $Y \in [a, b]$. In the next section, an analysis of a simulated data set provides a comparison of TMLE using the logistic fluctuation function and TMLE using this linear fluctuation.

7.3 Simulations

Two simulation studies illustrate the effects of employing a logistic vs. linear fluctuation function in the definition of the TMLE. These two studies evaluate practical performance with and without sparsity in the data, where a high degree of sparsity

corresponds to a target parameter that is borderline identifiable. As above, the parameter of interest is defined as the additive effect of a binary point treatment on the outcome, $\psi_0 = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$. We also implement three additional estimators: MLE, IPTW, and A-IPTW.

7.3.1 Estimators

In the simulation setting, Y is not bounded, so that we do not have an a priori a and b bound on Y. Instead of truncating Y and redefining the target parameter as the causal effect on the truncated Y, we still aim to estimate the causal effect on the original Y. Therefore, in the TMLE using a logistic fluctuation function we set $a = \min(Y)$, $b = \max(Y)$, and $Y^* = (Y - a)/(b - a)$. In this TMLE, the initial estimate $\bar{Q}_n^{0,Y^*}$ of $E_0(Y^*|A, W)$ needs to be represented as a logistic function of its logit transformation. Note that $\text{logit}(x)$ is not defined when $x = 0$ or 1. Therefore, in practice $\bar{Q}_n^{0,Y^*}$ needs to be bounded away from 0 and 1 by truncating at $(\alpha, (1 - \alpha))$ for some small $\alpha > 0$. In the reported simulations we used $\alpha = 0.005$. We also obtained results for $\alpha = 0.001$ or $\alpha = 0.01$, but no notable difference was observed.

In our simulations, we also included the A-IPTW estimator of ψ_0, defined as

$$\psi_n^{A-IPTW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{2A_i - 1}{g_n(A_i \mid W_i)}(Y_i - \bar{Q}_n^0(A_i, W_i)) + (\bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i))\right\}.$$

The two TMLEs and the A-IPTW estimator are double robust so that these estimators will be consistent for ψ_0 if either g_n or $\bar{Q}_n^0$ is consistent for g_0 and $\bar{Q}_0$, respectively. In addition, the two TMLEs and the A-IPTW estimator are asymptotically efficient if both g_n and $\bar{Q}_n^0$ consistently estimate the true g_0 and $\bar{Q}_0$, respectively.

In this simulation study we will use simple parametric maximum likelihood estimators as initial estimators $\bar{Q}_n^0$ and g_n, even though we recommend the use of super learning in practice. The goal of this simulation is to investigate the performance of the updating step under misspecified and correctly specified $\bar{Q}_n^0$, and for that purpose we can work with parametric maximum likelihood estimation fits.

We also report the MLE $\Psi(Q_n^0)$ of ψ_0 according to a parametric model for $\bar{Q}_0$, and an IPTW estimator of ψ_0 that uses g_n as estimator of g_0:

$$\psi_n^{MLE} = \frac{1}{n}\sum_{i-1}^{n}\left\{\bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i)\right\},$$

$$\psi_n^{IPTW} = \frac{1}{n}\sum_{i=1}^{n}(2A - 1)\frac{Y_i}{g_n(A_i, W_i)}.$$

The MLE of ψ_0 is included for the sake of evaluating the bias reduction step carried out by the TMLEs and the A-IPTW estimator.

7.3.2 Data-Generating Distributions

Covariates W_1, W_2, W_3 were generated as independent binary random variables: $W_1, W_2, W_3 \sim Bernoulli(0.5)$. Two treatment mechanisms were defined that differ only in the values of the coefficients for each covariate. They are of the form

$$g_0(1 \mid W) = \text{expit}(\beta W_1 + \delta W_2 + \gamma W_3).$$

We considered the following two settings for the treatment mechanism:

$$\beta_1 = 0.5, \delta_1 = 1.5, \gamma_1 = -1, \text{ and}$$
$$\beta_2 = 1.5, \delta_2 = 4.5, \gamma_2 = -3.$$

We refer to these two treatment mechanisms as $g_{0,1}$ and $g_{0,2}$, respectively. The observed outcome Y was generated as

$$Y = A + 2W_1 + 3W_2 - 4W_3 + e,\ e \sim N(0, 1).$$

For both simulations the true additive causal effect equals one: $\psi_0 = 1$. Treatment assignment probabilities based on mechanism $g_{0,1}$ range from 0.269 to 0.881, indicating no sparsity in the data for simulation 1. In contrast, treatment assignment probabilities based on mechanism $g_{0,2}$ range from (0.047 to 0.998). Simulation 2 poses a more challenging estimation problem in the context of sparse data.

Estimates were obtained for 1000 samples of size $n = 1000$ from each data-generating distribution. Treatment assignment probabilities were estimated using a correctly specified logistic regression model. In both simulations predicted values for $g_n(A \mid W)$ were bounded away from 0 and 1 by truncating at $(p, 1 - p)$, with $p = 0.01$. In one set of results a correctly specified main terms regression model was used to compute the initial estimate $\bar{Q}_n^0$, while in the other set of results the initial estimate was defined as the least squares regression Y on A only.

7.3.3 Results

Table 7.1 reports the average estimate, bias, empirical variance, and MSE for each estimator, under different specifications of the initial estimator $\bar{Q}_n^0$. In simulation 1, when $\bar{Q}_0$ is correctly estimated, all estimators perform quite well, though as expected IPTW is the least efficient. However, when $\bar{Q}_0$ is incorrectly estimated, the MLE is biased and has high variance relative to the other estimators. Since $g_n(A \mid W)$ is correctly specified, IPTW and A-IPTW provide unbiased estimates, as do both TMLEs: the TMLE_{Y^*} based on the logistic regression model is similar to the TMLE based on the linear regression model, as there is no sparsity in the data, and both are asymptotically efficient estimators.

