Title
Collaborative Targeted Maximum Likelihood Estimation

Author
Gruber, Susan

Publication Date
2011

Peer reviewed|Thesis/dissertation

Collaborative Targeted Maximum Likelihood Estimation

by

Susan Gruber

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark J. van der Laan, Chair
Professor Nicholas P. Jewell
Associate Professor Jasjeet S. Sekhon

Spring 2011

Collaborative Targeted Maximum Likelihood Estimation

Abstract

Collaborative Targeted Maximum Likelihood Estimation

by

Susan Gruber

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark J. van der Laan, Chair

Collaborative targeted maximum likelihood estimation is an extension to targeted maximum likelihood estimation (TMLE), first introduced in van der Laan and Rubin (2006a). TMLE is an efficient, double robust (DR), semi-parametric methodology for estimating a pathwise differential parameter of a statistical distribution given censored data. The TMLE procedure involves a parametric fluctuation of an initial estimate of the relevant factor of the density of the observed data (Q), that involves estimating the nuisance portion of the likelihood—censoring mechanism, g. DR estimators are consistent when at least one of these is estimated consistently, when assumptions described below are met. As an efficient estimator, under regularity conditions TMLE achieves the semi-parametric efficiency bound when both are estimated consistently. TMLE is also a substitution estimator, and as such respects global bounds on the parameter and the data. TMLE is especially valuable in high dimensional settings, where parametric approaches fail due to the curse of dimensionality, and the bias/variance tradeoff made by other semi-parametric approaches is not designed to be optimal for the parameter of interest.

Though the best approach to nuisance parameter estimation is a current topic of debate in the literature, methods typically rely on maximizing a likelihood, possibly constrained to ensure that predicted probabilities are bounded away from $(0, 1)$ (e.g., Cao et al. (2009); Tan (2006, 2010)). However, by establishing the collaborative double robustness of the efficient influence curve, van der Laan and Gruber (2010) provides a theoretical justification for moving away from the practice of external nuisance parameter estimation. That article also presents the collaborative TMLE (C-TMLE), and provides an algorithm for constructing the estimator. Starting with an initial estimate of the Q portion of the likelihood, C-TMLE builds a series candidate nuisance parameter estimates, $g_{n,k}$, based on their ability to improve the goodness of fit for Q, while simultaneously increasing the goodness of fit for g. These index a series of candidate TMLEs. The candidate selected by minimum loss-based cross-validation is the C-TMLE estimator. This approach to targeted maximum likelihood estimation is especially useful when the true censoring mechanism is unknown, and adjusting for a large number of correlated confounders leads to highly variable estimates. It is also particularly effective when there is sparsity in the data that renders a statistical parameter borderline identifiable.

This dissertation discusses C-TMLE, and extensions and applications of TMLE primarily in the context of binary point treatment effect estimation, but results generalize. For example, C-TMLE has already been extended to survival analysis (Stitelman and van der Laan, 2010). The dissertation is structured as follows. After a brief review of TMLE and defining the notation used throughout, Chapter 2 discusses collaborative double robustness of the efficient influence curve. The implications of this property provide an understanding of the requirements for nuisance parameter estimation that led to the development of C-TMLE. Chapter 3 defines the TMLE for bounded continuous outcomes that enforces known global constraints on the statistical model. The C-TMLE algorithm is presented in Chapter 4, along with simulation studies designed to illustrate the behavior of C-TMLE, and an analysis of genomics data where high correlations among mutations make it difficult to obtain meaningful estimates of fully-adjusted association measures. Chapter 5 considers estimator performance under sparse data conditions and model misspecification. This chapter clearly demonstrates that exploiting collaborative double robustness can yield big gains with respect to bias and variance. It also highlights another important aspect of TMLEs, their ability to incorporate data-adaptive estimation while still providing valid influence-curve based inference. Much of the work made use of super learning, an ensemble method for prediction that uses cross-validation to select the optimal convex combination of predictions from individual prediction algorithms (van der Laan et al., 2007). Data-adaptivity lessens the reliance on a priori specified models, and results in Chapter 5 demonstrate the benefits of this approach. In addition, Chapter 6 demonstrates that when these a priori specified models are included in the model space searched over by the super learner, results using the most aggressive procedures are identical (bias and variance) to those obtained by relying on a pre-specified parametric model that happens to be correct. The main focus of Chapter 6 is the application of TMLE to the meta-analysis of safety data. An appendix describes *tmle*, an R package for targeted maximum likelihood estimation.

Contents

List of Figures

List of Tables

Acknowledgments

Mark van der Laan is an extraordinary advisor, and I will always be grateful to him for all he has taught me, and all the opportunities he made available to me. Because of him these past three years have been the most intellectually stimulating and rewarding period of my life.

