The International Journal of Biostatistics

Manuscript 1308

The Relative Performance of Targeted Maximum Likelihood Estimators

Kristin E. Porter, *University of California, Berkeley*
Susan Gruber, *University of California, Berkeley*
Mark J. van der Laan, *University of California, Berkeley*
Jasjeet S. Sekhon, *University of California, Berkeley*

The Relative Performance of Targeted Maximum Likelihood Estimators

Kristin E. Porter, Susan Gruber, Mark J. van der Laan, and Jasjeet S. Sekhon

Abstract

There is an active debate in the literature on censored data about the relative performance of model based maximum likelihood estimators, IPCW-estimators, and a variety of double robust semiparametric efficient estimators. Kang and Schafer (2007) demonstrate the fragility of double robust and IPCW-estimators in a simulation study with positivity violations. They focus on a simple missing data problem with covariates where one desires to estimate the mean of an outcome that is subject to missingness. Responses by Robins, et al. (2007), Tsiatis and Davidian (2007), Tan (2007) and Ridgeway and McCaffrey (2007) further explore the challenges faced by double robust estimators and offer suggestions for improving their stability. In this article, we join the debate by presenting targeted maximum likelihood estimators (TMLEs). We demonstrate that TMLEs that guarantee that the parametric submodel employed by the TMLE procedure respects the global bounds on the continuous outcomes, are especially suitable for dealing with positivity violations because in addition to being double robust and semiparametric efficient, they are substitution estimators. We demonstrate the practical performance of TMLEs relative to other estimators in the simulations designed by Kang and Schafer (2007) and in modified simulations with even greater estimation challenges.

KEYWORDS: censored data, collaborative double robustness, collaborative targeted maximum likelihood estimation, double robust, estimator selection, inverse probability of censoring weighting, locally efficient estimation, maximum likelihood estimation, semiparametric model, targeted maximum likelihood estimation, targeted minimum loss based estimation, targeted nuisance parameter estimator selection

Author Notes: Kristin E. Porter and Susan Gruber contributed equally to this work. We would like to thank Zhiqiang Tan and Weihua Cao for providing us with software to implement their estimators. This work was supported by NIH grant R01AI074345-04 and by a National Institutes of Health NRSA Trainee appointment on grant number T32 HG 00047.

1 Introduction

The translation of a scientific question into a statistical estimation problem often involves the formulation of a full-data structure, a target parameter of the full-data probability distribution representing the scientific question of interest, and an observed data structure which can be viewed as a mapping on the full data structure and a censoring variable. One must identify the target parameter of the full-data distribution from the probability distribution of the observed data structure, which often requires particular modeling assumptions such as the coarsening at random assumption on the censoring mechanism (i.e., the conditional distribution of censoring, given the full-data structure). The statistical problem is then reduced to a pure estimation problem defined by the challenge of constructing an estimator of the estimand, defined by the identifiability result for the target parameter of the full-data distribution. The estimator should respect the statistical model implied by the posed assumptions on the censoring mechanism and the full-data distribution.

For semiparametric (e.g., nonparametric) statistical models, many estimators rely in one way or another on the inverse probability of censoring weights (IPCW). Such estimators can be biased and highly variable under practical or theoretical violations of the positivity assumption, which is a support condition on the censoring mechanism that is necessary to establish the identifiability of the target parameter (e.g., Robins (1986, 1987, 1999); Neugebauer and van der Laan (2005); Petersen et al. (2010)). A particular class of estimators are so called double robust estimators (e.g., van der Laan and Robins (2003)). Double robust (DR) estimators, which rely on both IPCW and a model of the full-data distribution, are not necessarily protected from the bias or inflated variance that can result from positivity violations, and in recent literature, there is much debate on the relative performance of DR estimators when the positivity assumption is violated. In particular, Kang and Schafer (2007) (KS) demonstrate the fragility of DR estimators in a simulation study with near, or practical, positivity violations. They focus on a simple missing data problem in which one wishes to estimate the mean of an outcome that is subject to missingness and all possible covariates for predicting missingness are measured. Responses by Robins et al. (2007), Tsiatis and Davidian (2007), Tan (2007) and Ridgeway and McCaffrey (2007) further explore the challenges faced by DR estimators and offer suggestions for improving their stability.

Under regularity conditions, DR estimators are asymptotically unbiased if either the model of the conditional expectation of the outcome given the covariates or the model of the conditional probability of missingness given the covariates is consistent. DR estimators are semiparametric efficient (for the nonparametric model for the full-data distribution) if both of these estimators are consistent. In their article, KS introduce a variety of DR estimators and compare them to non-DR IPCW estimators as well as a simple parametric model based ordinary least squares (OLS) estimator. As the KS simulation has practical positivity violations, some values of both the true and estimated missingness mechanism are very close to zero. In this situation, the IPCW will be extremely large for some observations of the sample. Therefore, DR and non-DR estimators that rely on IPCW may be unreliable. As a result, KS warn against the routine use of estimators that rely on IPCW, including DR estimators. This is in agreement with other literature analyzing the issue. For an overview of the issue, for example, see Robins (1986,1987, 1999); Robins and Wang (2000); van der Laan and Robins (2003)). For literature showing simulations demonstrating the extreme sparsity bias of IPCW-estimators, see for example, Neugebauer and van der Laan (2005)). Also, Petersen et al. (2010); Wang et al. (2006a); Moore et al. (2009); Cole and Hernan (2008); Kish (1992); Bembom and van der Laan (2008)) have focused on diagnosing violations of the positivity assumptions in response to this concern. Bembom and van der Laan (2008) presented data adaptive selection of the truncation constant to control the influence of weighting. In addition, van der Laan and Petersen (2007) and Petersen et al. (2010) discussed selecting parameters that are relying on realistic assumptions.

The particular simulation in KS also gives rise to a situation in which under dual misspecification, the OLS estimator outperforms all of the presented DR estimators. While this is an interesting issue, it is not the main focus of this article. In our view, dual misspecification brings up the need for other strategies for improving the robustness of estimators in general, such as incorporating data adaptive estimation instead of relying on parametric regression models for the missingness mechanism and the conditional distribution of responses, an idea echoed in the responses by Tsiatis and Davidian (2007) and Ridgeway and McCaffrey (2007), and standardly incorporated in the UC Berkeley literature on targeted maximum likelihood estimation (e.g., van der Laan and Rubin (2006); van der Laan et al. (2009)). In particular, we note that a statistical estimation problem is also defined by the statistical model, which, in this case, is defined by a nonparametric model: such models require data adaptive estimators in order to claim that the estimator is consistent. Nonetheless, we explicitly demonstrate the impact of the utilization of machine learning on the simulation results in a final section of this article.

In their response to the KS paper, Robins et al. (2007) point out that a desirable property of DR estimators is "boundedness," in that for a finite sample, estimators of the mean response fall in the parameter space with probability 1. Estimators that impose such a restriction can introduce new bias but avoid the challenges of highly variable weights. Robins et al. (2007) discuss ways in which to guarantee that "boundedness" holds and present two classes of bounded estimators–regression DR estimators and bounded Horvitz-Thompson DR estimators. We define examples of these estimators below, and we evaluate their relative performance. The response by Tsiatis and Davidian (2007) offers strategies for constructing estimators that are more robust under the circumstances in the KS simulations. In particular, to address positivity violations, they suggest an estimator that uses IPCW only for observations with missingness mechanism values that are not close to zero, while using regression predictions for the observations with very small missingness mechanism values. One might consider either a hard cutoff for dividing observations or weighting each part of the influence curve by the estimated missingness mechanism. Tan (2007) also points to an improved locally efficient double robust estimator (Tan (2006)) that is able to maintain double robustness as well as provides guaranteed improvement relative to an initial estimator, improving on such type of estimators that had an algebraic similar form but failed to guarantee both properties (Robins et al. (1994), and see also van der Laan and Robins (2003)). Many responders also make valuable suggestions regarding the dual misspecification challenge.

In the current paper, adapted in part from Sekhon et al. (2011), we add targeted maximum likelihood estimators (TMLEs), or more generally, targeted minimum loss based estimators (van der Laan and Rubin (2006)) to the debate on the relative performance of DR estimators under practical violations of the positivity assumption in the particular simple missing data problem set forth by KS. TMLEs involve a two-step procedure in which one first estimates the conditional expectation of the outcome, given the covariates, and then updates this initial estimator, targeting the parameter of interest, rather than the overall conditional mean of the outcome given the covariates. The second step requires specification of a loss-function (e.g., log-likelihood loss function) and a parametric submodel through the initial regression, so that one can fit the parametric sub-model by minimizing the empirical risk (e.g., maximizing the log-likelihood). The estimator of the target parameter is then defined as the corresponding substitution estimator. Because TMLEs are substitution estimators, they not only respect the global bounds of the parameter and data (and thus satisfy the "boundedness" property defined by Robins et al. (2007)), but, even more importantly, they respect the fact that the true parameter value is a particular function of the data generating probability distribution.