Table 7.1 Estimator performance for simulations 1 and 2 when the initial estimator of $\bar{Q}_0$ is correctly specified and misspecified. Results are based on 1000 samples of size $n = 1000$, g_n is consistent,and bounded at (0.01,0.99)

	$\bar{Q}_0$ correctly specified				$\bar{Q}_0$ misspecified			
	ψ_n	Bias	Var	MSE	ψ_n	Bias	Var	MSE
Simulation 1								
MLE	1.003	0.003	0.005	0.005	3.075	2.075	0.030	4.336
IPTW	1.006	0.006	0.009	0.009	1.006	0.006	0.009	0.009
A-IPTW	1.003	0.003	0.005	0.005	1.005	0.005	0.010	0.010
TMLE_{Y^*}	0.993	−0.007	0.005	0.005	0.993	−0.007	0.006	0.006
TMLE	0.993	−0.007	0.005	0.005	0.993	−0.007	0.006	0.006
Simulation 2								
MLE	1.001	0.001	0.009	0.009	4.653	3.653	0.025	13.370
IPTW	1.554	0.554	0.179	0.485	1.554	0.554	0.179	0.485
A-IPTW	0.999	−0.001	0.023	0.023	1.708	0.708	0.298	0.798
TMLE_{Y^*}	0.989	−0.011	0.037	0.037	0.722	−0.278	0.214	0.291
TMLE	0.986	−0.014	0.042	0.042	−0.263	−1.263	2.581	4.173

In simulation 2, all estimators except IPTW are unbiased when $\bar{Q}_0$ is correctly estimated. In this case, both TMLEs have higher variance than A-IPTW, even though all three are asymptotically efficient. All three are more efficient than IPTW but less efficient than MLE. Though asymptotically the IPTW estimator is expected to be unbiased in this simulation, since g_n is a consistent estimator of $g_{0,2}$, these results demonstrate that in finite samples, heavily weighting a subset of observations not only increases variance but can also bias the estimate.

When the model for $\bar{Q}_0$ is misspecified in simulation 2, MLE is even more biased than it was in simulation 1. The efficiency of all three double robust efficient estimators suffers in comparison with simulation 1 as well. Nevertheless, TMLE_{Y^*}, using the logistic fluctuation, has the lowest MSE of all estimators. Its superiority over TMLE, using linear least squares regression, in terms of bias and variance is clear. TMLE_{Y^*} also outperforms A-IPTW with respect to both bias and variance and performs much better than IPTW or MLE.

7.4 Discussion

For the sake of demonstration, we considered estimation of the additive causal effect. However, the same TMLE, using the logistic fluctuation, can be used to estimate other point-treatment causal effects, including parameters of a marginal structural model. The proposed quasi-log-likelihood loss function can be used to define a super learner for prediction of a bounded continuous outcome. It will be of interest to evaluate such a super learner relative to a super learner that does not incorporate these known bounds. The quasi-log-likelihood loss function and the logistic fluctu-

ation function can also be applied in a TMLE of the causal effect of a multiple time point intervention in which the final outcome is bounded and continuous. In this case, one uses the loss function and logistic fluctuation function to fluctuate the last factor of the likelihood of the longitudinal structure. Our simulations show that the proposed fluctuation function and loss function, and corresponding TMLEs, should also be used for continuous outcomes for which no a priori bounds are known. In this case, one simply uses the minimal and maximal observed outcome values. In this way, these choices naturally robustify the TMLEs by enforcing that the updated initial estimator will not predict outcomes outside the observed range. TMLE using the logistic fluctuation function can also be incorporated in C-TMLE (Chaps. 19–21 and 23) without modification.

Appendix

The following lemma proves that the quasi-log-likelihood loss function is indeed a valid loss function for the conditional mean $\bar{Q}_0$ of a continuous outcome in $[0, 1]$.

Lemma 7.1. *We have that*

$$\bar{Q}_0 = \underset{\bar{Q}}{\operatorname{argmin}}\, E_0 L(\bar{Q}),$$

where the minimum is taken over all functions of (A, W) *that map into* $[0, 1]$. *In addition, given a function* H^**, define the fluctuation function*

$$logit(\bar{Q}(\epsilon)) = logit(\bar{Q}) + \epsilon H^*.$$

For any function H^* *we have*

$$\frac{d}{d\epsilon} L(\bar{Q}(\epsilon))\Big|_{\epsilon=0} = H^*(A, W)(Y - \bar{Q}(A, W)).$$

Proof. Let $\bar{Q}_1$ be a local minimum of $\bar{Q} \to E_0 L(\bar{Q})(O)$, and consider the fluctuation function $\epsilon \to \bar{Q}_1(\epsilon)$ defined above. Then the derivative of $\epsilon \to E_0 L(\bar{Q}_1(\epsilon))$ at $\epsilon = 0$ equals zero. However, we also have

$$-\frac{d}{d\epsilon} L(\bar{Q}_1(\epsilon))\Big|_{\epsilon=0} = H^*(A, W)(Y - \bar{Q}_1(A, W)).$$

Thus, it follows that

$$E_0[H^*(A, W)(Y - \bar{Q}_1(A, W))] = E_0[H^*(A, W)(\bar{Q}_0 - \bar{Q}_1)(A, W)].$$

But this needs to hold for any function $H^*(A, W)$, which proves that $\bar{Q}_1 = \bar{Q}_0$ almost everywhere. The final statement follows as well. □