I have also been surrounded by brilliant, collaborative graduate students. I've learned so much from conversations with Jim Bullard, Eric Polley, Kelly Moore, Ori Stitelman, Wen Zheng, Iván Díaz, Jordan Brooks, Paul Chaffee, and Romain Neugebauer, and also from Oliver Bembom, whose paper on data adaptive adjustment set selection first got me interested in the problem of nuisance parameter estimation. I am especially grateful to Eric, Jim, and Romain for their reliable Super Learner and DSA software.

I am also grateful to Jas Sekhon, Maya Petersen, Sandrine Dudoit, Nick Jewell, Alan Hubbard, Steve Selvin, Maureen Lahiff, and David Freedman for their contributions to my education. Special thanks to Sandrine for advice and support when I needed it most, and to Jas for arranging the presentation of our work at the 2010 Atlantic Causal Conference, which turned out to be a pivotal event for me.

Versions of the material presented here have been published elsewhere. Chapters 3 and 4 appeared in *The International Journal of Biostatistics* as "A Targeted Maximum Likelihood Estimator for A Bounded Continuous Outcome," and "An Application of Collaborative Targeted Maximum Likelihood Estimation in Causal Inference and Genomics," co-authored with Mark van der Laan. Material from these two chapters and from Chapter 5, co-authored with Kristin Porter, Jasjeet S. Sekhon, and Mark van der Laan, is included in *Targeted Learning: Prediction and Causal Inference for Observational and Experimental Data,* by Mark van der Laan and Sherri Rose, Springer. I'd like to thank Thamban Valappil, Greg Soon, and Dan Rubin at the United States Food and Drug Administration for their input and for providing the data analyzed in Chapter 6, co-authored with Mark van der Laan. The article in the appendix describing an R package for targeted maximum likelihood estimation of binary point treatment effects, co-authored with Mark, is under review at *The Journal of Statistical Software*. Some of this material is presented in the introduction as well. I thank each of these individuals and publishers for their contributions to the work and permission to include it here.

I never would have gone back to school if it weren't for Judith Barnes, who demonstrated that we can make anything happen, at any age. I also want to acknowledge Patrick Diehl, Yomay Shyur, Christine Yang, Julia Lee, Alex Cohn, and Jordy Rose, whose lively minds allowed me to maintain a last, tenuous grip on sanity during the years when day-to-day life was considerably constrained. I hope there are more Nomic games in our future.

Jordy and Robert have been supportive of this whole journey. They progressed from middle school through high school and on to college in the six years I've been at Berkeley, and grown into thoughtful, caring adults. I feel supremely fortunate to have these two wonderful sons.

Dan has made all of this possible. He has always had faith in me, always wants what is best for me, and will make me feel fulfilled. This dissertation is dedicated to him, and finishing it signifies the start of an exciting new chapter in our life.

Chapter 1

Introduction

Collaborative targeted maximum likelihood estimation is an extension to targeted maximum likelihood estimation (TMLE), first introduced in van der Laan and Rubin (2006a). TMLE is an efficient, double robust (DR), semi-parametric methodology for estimating a pathwise differential parameter of a statistical distribution given censored data. The TMLE procedure involves a parametric fluctuation of an initial estimate of the relevant factor of the density of the observed data (Q), that involves estimating the nuisance portion of the likelihood—censoring mechanism, g. DR estimators are consistent when at least one of these is estimated consistently, when assumptions described below are met. As an efficient estimator, under regularity conditions TMLE achieves the semi-parametric efficiency bound when both are estimated consistently. TMLE is also a substitution estimator, and as such respects global bounds on the parameter and the data. TMLE is especially valuable in high dimensional settings, where parametric approaches fail due to the curse of dimensionality, and the bias/variance tradeoff made by other semi-parametric approaches is not designed to be optimal for the parameter of interest.

Though the best approach to nuisance parameter estimation is a current topic of debate in the literature, methods typically rely on maximizing a likelihood, possibly constrained to ensure that predicted probabilities are bounded away from $(0, 1)$ (e.g., Cao et al. (2009); Tan (2006, 2010)). However, by establishing the collaborative double robustness of the efficient influence curve, van der Laan and Gruber (2010) provides a theoretical justification for moving away from the practice of external nuisance parameter estimation. That article also presents the collaborative TMLE (C-TMLE), and provides an algorithm for constructing the estimator. Starting with an initial estimate of the Q portion of the likelihood, C-TMLE builds a series candidate nuisance parameter estimates, $g_{n,k}$, based on their ability to improve the goodness of fit for Q, while simultaneously increasing the goodness of fit for g. These index a series of candidate TMLEs. The candidate selected by minimum loss-based cross-validation is the C-TMLE estimator. This approach to targeted maximum likelihood estimation is especially useful when the true censoring mechanism is unknown, and adjusting for a large number of correlated confounders leads to highly variable estimates. It is also particularly effective when there is sparsity in the data that renders a statistical parameter borderline identifiable.