TMLEs are double robust and asymptotically efficient. Moreover, TMLEs can incorporate data-adaptive likelihood or loss based estimation procedures to estimate both the conditional expectation of the outcome and the missingness mechanism.The TMLE also allows the incorporation of targeted estimation of the censoring/treatment mechanism, as embodied by the collaborative TMLE (C-TMLE), thereby fully confronting a long standing problem of how to select covariates in the propensity score/missingness mechanism of DR-estimators. In this article, we compare the performance of TMLEs to other DR estimators in the literature using the exact simulation study presented in the KS paper. We also make slight modifications to the KS simulation, in order to make the estimation even more challenging.

The remainder of this article is organized as follows. Section 2 presents notation, which deviates from that presented in KS, for the data structure and parameter of interest. Section 3 formally defines the positivity assumption and gives an overview of causes, diagnostics and responses to violations. Section 4 defines the estimators on which we focus in this paper, including a sample of estimators in the literature and TMLEs. Section 5 compares estimator performance in the original and modified KS simulations. Section 6 then looks at coupling TMLEs with machine learning. Section 7 concludes with a discussion of the findings.

2 Data Structure, Statistical Model, and Parameter of Interest

Consider an observed data set consisting of n independent and identically distributed (i.i.d) observations of $O = (W, \Delta, \Delta Y) \sim P_0$. W is a vector of covariates, and $\Delta = 1$ indicates whether Y, a continuous outcome, is observed. P_0 denotes the true distribution of O, from which all observations are sampled. We view O as a missing data structure on a hypothetical full data structure $X = (W, Y)$, which contains the true, or potential, value of Y for all observations, as if no values are missing. We assume Y is missing at random (MAR) such that $P_0(\Delta = 1 \mid X) = g_0(1 \mid W)$. In other words, we assume there are no unobserved confounders of the relationship between missingness Δ and the outcome Y.

We define $Q_0 = \{Q_{0,W}, \bar{Q}_0\}$, where $Q_{0,W}(w) \equiv P_0(W = w)$ and $\bar{Q}_0(W) \equiv E_0(Y \mid \Delta = 1, W)$. We make no assumptions about Q_0. The generalized Cramer-Rao information bound for any parameter of Q_0 does not depend on the statistical model for the missingness mechanism g_0. The parameter of interest is the mean outcome $E_0(Y)$ for the sampled population, as if there were not missing observations of Y. Due to the MAR assumption and the positivity assumption defined below, our target parameter is identified from P_0 by the following mapping from Q_0:

$$\mu(P_0) = E_0(Y) = E_0(\bar{Q}_0(W)).$$

3 The Positivity Assumption

The identifiability of the parameter of interest $\mu(P_0)$ requires MAR and adequate support in the data. Regarding the latter, it requires that within each stratum of W, there is positive probability that Y is not missing. This requirement is often referred to as the positivity assumption. Formally, for our target parameter, the positivity assumption requires that:

$$g_0(\Delta = 1 \mid W) > 0 \ P_0\text{-almost everywhere.} \tag{1}$$

The positivity assumption is specific to the the target parameter. For example, the positivity assumption of the target parameter $E_0\{E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)\}$ of the probability distribution of $O = (W, A, Y)$, representing the additive causal effect under causal assumptions, requires that within each stratum there is a positive probability for all possible treatment assignments. For example, if A is a binary treatment, then positivity requires that $0 < g_0(A = 1 \mid W) < 1$. (The assumption is often referred to as the experimental treatment assignment (ETA) assumption for causal parameters.) In addition to being parameter-specific, the positivity assumption is also model-specific. Parametric model assumptions, which extrapolate to regions of the joint distribution of (A,W) that may not be supported in the data, allow for weakening the positivity assumption (Petersen et al. (2010)). However, analysts need to be sure that their parametric assumptions actually hold true, which may be difficult if not impossible.

Violations and near violations of the positivity assumption can arise for two reasons. First, it may be theoretically impossible or highly unlikely for the outcome Y to be observed for certain covariate values in the population of interest. The threat to identifiability due to such structural violations of positivity exists regardless of the sample size. Second, given a finite sample, the probability of the outcome being observed for some covariate values might be so small that the observed sample cannot be distinguished from a sample drawn under a theoretical violation of the positivity assumption. The effect of such practical violations of the positivity assumption are sample size specific, and the resulting sparse data bias and inflated variance are often as dramatic as under structural violations.

Several approaches for diagnosing bias due to positivity violations have been suggested (see Petersen et al. (2010) for an overview). Analysts may assess the distribution of Δ within covariate strata (or in the case of causal parameters, the distribution of treatment assignment), but this method is not practical with high dimensional covariate sets or with continuous or multi-level covariates, and also provides no quantitative measure of the resulting sparse-data bias. Analysts may also assess the distribution of the estimated missingness mechanism scores, $g_n(\Delta = 1 \mid W)$, or inverse probability weights. While this approach may indicate positivity violations, it does not provide any information on the extent of potential bias of the chosen estimator. Wang et al. (2006b) introduce and Petersen et al. (2010) further discuss a diagnostic that provides an estimate of positivity bias for any candidate estimator, which is based on a parametric bootstrap. Bias estimates of similar or larger magnitude than an estimate's standard error can raise a red flag to analysts that inference for their target parameter is threatened by lack of positivity.

When censoring probabilities are close to 0 (or 1 in the case of an effect parameter), a common practice is to truncate the probabilities or the resulting inverse probability weights, either at fixed levels or at percentiles (Petersen et al. (2010); Wang et al. (2006a); Moore et al. (2009); Cole and Hernan (2008); Kish (1992); Bembom and van der Laan (2008)). The practice limits the influence of observations with large unbounded weights, which may reduce positivity bias and rein in inflated variance. However, this practice may also introduce bias, due to misspecification of the missingness mechanism g_n. The extent to which truncating g_n hurts or helps the performance of an estimator depends on the level of truncation, the estimator and the distribution of the data. In our simulations below, we examine the effect of truncating missingness probabilities for all estimators that we introduce in the next section.

4 Estimators of a Mean Outcome when the Outcome is Subject to Missingness

4.1 Estimators in the Literature

As a benchmark, KS compare all estimators in their paper to the ordinary least squares (OLS) estimator. For the target parameter, the OLS estimator is equivalent to the G-computation estimator based on a linear regression model. It is defined as:

$$\mu_{n,OLS} = \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n^0(W_i).$$

where $\bar{Q}_n^0 = m_{\beta_n}$ is a linear regression initial fit of $\bar{Q}_0$, and β_n is given by:

$$\beta_n = \arg\min_{\beta} \sum_{i=1}^{n} \Delta_i (Y_i - m_\beta(W_i))^2.$$

(Note that in our notation, the subscript $_n$ refers to an estimation, and the superscript indicates whether the estimation is from an initial fit $\binom{0}{n}$, or as we introduce below, a refit $\binom{\prime}{n}$ or a fluctuated fit $\binom{*}{n}$.) Under violation of the positivity assumption, the OLS estimator, when defined, extrapolates from strata of W in which there is support to strata of W that lack adequate support. The extrapolation depends on the validity of the linear regression model, and misspecification leads to bias.

KS present comparisons of several DR (and non-DR) estimators. We focus on just a couple of them here. Using our terminology with the terminology and abbreviations from KS in parenthesis the estimators we compare are: the weighted least squares (WLS) estimator (regression estimation with inverse-propensity weighted coefficients, $\mu_{n,WLS}$) and the augmented IPCW (A-IPCW) estimator (regression estimation with residual bias correction, $\mu_{n,BC-OLS}$). Both of these DR estimators are defined below.

The WLS estimator is defined as:

$$\mu_{n,WLS} = \frac{1}{n}\sum_{i=1}^{n} m_{\beta_n}(W_i),$$

where

$$\beta_n = \arg\min_{\beta}\sum_{i=1}^{n}\frac{\Delta_i}{g_n(1 \mid W_i)}(Y_i - m_\beta(W_i))^2.$$

The A-IPCW estimator, introduced by J.M. Robins and Zhao (1994), is then defined as:

$$\mu_{n,A-IPCW} = \bar{Q}_n^0(W_i) + \frac{1}{n}\sum_{o=1}^{n}\frac{\Delta_i}{g_n(1 \mid W_i)}(Y_i - \bar{Q}_n^0(W_i)).$$

Both of these estimators rely on estimators of $\bar{Q}_0$ and g_0. They are consistent if $\bar{Q}_n^0$ or g_n is consistent, and efficient if both are consistent. Under positivity violations, however, these estimators rely on the consistency of $\bar{Q}_n^0$, and require that g_n converges to a limit that satisfies the positivity assumption (see e.g., van der Laan and Robins (2003)).

Additionally, in comments on KS, Robins et al. (2007) introduce bounded Horvitz-Thompson (BHT) estimators, which, as the name suggests, are bounded, in that for finite sample sizes the estimates are guaranteed to fall in the parameter space. A BHT estimator is defined as:

$$\mu_{n,BHT} = \bar{Q}_n^0(W) + \frac{1}{n}\sum_{i}\frac{\Delta_i}{g_{n_{EXT}}(1 \mid W_i)}(Y_i - \bar{Q}_n^0(W_i)).$$

This is equivalent to the A-IPTW estimator, but estimating $g_0(1 \mid W)$ by fitting the following logistic regression model:

$$\text{logit} P_{EXT}(\Delta = 1 \mid W) = \alpha^T W + \phi h_n(W),$$

and $h_n(W) = \bar{Q}_n^0(W) - \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n^0(W_i)$.