This dissertation discusses C-TMLE, and extensions and applications of TMLE primarily in the context of binary point treatment effect estimation, but results generalize. For example, C-TMLE has already been extended to survival analysis (Stitelman and van der Laan, 2010). The dissertation is structured as follows. After a brief review of TMLE and defining the notation used throughout, Chapter 2 discusses collaborative double robustness of the efficient influence curve. The implications of this property provide an understanding of the requirements for nuisance parameter estimation that led to the development of C-TMLE. Chapter 3 defines the TMLE for bounded continuous outcomes that enforces known global constraints on the statistical model. The C-TMLE algorithm is presented in Chapter 4, along with simulation studies designed to illustrate the behavior of C-TMLE, and an analysis of genomics data where high correlations among mutations make it difficult to obtain meaningful estimates of fully-adjusted association measures. Chapter 5 considers estimator performance under sparse data conditions and model misspecification. This chapter clearly demonstrates that exploiting collaborative double robustness can yield big gains with respect to bias and variance. It also highlights another important aspect of TMLEs, their ability to incorporate data-adaptive estimation while still providing valid influence-curve based inference. Much of the work made use of super learning, an ensemble method for prediction that uses cross-validation to select the optimal convex combination of predictions from individual prediction algorithms (van der Laan et al., 2007). Data-adaptivity lessens the reliance on a priori specified models, and results in Chapter 5 demonstrate the benefits of this approach. In addition, Chapter 6 demonstrates that when these a priori specified models are included in the model space searched over by the super learner, results using the most aggressive procedures are identical (bias and variance) to those obtained by relying on a pre-specified parametric model that happens to be correct. The main focus of Chapter 6 is the application of TMLE to the meta-analysis of safety data. Appendix A includes derivations of the influence curve for the log relative risk and log odds ratio parameters, and the contribution to the influence curve from the estimation of the censoring mechanism for causal effect parameters commonly reported in the literature: additive effect, relative risk, log relative risk, odds ratio, log odds ratio. R software for TMLE and C-TMLE is available on the world wide web (Gruber, 2010b,a). A document describing the *tmle* package for the R statistical programming environment (Team, 2010) is in Appendix B.

1.1 TMLE Review

Consider observations $O = (O_1, \ldots O_n) \sim P_0$, a true underlying probability distribution that gives rise to the data. Often, a particular feature of P_0, $\Psi(P_0)$, is of scientific interest. Global maximum likelihood estimation procedures for estimating P_0 can make a bias/variance tradeoff that is sub-optimal for estimation of $\Psi(P_0)$. The goal of TMLE is to improve this tradeoff.

Consider the application of TMLE to causal effect estimation. The counterfactual framework discussed in Rubin (1974) frames the estimation of causal effects as a missing data problem. Suppose we are interested in assessing the marginal difference in an outcome, Y, if everyone received treatment ($A = 1$) vs. everyone not receiving treatment ($A = 0$). If we could actually

measure the outcome under both scenarios for all individuals, the full data would be given as $X^{Full} = (Y_1, Y_0, W)$, where Y_1 is the counterfactual outcome corresponding to treatment $(A = 1)$, Y_0 is the counterfactual outcome under no treatment $(A = 0)$, and W is a vector of baseline covariates. A causal quantity of interest could be the additive causal effect $E_0Y_1 - E_0Y_0$. This parameter is defined non-parametrically on full data X^{Full} as $\psi_0^F = E_0Y_1 - E_0Y_0$, and identified from the observed data $O = (W, A, Y = Y_A)$ as $\Psi(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$ under the randomization assumption and positivity assumption. Here ψ_0^F denotes the causal quantity of interest, and ψ_0 is the statistical counterpart that can be interpreted as the causal effect ψ_0^F under these assumptions. We note that Ψ represents a mapping from a probability distribution of O into a real number, and Ψ is called the target parameter mapping.