We also include another important class of doubly robust, locally efficient, regression-based estimators introduced by Scharfstein et al. (1999), further discussed in Robins (1999) and compared to the TMLE in Rosenblum and van der Laan (2010). This estimator is based on a parametric regression model, which includes a "clever covariate" that incorporates inverse probability weights. The estimator behaves similarly to the TMLE using a linear fluctuation (and is identical if the TMLE using a linear fluctuation uses this clever parametric regression as initial estimator). We use the abbreviation PRC. The estimator is defined as:

$$\mu_{n,PRC} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}'_n(W_i),$$

where $Q'_n(W) = m_{\beta_n,\epsilon_n}(W)$ and $m_{\beta,\epsilon}(W)$ is a parametric model, which includes the clever covariate $H^*_{g_n}(W) = \frac{1}{g_n(1|W)}$, and (β_n, ϵ_n) is the OLS.

Cao et al. (2009) presents a DR estimator that achieves minimum variance among a class of DR estimators indexed by all possible linear regressions for the initial estimator, when the estimator of missingness mechanism is correctly specified (see also Rubin and van der Laan (2008) for empirical efficiency maximization), while it preserves the double robustness. They also address the effect of large IPCW by enhancing the missingness mechanism estimator in order to constrain the predicted values. Their estimator is defined as:

$$\mu_{n,Cao} = \sum_{i=1}^{n} \frac{\Delta_i Y_i}{g_n(1 \mid W_i)} - \frac{\Delta_i - g_n(1 \mid W_i)}{g_n(1 \mid W_i)} m(W_i, \beta_n).$$

Cao's enhanced missingness mechanism estimator is given by:

$$g_n(1 \mid W) = \pi^{en}(W, \delta_n, \gamma_n) = 1 - \frac{exp(\delta_n + \tilde{W}\gamma_n)}{1 + exp(\tilde{W}\gamma_n)}.$$

Here $\tilde{W} = [1, W]$, and the parameters γ and δ are estimated subject to the constraints $0 < \pi(W, \delta, \gamma) < 1$ and $\sum_{i=1}^{n} \Delta_i/\pi^{en}(W_i, \delta_n, \gamma_n) = n$. A quasi-Newton method implemented in the *constrOptim.nl* function in the R package *alabama* was used to estimate (δ_n, γ_n) (Varadhan, 2010). We used OLS to estimate β_n, which corresponds to Cao's $\hat{\mu}^{en}_{usual}$.

Tan (2010) presents an augmented likelihood estimator that is a more robust version of estimators originally introduced in Tan (2006) that respect boundedness and is semi-parametric efficient. This estimator is defined as:

$$\mu_{n,Tan} = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i Y_i}{\omega(W;\tilde{\lambda}_{step2})},$$

where $\omega(W;\tilde{\lambda}_{step2})$ is an enhanced estimate of the missingness mechanism based on an initial estimate, $\pi_{ML}(W)$. Specifically, $\omega(W;\lambda) = \pi_{ML}(W) + \lambda^T h_n(W)$, where $h_n = (h_{n,1}^T, h_{n,2}^T)$,

$$\begin{aligned} h_{n,1} &= (1 - \pi_{n,ML}(W))\nu_n(W), \\ h_{n,2} &= \frac{\partial \pi}{\partial \gamma_{n,ML}}(W;\gamma_{n,ML}), \\ \nu_n(W) &= [1, \bar{Q}_n^0(W)]^T, \end{aligned}$$

and $\gamma_{n,ML}$ is a maximum likelihood estimator for the propensity score model parameter. An estimate λ_n that respects the constraint $0 < \omega(W_i, \lambda)$ if $\Delta_i = 1$ can be obtained using a two-step procedure outlined in Tan's article. Following Tan's recommendation, non-linear optimization was carried out using the R *trust* package (Geyer, 2009). We consider the two variants of Tan's LIK2 augmented likelihood estimator that performed best in Tan's simulations under misspecification of Q. Our estimator TanWLS relies on a weighted least squares estimate of $\bar{Q}_n^0$. TanRV relies on the empirical efficiency maximization estimator of Rubin and van der Laan (Rubin and van der Laan, 2008),

$$\begin{aligned} \bar{Q}_{n,RV} &= \sum_{i=1}^{n} \frac{\Delta_i}{g(1 \mid W_i)}(Y_i - m(W;\beta_n)) + m(W;\beta_n), \\ \beta_n &= \arg\min_{\beta} \sum_{i=1}^{n} \frac{\Delta_i(1 - g_n(1 \mid W_i))}{g_n(1 \mid W_i)^2}(Y_i - m_\beta(W_i))^2. \end{aligned}$$

4.2 TMLEs

The targeted maximum likelihood procedure was first introduced in van der Laan and Rubin (2006). For a compilation of current and past work on targeted maximum likelihood estimation, see van der Laan et al. (2009).

In contrast to the estimating equation-based DR estimators defined above (WLS, A-IPCW, BHT, Cao, and Tan), the PRC estimator and TMLEs are DR *substitution* estimators. TMLEs are based on an update of an initial estimator of P_0 that fluctuates the fit with a fit of a clever parametric submodel. Assuming a valid parametric submodel is selected, TMLEs do not only respect the bounds on the outcome implied by the statistical model or data, but also respect that the true target parameter value is a specified function of the data generating distribution. Due to respecting this information, the TMLE does not only respect the local bounds of the statistical model by being asymptotically (locally) efficient (as the other DR estimators), but also respect the global constraints of the statistical model. Being a substitution estimator is particularly important under sparsity, as implied by violations of the positivity assumption.

Although our target parameter involves a continuous Y, to introduce the TMLE for the mean outcome, we begin by defining the TMLE for a binary Y. In this case, the TMLE is defined as:

$$\mu_{n,TMLE} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(W_i), \tag{2}$$

where we use the logistic regression submodel:

$$\text{logit}\bar{Q}_n^0(\epsilon) = \text{logit}\bar{Q}_n^0 + \epsilon H_{g_n}^*,$$

the clever covariate is defined as $H_{g_n}^*(W) = \frac{1}{g_n(1|W)}$, and ϵ, the fluctuation parameter, is estimated by maximum likelihood in which the loss function is thus the log-likelihood loss function:

$$-L(\bar{Q})(O) = \Delta\left\{Y\log\bar{Q}(W) + (1-Y)\log(1-\bar{Q}(W))\right\}. \tag{3}$$

Thus ϵ_n is fitted with univariate logistic regression, using the initial regression estimator $\bar{Q}_n^0$ as an off-set:

$$\epsilon_n = \arg\min_{\epsilon}\sum_{i=1}^{n} L(\bar{Q}_n^0(\epsilon))(O_i).$$

The TMLE of $\bar{Q}_0$ is defined as $\bar{Q}_n^* = \bar{Q}_n(\epsilon_n)$, and $\mu(Q_n^*)$ is the corresponding TMLE of μ_0.

For estimators $\bar{Q}_n^0$ and g_n, one may specify a parametric model or use machine learning or even super learner, which uses loss-based cross-validation to select weighted combination of candidate estimators (van der Laan et al. (2007)).

Next, consider that Y is continuous, but bounded by 0 and 1. In this case, we can implement the same TMLE as we would for binary Y in (2). That is, we use the same logistic regression submodel, and the same loss function (3), and the same standard software for logistic regression to fit ϵ, simply ignoring that Y is not binary. The same loss function is still valid for the conditional mean $\bar{Q}_0$ (Wedderburn (1974); Gruber and van der Laan (2010a)):

$$\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(\bar{Q}).$$

Finally, given a continuous $Y \in [a, b]$ we can define $Y^* = (Y - a)/(b - a)$ so that $Y^* \in [0, 1]$. Then, let $\mu^*(P_0) = E_0(E_0(Y^* \mid \Delta = 1, W))$. This approach requires setting a range $[a, b]$ for the outcomes Y. If such knowledge is available, one simply uses the known values. If Y would not be subject to missingness, then one would use the minimum and maximum of the empirical sample which represents a very accurate estimator of the range. In these simulations, Y is subject to informative missingness, so that the minimum or maximum of the biased sample represents a biased estimate of the range, resulting in a small unnecessary bias in the TMLE (asymptotically negligible relative to MSE). We enlarged the range of the complete observations on Y by setting a to 0.9 times the minimum of the observed values, and b to 1.1 times the maximum of the observed values, which seemed to remove most of the unnecessary bias. We expect that some improvements can be obtained by incorporating a valid estimator of the range that takes into account the informative missingness, but such second order improvements are outside the scope of this article. We now compute the above TMLE of $\mu^*(P_0)$, denoted as TMLE_{Y*}, and we use the relation $\mu(P_0) = (b - a)\mu^*(P_0) + a$.

We note that the estimator proposed by (Scharfstein et al., 1999) and discussed in the KS debate is a particular special case of a TMLE (Rosenblum and van der Laan (2010)). It defines a clever parametric initial regression for which the update step of the general TMLE-algorithm introduced in van der Laan and Rubin (2006) results in a zero-update, and is thus not needed. Such a TMLE falls in the class of TMLEs defined by an initial regression estimator, a squared error loss function and univariate linear regression sub-model (coding the fluctuations of the initial regression estimator for the TMLE-update step). Such TMLEs for continuous outcomes (contrary to the excellent robustness of the TMLE for binary outcome based on the log-likelihood loss function and logistic regression submodel) suffer from great sensitivity to violations of the positivity assumptions, as was also observed in the simulations presented in the Kang and Schafer debate. As explained in (Gruber and van der Laan (2010a)) the problem with this TMLE defined by the squared error loss function and univariate linear regression submodel is that its updates are not subject to any bounds implied by the statistical model or data: that is, it is not using a parametric *sub*-model, an important principle of the general TMLE algorithm. The valid TMLE for continuous outcomes above, defined by the quasi-binary-log-likelihood loss and a univariate logistic regression parametric submodel, was recently presented (Gruber and van der Laan (2010a)), and in the latter article it was demonstrated that the previously observed sensitivity of these two estimators to the positivity assumption was due to those specific choices.

Finally, a natural extension of all of the above TMLEs is to make a more sophisticated estimate of g_0. Therefore, estimator $\mu_{n,C-TMLE_{Y*}}$ is defined by (2) as well, but the algorithm for computing Q_n^* differs. For the C-TMLE, we generate a sequence of nested-logistic regression model fits of g_0, $g_{n,1}, \ldots, g_{n,K}$, and we create a corresponding sequence of candidate TMLEs $Q^*_{k,g_{n,k}}$, using $g_{n,k}$ in the targeted MLE step, $k = 1, \ldots, K$, such that the loss-function (e.g., log-likelihood) specific fit of $Q^*_{k,g_{n,k}}$ is increasing in k. Finally, we use loss-function specific cross-validation to select k. The precise algorithm is presented in Gruber and van der Laan (2010b) and the software is available, and posted on `http://www.stat.berkeley.edu/~laan`. As a result, the resulting estimator g_n used in the TMLE is aimed to only include covariates that are effective in removing bias with respect to the target parameter: the theoretical underpinnings in terms of collaborative double robustness of the efficient influence curve is presented in van der Laan and Gruber (2009).

5 Simulation Studies

In this section, we compare the performance of TMLEs to the estimating equation-based DR estimators (WLS, A-IPTW, BHT, Cao, TanWLS, TanRV) as well as PRC and OLS, in the context of positivity violations. The goal of the original simulation designed by KS was to highlight the stability problems of DR estimators. We explore the relative performance of the estimators under the original KS simulation and a number of alternative data generating distributions that involve stronger and different types of violations of the positivity assumption. These new simulation settings were designed to provide more diverse and even more challenging test cases for evaluating robustness and thereby finite sample performance of the different estimators.

For the four simulations described below, all estimators were used to estimate $\mu(P_0)$ from 250 samples of size 1000. We include TMLE$_{Y*}$ and C-TMLE$_{Y*}$ estimators based on the quasi-log-likelihood loss function and the logistic regression submodel. We evaluated the performance of the estimators by their bias, variance and mean squared error (MSE).

We compared the estimators of $\mu(P_0)$ using different specifications of the estimators of $\bar{Q}_0$ and g_0. In each of the tables presented below, "Qcgc" indicates that the estimators of both were specified correctly; "Qcgm" indicates that the estimator of $\bar{Q}_0$ was correctly specified, but the estimator of g_0 was misspecified ; "Qmgc" indicates that the estimator of $\bar{Q}_0$ was misspecified, but the estimator of g_0 was correctly specified; and "Qmgm" indicates that both estimators were misspecified. For the modified simulations we present results for the "Qmgc" specification only, in order to focus on the performance of each estimator when reliance on g_n is essential. Additional results for the other model specifications are available as supplemental materials.

Also, for all estimators, we compared results with no lower bound on $g_n(1 \mid W)$ with truncating $g_n(1 \mid W)$ at a lower bound set at 0.025. We note that neither KS nor Robins et al. (2007) included bounding $g_n(1 \mid W)$ when applying their estimators. Although, not bounding $g_n(1, W)$ has the advantage that in any given application it is difficult to determine which bounds to use, the theory teaches us that the DR estimators can only be consistent if g_n is bounded from below, even if in truth g_0 is unbounded. In addition, some of the estimators above incorporate implicit bounding of g_n, so that such estimators would appear to be particularly advantageous, while the gain in performance might all be due to the implicit bounding of g_n

(which would be good to know). Additional results when g_n is bounded from below at 0.01 and 0.05 demonstrate similar behavior, and are also available on the web.

5.1 Kang and Schafer Simulation

Kang and Schafer (2007) consider n i.i.d. units of $O = (W, \Delta, \Delta Y) \sim P_0$, where W is a vector of 4 baseline covariates, and Δ is an indicator of whether the continuous outcome, Y, is observed. Kang and Schafer are interested in estimating the following parameter:

$$\mu(P_0) = E_0(Y) = E_0(E_0(Y \mid \Delta = 1, W)).$$

Let $(Z_1, \ldots, Z_4)$ be independent normally distributed random variables with

mean zero and variance 1. The covariates W we actually observe are generated as follows:

$$\begin{aligned} W_1 &= exp(Z_1/2) \\ W_2 &= Z_2/(1 + exp(Z_1)) + 10 \\ W_3 &= (Z_1 Z_3/25 + 0.6)^3 \\ W_4 &= (Z_2 + Z_4 + 20)^2. \end{aligned}$$

The outcome Y is generated as:

$$Y = 210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4 + N(0, 1).$$

From this, one can determine that the conditional mean $\bar{Q}_0(W)$ of Y, given W, which equals the same linear regression in $Z_1(W), \ldots, Z_4(W)$, where $Z_j(W)$, $j = 1, \ldots, 4$, are the unique solutions of the 4 equations above in terms of $W = (W_1, \ldots, W_4)$. Thus, if the data analyst would have been provided the functions $Z_j(W)$, then the true regression function is linear in these functions, but the data analyst is measuring the terms W_j instead. The other complication of the data generating distribution is that Y is subject to missingness, and the true censoring mechanism, denoted by $g_0(1 \mid W) \equiv P_0(\Delta = 1 \mid W)$, is given by:

$$g_0(1 \mid W) = \text{expit}(-Z_1(W) + 0.5Z_2(W) - 0.25Z_3(W) - 0.1Z_4(W)).$$

With this data generating mechanism, the average response rate is 0.50. Also, the true population mean is 210, while the mean among respondents is 200. These values indicate a small selection bias.

In these simulations, a linear main term model in the main terms $(W_1, \ldots, W_4)$ for either the outcome-regression or missingness mechanism is misspecified, while a linear main term model in the main terms $(Z_1(W), \ldots, Z_4(W))$ would be correctly specified. Note that in the KS simulation, there are finite sample violations of the positivity assumption. Specifically, we find $g_0(\Delta = 1 \mid W) \in [0.01, 0.98]$ and the estimated missingness probabilities $g_n(\Delta = 1 \mid W)$ were observed to fall in the range $[4 \times 10^{-6}, 0.97]$.

Figure 1 and Table 1 present the simulation results without any bounding of g_n. Tan's estimator imposes internal bounds on the estimated missingness mechanism, however we report performance of TanWLS and TanRV estimators when given an initial estimate g_n that is not bounded away from 0. All estimators have similar performance when $\bar{Q}_n^0$ is correctly specified. When both models are misspecified Cao's estimator performs as well as OLS. OLS, CAO and C-TMLE$_{Y*}$ are least biased, and TanRV has the smallest MSE. The performance of all other estimators degrades under dual misspecification. Arguably, the most interesting test case for all estimators (given that they are all enforced to use parametric models) is Qmgc. TanWLS, TanRV, C-TMLE$_{Y*}$, WLS have the smallest MSE, and TanRV, TanWLS are least biased. The performance of both Tan estimators is unaffected by externally bounding g_n due to their internal bounding of g_n.

Figure 2 and Table 2 compare the results for each estimator when g_n is bounded from below at 0.025. Bounding g_n appears to be crucial for PRC in the case of Qmgm, and improves the performance of Cao's estimator for the Qmgc specification, but has little effect on the performance of the other estimators. However, this result does not generalize to other data generating distributions, where the selection bias is greater and sparsity is more extreme, as the next simulation demonstrates.

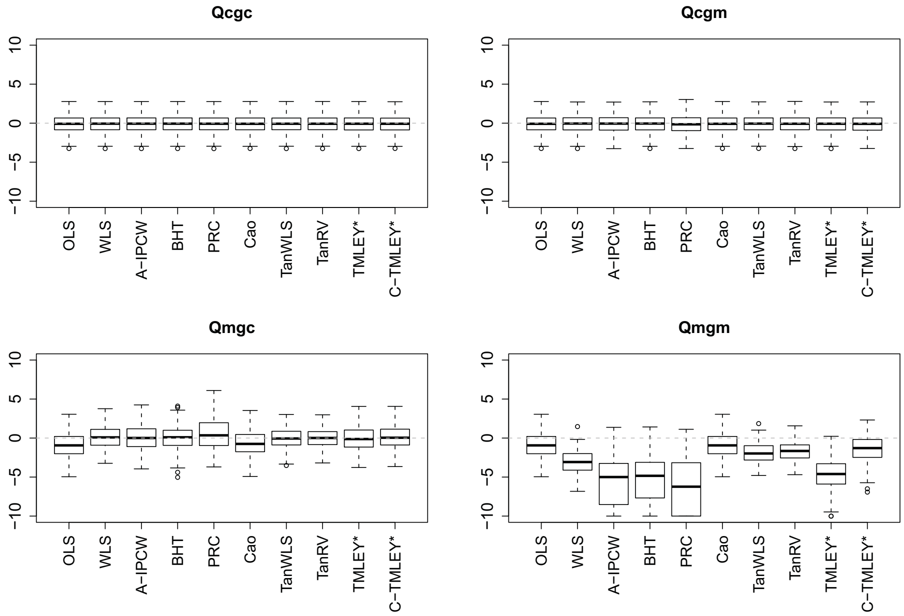

Figure 1: Sampling distribution of $(\mu_n - \mu_0)$ with no bounding of g_n, Kang and Schafer simulation.