Non-parametric structural equation modeling (NPSEM) provides an alternative paradigm for defining causal effect parameters (Pearl, 2010b). The following system of equations expresses the knowledge about the data generating mechanism:

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned}$$

where U_W, U_A, and U_Y are exogenous error terms. This NPSEM allows the definition of counterfactual outcomes $Y_a = f_Y(W, a, U_Y)$, corresponding with the intervention that sets the treatment node A equal to a, and thereby the causal quantity of interest, such as $E_0Y_1 - E_0Y_0$. The functions f_W, f_A, f_Y may be unspecified, one might assume exclusion restriction assumptions, or one might even assume parametric forms.

The statistical association measure ψ_0 can be interpreted as the additive causal effect of A on Y providing two assumptions are met: 1) Coarsening at random (CAR) is an assumption of conditional independence between treatment assignment and the full data given measured covariates, $A \perp X \mid W$ (Heitjan and Rubin (1991), Jacobsen and Keiding (1995), Gill et al. (1997)). This assumption indicates there are no unmeasured confounders of the effect of treatment on the outcome, i.e., U_A is independent of U_Y in the NPSEM. 2) The positivity assumption, also known as the experimental treatment assignment assumption (ETA), is that $\forall a \in \mathcal{A}, P(A = a \mid W) > 0$. In other words, if no observations within some stratum defined by W receive treatment at level $A = a$, then the data do not provide sufficient information to compare the effect of treatment at level a with no treatment, or with treatment at some other level. The parameter is borderline identifiable when there is a practical ETA violation, $\exists a \in \mathcal{A} : P(A = a \mid W) < \epsilon$, for some small ϵ relative to sample size.

The NPSEM approach and the counterfactual framework offer distinct formulations for discussing causality, yet each provides an equivalent foundation for defining causal effects as parameters of statistical distributions. When causal assumptions are not met, the statistical parameter represents an association measure that may still be of scientific interest.

A number of estimation procedures have been applied to causal effect estimation, including the maximum likelihood-based G-computation estimator (Robins, 1986), the inverse-probability-

of-treatment-weighted (IPTW) estimator (Hernan et al., 2000b; Robins, 2000b), the augmented IPTW estimator (Robins and Rotnitzky, 2001; Robins et al., 2000; Robins, 2000a). Scharfstein et al. (1999) presented a doubly robust regression-based estimator for the treatment specific mean, later extended to time-dependent censoring (Bang and Robins, 2005). See Rosenblum and van der Laan (2010) for a discussion of TMLE in relation to these other estimators. TMLE is a maximum likelihood based G-computation estimator that targets the fit of the data generating distribution towards reducing bias in the parameter of interest, generally one particular low-dimensional feature of the true underlying distribution.

TMLE is more generally referred to as Targeted Minimum Loss-based Estimation. At its core, in the above application, TMLE methodology involves fluctuating an initial estimate of the conditional mean outcome, and minimizing a loss function to select the magnitude of the fluctuation. The targeting fluctuation is parameter-specific. The loss function is not unique, but must be chosen with care to ensure that the fluctuated estimate is a parametric sub-model $M \in \mathcal{M}$, and that the risk of the loss function is indeed minimized at the truth. Targeted *maximum likelihood* estimation corresponds with choosing the negative log-likelihood loss function.

An orthogonal factorization of the likelihood of the data is given by

$$\mathcal{L}(O) \;=\; P(Y \mid A, W)P(A \mid W)P(W).$$

We refer to $P(W)$ and $P(Y \mid A, W)$ as the Q portion of the likelihood, $Q = (Q_W, Q_Y)$, and $P(A \mid W)$ as the g portion of the likelihood. Further define

$$\begin{aligned} \bar{Q}_0(A, W) &\equiv E_0(Y \mid A, W) \\ g_0(1 \mid W) &\equiv P_0(A = 1 \mid W) \end{aligned}$$

where the subscript '0' denotes the truth, and a subscript 'n' will denote the corresponding quantity estimated from data. $P_0(W)$ is estimated by the empirical distribution on W, the non-parametric MLE. $\bar{Q}_n(A, W)$ can be obtained by regressing Y on A and W. For some applications g_0 may be known, (e.g., treatment assignment in randomized controlled trials), so that consistent estimation will be guaranteed. It has been shown that estimation of g_0 leads to increased efficiency even when the true g_0 is known (van der Laan and Robins, 2003).

The parameter of interest of the probability distribution P_0 of O is therefore defined non-parametrically as $\psi_0 = E_W(E(Y \mid A = 1, W) - E(Y \mid A = 0, W))$. Under the appropriate causal graph assumptions ψ_0 corresponds with the G-computation formula for the marginal additive causal effect.