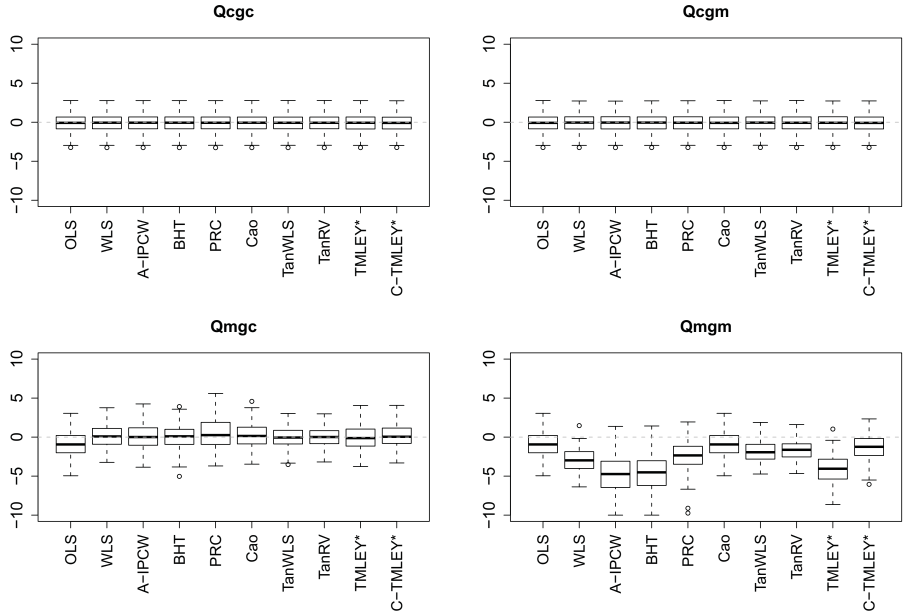

Figure 2: Sampling distribution of $(\mu_n - \mu_0)$ with g_n bounded at 0.025, Kang and Schafer simulation.

Table 1: Kang and Schafer simulation results with no bounding of g_n.

	Qcgc			**Qcgm**			**Qmgc**			**Qmgm**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	−0.09	1.40	1.41	−0.09	1.40	1.41	−0.93	1.97	2.82	−0.93	1.97	2.82
WLS	−0.09	1.40	1.41	−0.09	1.41	1.41	0.10	1.84	1.84	−3.04	2.08	11.33
A-IPCW	−0.09	1.40	1.41	−0.10	1.45	1.45	0.04	2.52	2.51	−8.81	2.3e+2	3.1e+2
BHT	−0.09	1.40	1.41	−0.09	1.41	1.41	0.01	2.34	2.33	−7.08	62.47	1.1e+2
PRC	−0.09	1.40	1.40	−0.12	1.44	1.45	0.56	3.61	3.91	−37.24	4.9e+4	5.0e+4
Cao	−0.09	1.40	1.41	−0.09	1.40	1.41	−0.69	2.27	2.74	−0.93	1.97	2.82
Tan.WLS	−0.09	1.40	1.40	−0.09	1.40	1.41	−0.01	1.55	1.54	−1.93	1.62	5.33
Tan.RV	−0.09	1.40	1.40	−0.09	1.40	1.40	0.03	1.44	1.44	−1.67	1.51	4.31
TMLE$_{Y*}$	−0.10	1.40	1.41	−0.11	1.40	1.40	−0.09	2.12	2.12	−4.61	3.62	24.84
C-TMLE$_{Y*}$	−0.10	1.40	1.41	−0.11	1.40	1.40	0.09	1.77	1.77	−1.49	2.76	4.97

Table 2: Kang and Schafer simulation results, g_n bounded at 0.025.

	Qcgc			**Qcgm**			**Qmgc**			**Qmgm**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	−0.09	1.40	1.41	−0.09	1.40	1.41	−0.93	1.97	2.82	−0.93	1.97	2.82
WLS	−0.09	1.40	1.41	−0.09	1.41	1.41	0.10	1.84	1.84	−2.94	1.97	10.59
A-IPCW	−0.09	1.40	1.41	−0.09	1.41	1.41	0.04	2.44	2.43	−4.85	6.10	29.64
BHT	−0.09	1.40	1.41	−0.09	1.41	1.41	0.03	2.20	2.19	−4.65	5.35	26.95
PRC	−0.09	1.40	1.40	−0.09	1.40	1.41	0.51	3.47	3.72	−2.40	3.08	8.85
Cao	−0.09	1.40	1.41	−0.09	1.40	1.41	0.18	2.17	2.20	−0.93	1.97	2.83
Tan.WLS	−0.09	1.40	1.40	−0.09	1.40	1.41	−0.01	1.55	1.54	−1.91	1.63	5.25
Tan.RV	−0.09	1.40	1.40	−0.09	1.40	1.41	0.03	1.44	1.44	−1.66	1.52	4.26
TMLE$_{Y*}$	−0.10	1.40	1.41	−0.10	1.41	1.41	−0.09	2.10	2.10	−4.12	3.10	20.04
C-TMLE$_{Y*}$	−0.10	1.40	1.41	−0.10	1.40	1.41	0.11	1.74	1.74	−1.37	2.30	4.16

5.2 Modification 1 of Kang and Schafer Simulation

In the KS simulation, when $\bar{Q}_0$ or g_0 are misspecified the misspecifications are small, and the selection bias is small. Therefore, we modified the KS simulation in order to increase the degree of misspecification and selection bias. This creates a greater challenge for estimators, and better highlights their relative performance.

As before, let Z_j be i.i.d. $N(0,1)$. The outcome Y is generated as $Y = 210 + 50Z_1 + 25Z_2 + 25Z_3 + 25Z_4 + N(0,1)$. The covariates actually observed by the data analyst are now given by the following functions of $(Z_1, \ldots, Z_4)$:

$$\begin{aligned} W_1 &= exp(Z_1^2/2) \\ W_2 &= 0.5Z_2/(1+exp(Z_1^2))+3 \\ W_3 &= (Z_1^2 Z_3/25+0.6)^3+2 \\ W_4 &= (Z_2+0.6Z_4)^2+2. \end{aligned}$$

From this one can determine the true regression function $\bar{Q}_0(W) = E_0(E(Y \mid Z) \mid W)$. The missingness indicator is generated as follows:

$$g_0(1 \mid W) = \text{expit}(-2Z_1 + Z_2 - 0.5Z_3 - 0.2Z_4).$$

A misspecified fit is now obtained by fitting a linear or logistic main term regression in $W_1, \ldots, W_4$, while a correct fit is obtained by providing the user with the terms $Z_1, \ldots, Z_4$, and fitting a linear or logistic main term regression in $Z_1, \ldots, Z_4$. With these modifications, the population mean is again 210, but the mean among respondents is 184.4. With these modifications, we have a higher degree of practical violation of the positivity assumption: $g_0(\Delta = 1 \mid W) \in [1.1 \times 10^{-5}, 0.99]$ while the estimated probabilities, $g_n(\Delta = 1 \mid W)$, were observed to fall in the range $[2.2 \times 10^{-16}, 0.87]$.

Figure 3 and Table 3 presents results for misspecified $\bar{Q}_n^0$ without bounding g_n and with g_n bounded at 0.025. Bounding dramatically reduces the variance of all estimators, except OLS, Tan.WLS and Tan.RV, but recall that Tan estimators always internally bound g_n away from 0. This improved efficiency comes at the cost of a slight increase in bias for all estimators except PRC.The variance and MSE of C-TMLE$_{Y*}$ is less than half of the other non-TMLE estimators. In contrast to the results on the previous simulation, Cao, Tan.WLS, and Tan.RV exhibit a lack of robustness at this level of sparsity when forced to rely on g_n at misspecified $\bar{Q}_n^0$.

Table 3: Modification 1 of Kang and Schafer simulation, Q misspecified.

		Qmgc			**Qmgm**		
	lb on g_n	Bias	Var	MSE	Bias	Var	MSE
OLS	0	−35.56	16.58	1.3e+3	−35.56	16.58	1.3e+3
	0.025	−35.56	16.58	1.3e+3	−35.56	16.58	1.3e+3
WLS	0	−4.40	41.95	61.15	−34.67	15.95	1.2e+3
	0.025	−5.52	31.62	61.93	−34.67	15.95	1.2e+3
A-IPCW	0	−1.83	1.9e+2	2.0e+2	−34.75	17.19	1.2e+3
	0.025	−5.88	42.63	77.09	−34.75	17.19	1.2e+3
BHT	0	−3.04	64.63	73.59	−34.75	17.17	1.2e+3
	0.025	−5.03	32.89	58.02	−34.75	17.17	1.2e+3
PRC	0	80.64	8.7e+3	1.5e+4	1.25e+11	1.74e+25	1.75e+25
	0.025	9.27	2.2e+2	3.0e+2	-34.38	15.28	1.2e+3
Cao	0	−6.17	44.68	82.52	−35.57	16.58	1.3e+3
	0.025	−24.25	21.79	6.1e+2	−35.50	17.87	1.3e+3
Tan.WLS	0	−3.59	24.29	37.07	−33.64	42.37	1.2e+3
	0.025	−3.64	22.95	36.09	-33.49	50.00	1.2e+3
Tan.RV	0	5.22	93.77	1.2e+2	−34.69	63.16	1.3e+3
	0.025	5.28	94.11	1.2e+2	−34.65	64.21	1.3e+3
TMLE_{Y*}	0	−0.04	89.33	88.98	−33.74	6.48	1.1e+3
	0.025	1.00	22.05	22.96	−33.74	6.48	1.1e+3
C-TMLE_{Y*}	0	−0.64	15.55	15.90	−34.26	6.66	1.2e+3
	0.025	−1.50	11.96	14.17	−34.19	6.82	1.2e+3

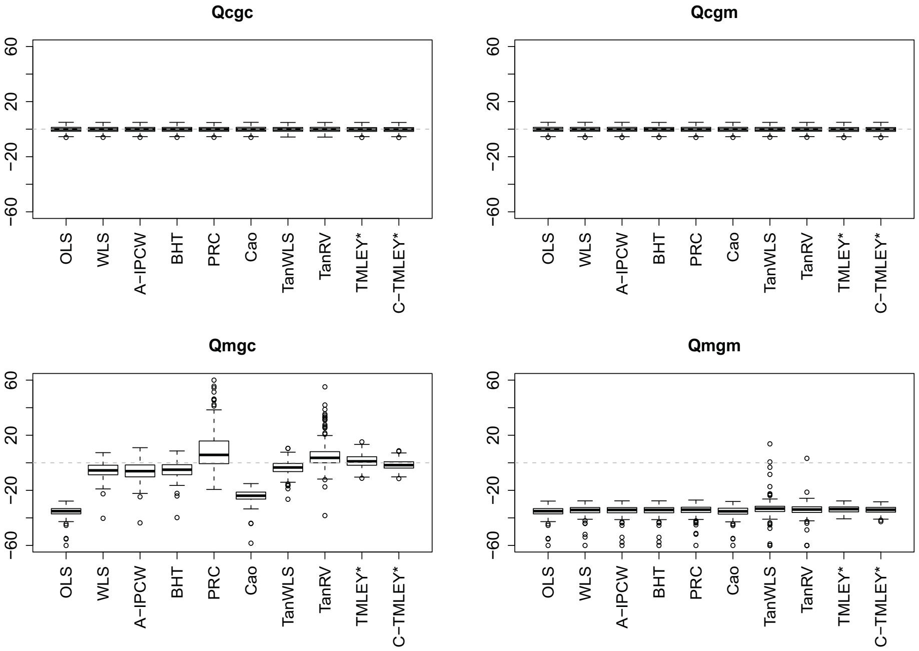

Figure 3: Sampling distribution of $(\mu_n - \mu_0)$ with g_n bounded at 0.025, Modification 1 of Kang and Schafer simulation.

5.3 Modification 2 of Kang and Schafer Simulation

For this simulation, we made one additional change to Modification 1: we set the coefficient in front of Z_4 in the true regression of Y on Z equal to zero. Therefore, while Z_4 is still associated with missingness, it is not associated with the outcome, and is thus not a confounder. Given $(W_1, \ldots, W_3)$, W_4 is not associated with the outcome either, and therefore as misspecified regression model of $\bar{Q}_0(W)$ we use a main term regression in (W_1, W_2, W_3).

This modification to the KS simulation enables us to take the debate on the relative performance of DR estimators one step further, by addressing a second key challenge of the estimators: that they often include non-confounders in the censoring mechanism estimator. Though such an estimator remains asymptotically unbiased, this unnecessary inclusion can increase asymptotic variance, and may unnecessarily introduce positivity violations leading to finite sample bias and inflated variance (Neugebauer and van der Laan, 2005; Petersen et al., 2010).

Figure 4 and Table 4 reveal that C-TMLE$_{Y*}$ has superior performance relative to estimating equation-based DR estimators when not all covariates are associated with Y. As discussed earlier, the C-TMLE algorithm provides an innovative black-box approach for estimating the censoring mechanism, preferring covariates that are associated with the outcome and censoring, without "data-snooping."

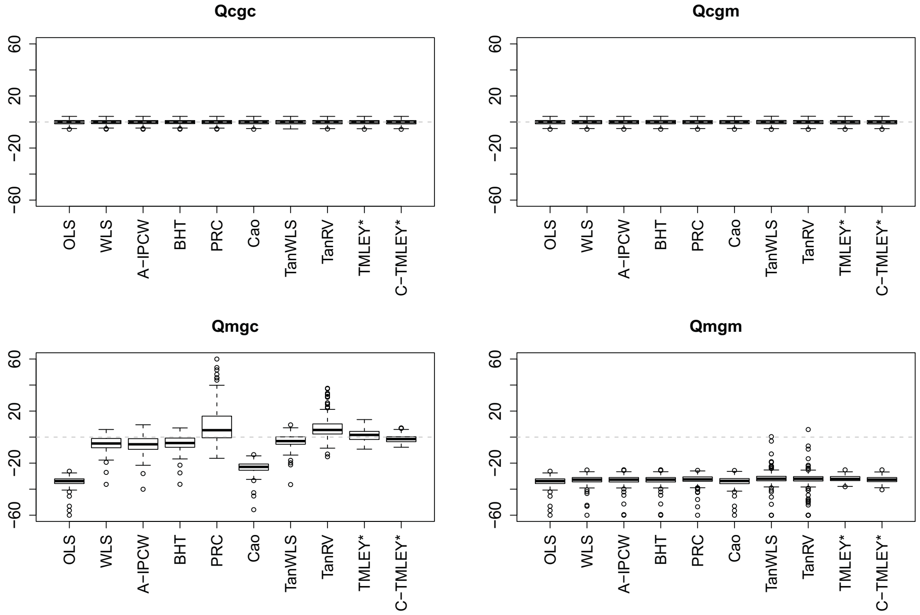

Figure 4: Sampling distribution of $(\mu_n - \mu_0)$ with g_n bounded at 0.025, Modification 2 of Kang and Schafer simulation.

5.4 Modification 3 of Kang and Schafer Simulation

In some rare cases, a C-TMLE can be a super efficient estimator because they use a collaborative estimator g_n that takes into account the fit of the initial estimator $\bar{Q}_n^0$ (we refer to Rotnitzky et al. (2010) and van der Laan and Gruber (2009) for a detailed discussion). As a consequence, it is of particular interest to investigate the behavior of C-TMLE$_{Y*}$ in the previous simulation but with the coefficient in front of Z_4 set equal to $C/\sqrt{n}$, for a number of values of C, in the data generating mechanism for the outcome, $Y = 210 + 50Z_1 + 25Z_2 + 25Z_3 + C/\sqrt{n}Z_4 + N(0,1)$. We report the results

Table 4: Modification 2 of Kang and Schafer simulation, Q misspecified.

		Qmgc			**Qmgm**		
	lb on g_n	Bias	Var	MSE	Bias	Var	MSE
OLS	0	−34.25	15.24	1.2e+3	−34.25	15.24	1.2e+3
	0.025	−34.25	15.24	1.2e+3	−34.25	15.24	1.2e+3
WLS	0	−3.64	39.52	52.61	−33.09	15.18	1.1e+3
	0.025	−4.92	28.65	52.75	−33.09	15.18	1.1e+3
A-IPCW	0	−1.11	$1.8e + 2$	1.8e+2	−33.14	16.47	1.1e+3
	0.025	−5.39	39.01	67.89	−33.14	16.47	1.1e+3
BHT	0	−2.27	72.06	76.91	−33.14	16.43	1.1e+3
	0.025	−4.57	29.73	50.49	−33.14	16.43	1.1e+3
PRC	0	77.78	7.7e+3	1.4e+4	5.4e+11	4.5e+25	4.5e+25
	0.025	9.11	2.0e+2	2.8e+2	−32.79	14.13	1.1e+3
Cao	0	−5.55	40.60	71.21	−34.25	15.25	1.2e+3
	0.025	−23.37	20.54	5.7e+2	−34.16	16.48	1.2e+3
Tan.WLS	0	−2.95	23.74	32.32	−32.02	49.66	1.1e+3
	0.025	−3.11	23.32	32.91	−32.02	43.37	1.1e+3
Tan.RV	0	6.87	65.77	1.1e+2	−32.95	89.67	1.2e+3
	0.025	6.94	65.02	1.1e+2	−32.87	71.78	1.2e+3
TMLE$_{Y*}$	0	0.15	76.03	75.75	−31.99	5.64	1.0e+3
	0.025	1.26	17.77	19.29	−32.00	5.60	1.0e+3
C-TMLE$_{Y*}$	0	−0.88	10.69	11.42	−32.58	5.83	1.1e+3
	0.025	−1.37	8.48	10.34	−32.68	8.48	1.1e+3

for $C = \{10, 20, 50\}$. Table 5 provides the results at each value of C for all estimators when $\bar{Q}_n^0$ is correctly specified and g_n is misspecified, and when $\bar{Q}_n^0$ is misspecified and g_n is both correctly and mis-specified. In each case, g_n is bounded at 0.025. We note that C-TMLE$_{Y*}$ does not break down, even under these particularly challenging conditions (nor under other simulated scenarios presented in Gruber and van der Laan (2010b)). It is an open question how a C-TMLE performs at different local data generating distributions when it is superefficient, and further research is warranted.

Table 5: Modification 3 to Kang and Schafer simulation, $C/\sqrt{n}$ perturbation, g_n bounded at 0.025.

	C = 10			**C = 20**			**C = 50**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
				Qcgm					
OLS	−0.06	3.94	3.92	−0.06	3.94	3.93	−0.06	3.94	3.93
WLS	−0.06	3.94	3.93	−0.06	3.94	3.93	−0.06	3.95	3.94
A-IPCW	−0.06	3.94	3.93	−0.06	3.94	3.93	−0.06	3.95	3.94
BHT	−0.06	3.97	3.96	−0.06	3.97	3.96	−0.06	3.98	3.97
PRC	−0.06	3.94	3.93	−0.06	3.94	3.93	−0.06	3.95	3.94
Cao	−0.06	3.93	3.92	−0.06	3.93	3.92	−0.06	3.94	3.93
Tan.WLS	−0.06	4.03	4.01	−0.06	4.03	4.02	−0.07	4.03	4.02
Tan.RV	−0.06	4.03	4.01	−0.06	4.03	4.01	−0.07	4.03	4.02
TMLE$_{Y*}$	−0.10	3.92	3.92	−0.11	3.92	3.92	−0.11	3.93	3.93
C-TMLE$_{Y*}$	−0.10	3.92	3.92	−0.10	3.92	3.92	−0.11	3.93	3.93
				Qmgc					
OLS	−34.28	15.25	1.2e+3	−34.29	15.25	1.2e+3	−34.34	15.24	1.2e+3
WLS	−5.13	28.24	54.44	−5.13	28.25	54.50	−5.15	28.28	54.68
A-IPCW	−5.47	38.63	68.38	−5.47	38.64	68.45	−5.49	38.69	68.67
BHT	−4.62	29.60	50.85	−4.63	29.61	50.90	−4.64	29.63	51.08
PRC	9.21	2.0e+2	2.8e+2	9.21	2.0e+2	2.8e+2	9.21	2.0e+2	2.8e+2
Cao	−23.42	20.47	5.7e+2	−23.43	20.47	5.7e+2	−23.46	20.48	5.7e+2
Tan.WLS	−3.25	21.00	31.45	−3.25	20.94	31.42	−3.26	20.78	31.35
Tan.RV	6.94	64.90	112.84	6.93	65.23	1.1e+2	6.88	66.37	1.1e+2
TMLE$_{Y*}$	1.17	18.03	19.34	1.17	18.02	19.32	1.16	18.02	19.29
C-TMLE$_{Y*}$	−1.63	8.01	10.64	−1.66	8.49	11.21	−1.68	8.83	11.63
				Qmgm					
OLS	−34.28	15.25	1.2e+3	−34.29	15.25	1.2e+3	−34.34	15.24	1.2e+3
WLS	−33.00	14.79	1.1e+3	−33.03	14.79	1.1e+3	−33.09	14.78	1.1e+3
A-IPCW	−33.05	16.39	1.1e+3	−33.07	16.38	1.1e+3	−33.13	16.35	1.1e+3
BHT	−33.05	16.36	1.1e+3	−33.07	16.35	1.1e+3	−33.13	16.32	1.1e+3
PRC	−32.39	14.45	1.1e+3	−32.42	14.44	1.1e+3	−32.49	14.40	1.1e+3
Cao	−34.18	16.50	1.2e+3	−34.20	16.49	1.2e+3	−34.25	16.48	1.2e+3
Tan.WLS	−32.76	73.05	1.1e+3	−32.72	76.88	1.1e+3	−32.75	76.83	1.1e+3
Tan.RV	−33.29	71.11	1.2e+3	−33.13	55.13	1.2e+3	−33.17	62.77	1.2e+3
TMLE$_{Y*}$	−32.03	5.57	1.0e+3	−32.05	5.56	1.0e+3	−32.12	5.54	1.0e+3
C-TMLE$_{Y*}$	−32.64	5.82	1.1e+3	−32.74	5.94	1.1e+3	−32.75	6.22	1.1e+3

6 TMLEs with Machine Learning for Dual Misspecification

The KS simulation with dual misspecification ($Qmgm$) can illustrate the benefits of coupling data-adaptive (super) learning with targeted maximum likelihood estimation. C-TMLE$_{Y*}$ constrained to use a main terms regression model with misspecified covariates (W_1, W_2, W_3, W_4) has smaller variance than $\mu_{n,OLS}$, but is more biased. The MSE of the TMLE$_{Y*}$ is larger than the MSE of C-TMLE$_{Y*}$, with increased bias and variance. We ask how the estimation process should be affected if we assume that parametric models are seldom correctly specified and that main term regression techniques generally fail in capturing the true relationships between predictors and an outcome. Our answer is that the estimation process should incorporate data-adaptive machine learning.

We coupled super learning with TMLE$_{Y*}$ and C-TMLE$_{Y*}$ to estimate both $\bar{Q}_0$ and g_0. For C-TMLE$_{Y*}$, four missingness-mechanism score-based covariates were created based on different truncation levels of the propensity score estimate $g_n(1 \mid W)$: no truncation, and truncation from below at the 0.01, 0.025, and 0.05-percentile. These four scores were supplied along with the misspecified main terms $W_1, \ldots, W_4$ to the targeted forward selection algorithm in the C-TMLE$_{Y*}$ used to build a series of candidate nested logistic regression estimators of the missingness mechanism and corresponding candidate TMLEs. The C-TMLE$_{Y*}$ algorithm used 5-fold cross-validation to select the best estimate from the eight candidate TMLEs. This allows the C-TMLE algorithm to build a logistic regression fit of g_0 that selects among the misspecified main-terms and super-learning fits of the missingness mechanism score $g_n(1 \mid W)$ at different truncation levels.

An important aspect of super learning is to ensure that the library of prediction algorithms includes a variety of approaches for fitting the true function $\bar{Q}_0$ and g_0. For example, it is sensible to include a main terms regression algorithm in the super learner library. Should that algorithm happen to be correct, the super learner will behave as the main terms regression algorithm. It is also recommended to include algorithms that search over a space of higher order polynomials, non-linear models, and, for example, cubic splines. For binary outcome regression, as required for fitting g_0, classification algorithms such as classification and regression trees (Breiman et al., 1984), support vector machines (Cortes and Vapnik, 1995)), and k-nearest-neighbor algorithms (Friedman (1994)), could be added to the library. The point of super-learning is that we cannot know in advance which procedure will be most successful for a given prediction problem. Super learning relies on the oracle property of V-fold cross-validation to asymptotically select the optimal convex combination of estimates obtained from these disparate procedures (van der Laan and Dudoit (2003); van der Laan et al. (2004), van der Laan et al. (2007)).

Consider the misspecified scenario proposed by KS. The true full-data distribution and the missingness mechanism are captured by main terms linear regression of the outcome on Z_1, Z_2, Z_3, Z_4. This simple model is virtually impossible to discover through the usual model selection approaches when the observed data consists of misspecified covariates $O = (W_1, W_2, W_3, W_4, \Delta, \Delta Y)$, given

$$\begin{aligned} Z_1 &= 2log(W_1), \\ Z_2 &= (W_2 - 10)(1 + 2W_1), \\ Z_3 &= \frac{25(W_3 - 0.6)}{2log(W_1)}, \\ Z_4 &= \sqrt[3]{W_4} - 20 - (W_2 - 10)(1 + 2W_1). \end{aligned}$$

This complexity illustrates the importance of including prediction algorithms that attack the estimation problem from a variety of directions. The super learner library we employed contained the algorithms listed below. The analysis was carried out in the R statistical programming environment v2.10.1 (Team, 2010), using algorithms included in the base installation or in the indicated package.

- **glm** (base) main terms linear regression.
- **step** (base) stepwise forward and backward selection using the AIC criterion (Hastie and Pregibon, 1992).
- **ipredbagg** (ipred) bagging for classification, regression and survival trees (Peters and Hothorn, 2009; Breiman, 1996).
- **DSA** (DSA) Deletion/Selection/Addition algorithm for searching over a space of polynomial models or order k (k set to 2). (Neugebauer and Bullard, 2010; Sinisi and van der Laan, 2004)

- **earth** (earth) Building a regression model using multivariate adaptive regression splines (MARS) (Milborrow, 2009; Friedman, 1991, 1993).
- **loess** (stats) Local polynomial regression fitting (W. S. Cleveland and Shyu, 1992).
- **nnet** (nnet) Single-hidden-layer neural network for classification (Venables and Ripley, 2002; Ripley, 1996).
- **svm** (e1071) Support vector machine for regression and classification (Dimitriadou et al., 2010; Chang and Lin, 2001).
- ***k*-nearest-neighbors*** (class) classification using most common outcome among identified k nearest nodes (k set to 10) (Venables and Ripley, 2002; Friedman, 1994)

* binary outcomes only, added to library for estimating g

Table 6 reports the results when super learning is incorporated into TMLE_{Y*} and C-TMLE_{Y*} estimation procedures, based on 250 samples of size 1000, with predicted values for $g_n(1 \mid W)$ truncated from below at 0.025. Using the data-adaptive estimator approach improved bias and variance of both estimators. TMLE_{Y*} efficiency improved by a factor of 8.5, and C-TMLE_{Y*} efficiency improved by a factor of 1.5. In addition, the MSE for both data-adaptive estimators is smaller than the MSE of the estimator that performed the best when both Q and g were misspecified, $\mu_{n,OLS}$ (MSE = 2.82).

Table 6: Results with and without incorporating super learning into TMLE_{Y*} and C-TMLE_{Y*}, $Qmgm$, g_n truncated at 0.025.

	Bias	**Var**	**MSE**
TMLE_{Y*}	-4.12	3.10	20.0
TMLE_{Y*} + SL	-0.77	1.51	2.10
C-TMLE_{Y*}	-1.37	2.30	4.16
C-TMLE_{Y*} + SL	-1.05	1.54	2.64

7 Discussion

By mapping continuous outcomes into [0,1] and using a logistic fluctuation, TMLE$_{Y*}$ and C-TMLE$_{Y*}$ are more robust to violations of the positivity assumption than the TMLEs using the linear fluctuation function. By being a substitution estimator, it follows that the impact of a single observation on TMLE$_{Y*}$ is bounded by $1/n$ while many of the other estimators do not have such a robustness property. We show that C-TMLE$_{Y*}$ has superior performance relative to estimating equation-based DR estimators when there are covariates that are strongly associated with the missingness indicator, while weakly or not at all associated with the outcome Y. The C-TMLE algorithm provides an innovative approach for estimating the censoring mechanism, preferring covariates that are associated with the outcome Y and missingness, Δ. C-TMLEs avoid data snooping concerns because the estimation procedure is fully specified before the analyst observes any data (or at least, not any data beyond some ancillary statistics). Even in cases in which *all* observed covariates are associated with Y, C-TMLE still performs well.

Related work is also being done with respect to other parameters of interest. For example, both Cao et al. (2009) and Tan (2006) include discussions on applying their estimators to causal effect parameters. In addition, Freedman and Berk (2008), focus on a causal effect parameter, and demonstrate that DR estimators (and the WLS estimator in particular) can increase variance and bias when IPCW are large.

Overall, comparisons of estimators, beyond theoretical studies of asymptotics as well as robustness, will need to be based on large scale simulation studies including all available estimators, and cannot be tailored towards one particular simulation setting. Future research should be concerned with setting up such a large scale objective comparison based on publicly available software, and we are looking forward to contributing to such an effort.

The research underlying TMLEs was motivated, in part, by the goal of increasing the stability of DR estimators, and the KS simulations provide a demonstration of the merits of TMLEs under violations of the positivity assumption. TMLEs are estimators defined by the choice of loss function, and parametric submodel, both chosen so that the linear span of the scores at zero fluctuation with respect to the loss function includes the efficient influence curve/efficient score. All such TMLEs are double robust, asymptotically efficient under correct specification, and substitution estimators, but the choice of submodel can affect the finite sample robustness if the submodel does not respect any bounds such as the linear regression submodel for the TMLE.In addition, TMLEs can be combined with super learning and empirical efficiency maximization (Rubin and van der Laan (2008) and van der Laan and Gruber (2009)) to further enhance their performance in practice. We hope that by showing that these estimators perform well in simulations and settings created by other researchers for the purposes of showing the weaknesses of DR estimators, as well as in modified simulations that make estimation even more challenging, we provide probative evidence in support of TMLEs. Of course, much can happen in finite samples, and we look forward to further exploring how these estimators perform in other settings.

References

O. Bembom and M.J. van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report 230, Division of Biostatstics, University of California, Berkeley, 2008. URL `www.bepress.com/ucbbiostat/paper230/`.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees.* The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

W. Cao, A.A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96,3:723–734, 2009.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines (version 2.31). Technical report, 2001. `http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm2.ps.gz`.

S.R. Cole and M.A. Hernan. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168: 656–664, 2008.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20: 273–297, December 1995.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, , and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2010. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.5-24.

D.A. Freedman and R.A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32,4:392–409, 2008.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):pp. 1–67, 1991. ISSN 00905364. URL `http://www.jstor.org/stable/2241837`.

J. H. Friedman. Fast MARS. Technical report, Department of Statistics, Stanford University, 1993.

J. H. Friedman. Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University, 1994.

C. J. Geyer. *trust: Trust Region Optimization*, 2009. URL `http://CRAN.R-project.org/package=trust`. R package version 0.1-2.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Technical Report 265, UC Berkeley, 2010a.

S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6,1(18), 2010b.

T.J. Hastie and D. Pregibon. Generalized linear models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 6. Wadsworth & Brooks/Cole, 1992.

A. Rotnitzky J.M. Robins and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.*, 89:846–66, 1994.

J. Kang and J. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:523–39, 2007.

L. Kish. Weighting for unequal p_i. *Journal of Official Statistics*, 8:183–200, 1992.

S Milborrow. *earth: Multivariate Adaptive Regression Spline Models*, 2009. URL `http://CRAN.R-project.org/package=earth`. R package version 2.4-0.

K.L. Moore, R.S. Neugebauer, M.J. van der Laan, and I.B. Tager. Causal inference in epidemiological studies with strong confounding. Technical Report 255, Division of Biostatistics, University of California, Berkeley, 2009. URL `www.bepress.com/ucbbiostat/paper255/`.

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.

R. Neugebauer and M.J. van der Laan. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, 129, Issues 1-2:405–426, 2005.

A Peters and T Hothorn. *ipred: Improved Predictors*, 2009. URL `http://CRAN.R-project.org/package=ipred`. R package version 0.8-8.

M.L. Petersen, K. Porter, S. Gruber, Y. Wang, and M.J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 2010.

G. Ridgeway and D. McCaffrey. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22: 540–43, 2007.

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, New York, 1996.

J. M. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, 22:544–559, 2007.

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 Modelling*, 7:1393–1512, 1986.

J.M. Robins. Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987. ISSN 0097-4943.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association: Section on Bayesian Statistical Science*, pages 6–10, 1999.

J.M. Robins. Commentary on using inverse weighting and predictive inference to estimate the effecs of time-varying treatments on the discrete-time hazard. *Statistics in Medicine*, (21):1663–1680, 1999.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–66, September 1994.

M. Rosenblum and M. J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(19), 2010.

A. Rotnitzky, L. Li, and X. Li. A note on overadjustment in inverse probability weighted estimation. *Biometrika*, 97(4):997–1001, 2010.

D.B. Rubin and M.J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, Vol. 4, Iss. 1, Article 5, 2008.

D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association*, (94): 1096–1120 (1121–1146), 1999.

J.S. Sekhon, S. Gruber, K. Porter, and M.J. van der Laan. Propensity-score-based estimators and C-TMLE. In M.J. van der Laan and S. Rose, *Targeted Learning: Prediction and Causal Inference for Observational and Experimental Data*, chapter 21. Springer, New York, 2011.

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.

Z. Tan. A distributional approach for causal inference using propensity scores. *J. Am. Statist. Assoc.*, 101:1619–37, 2006.

Z. Tan. Comment: Understanding OR, PS and DR. *Statistical Science*, 22: 560–568, 2007.

Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97,3:661–682, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

A. Tsiatis and M. Davidian. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:569–73, 2007.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003.

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 2009.

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

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. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. Technical report 142, Division of Biostatistics, University of California, Berkeley, February 2004.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

M.J. van der Laan, S. Rose, and S. Gruber. Readings on targeted maximum likelihood estimation. *Technical report, working paper series http://www.bepress.com/ucbbiostat/paper254*, 2009.

R. Varadhan. *alabama: Constrained nonlinear optimization*, 2010. URL `http://CRAN.R-project.org/package=alabama`. R package version 2010.10-1.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. Springer, New York, 4th edition, 2002.

E. Grosse W. S. Cleveland and W. M. Shyu. Local regression models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 6. Wadsworth & Brooks/Cole, 1992.

Y. Wang, M. Petersen, D. Bangsberg, and M.J. van der Laan. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Technical Report 211, Division of Biostatistics, University of California, Berkeley, 2006a.

Y. Wang, M. Petersen, and M.J. van der Laan. A statistical method for diagnosing ETA bias in IPTW estimators. Technical report, Division of Biostatistics, University of California, Berkeley, 2006b.

R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 1974.