The probability distribution/density of O can be factored as $P_0(O) = Q_0(O)g_0(A \mid W)$, where $Q_0(O) = Q_{Y0}(Y \mid A, W)Q_{W0}(W)$ and $g_0(1 \mid W) = P_0(A = 1 \mid W)$. We used the notation Q_Y for a conditional distribution of Y, given A, W, and Q_W for the marginal distribution of W. For notational convenience, let $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ be the true conditional mean of Y, given A, W, which is thus a parameter of Q_{Y0}. We note that $\psi_0 = \Psi(Q_0)$ only depends on the data generating distribution P_0 through its Q_0-factor. The targeted maximum

likelihood estimator of ψ_0 is a particular substitution estimator

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

where $\bar{Q}_n(A, W)$ is an estimated conditional mean of Y given A, W, and the marginal distribution Q_{W0} is estimated with its empirical probability distribution.

Targeted maximum likelihood estimation involves obtaining an initial estimate of the true conditional mean of Y given A and W, and subsequently fluctuating this estimate in a manner designed to reduce bias in the estimate of the parameter of interest. Let $\bar{Q}_n^0(A, W)$ be the initial estimate of the true conditional mean $\bar{Q}_0(A, W)$. For example, if Y is binary, then one constructs a parametric (least favorable) model $\text{logit}(\bar{Q}_n^0(\epsilon)(A, W)) = \text{logit}(\bar{Q}_n^0 + \epsilon H^*)$, fluctuating the initial estimate $\bar{Q}_n^0$, where ϵ is the fluctuation parameter. The function $H^*(A, W)$, known as the "clever covariate", depends on the treatment assignment mechanism g_0, and is given by

$$H^*(A, W) = \frac{I(A = 1)}{g_0(1 \mid W)} - \frac{I(A = 0)}{g_0(0 \mid W)}.$$

The theoretical basis for this choice of clever covariate is given in van der Laan and Rubin (2006a). In particular, it has the bias-reduction property that if one estimates ϵ with the parametric maximum likelihood estimator, and one sets $\bar{Q}_n^1$ equal to the resulting update, then the resulting substitution estimator $\Psi(\bar{Q}_n^1)$ is asymptotically unbiased, even if the initial estimator $\bar{Q}_n^0$ is inconsistent. These results indicate that estimating g_0 is crucial for reducing bias.

1.1.1 Influence curve-based inference

TMLEs are asymptotically normally distributed with mean $\mu = \psi_0$ and variance σ^2/n, where σ^2 is the variance of the influence curve for $\Psi(\bar{Q})$. For the parameter of interest specified above, σ^2 is estimated from the data as:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\widehat{IC}^2(O_i)\\ \widehat{IC}(Q_n^*, g, \Psi(Q_n^*)) &= H^*(A, W)(Y - Q_n^*(A, W)) + Q_n^*(1, W) - Q_n^*(0, W) - \psi_n(Q_n^*).\end{aligned}$$

Ninety-five percent confidence intervals are calculated as $\psi_n(Q_n^*) \pm 1.96\hat{\sigma}/\sqrt{n}$. A test statistic can be used to test a null hypothesis of the form $H_0 : \psi_0 = 0$:

$$T = \frac{\psi_n}{\sqrt{\hat{\sigma}^2/n}}.$$

1.2 Further Remarks

The fundamental themes underlying C-TMLE were present in van der Laan and Rubin's seminal paper: building a sequence of TMLEs that are increasing in likelihood, targeting nuisance parameter estimation towards the parameter of interest, and using cross-validation to select among TMLEs. For example, this quote from Section 2.2 of the technical report version of the paper (van der Laan and Rubin, 2006b) describes constructing a sequence of TMLEs with increasing likelihood:

> As a potentially useful practical modification of this targeted MLE algorithm $\Phi^*(P_n)$ for ψ_n we suggest that at each step one does not necessarily need to select the maximizer $\epsilon(P_n \mid p_n^i)$, but instead one might simply select an ϵ so that $P_n \log p_n^k(\epsilon) > P_n \log p_n^o$, thereby still guaranteeing that the likelihood increases at each step. The important property driving the asymptotics of the resulting estimator is that the algorithm is such that for k converging to infinity the likelihood increases at each step, and (as a consequence) the maximizer of $\epsilon \to P_n \log p_n^k(\epsilon)$ converges to zero so that in the limit $\lim_k P_n D(p_n^k)) = 0$.

where $\Phi(P_n)$ is a density estimator, $\Phi^*(P_n)$ is a *targeted* density estimator, and $D(p)$ is the efficient influence curve.

Section 2.3 discusses targeting the nuisance parameter estimate:

> Although the efficient influence curves dependence on the [nuisance parameter] is such that the properties of [the nuisance parameter estimate] only affects the second order terms in the resulting targeted ML estimator of ψ_0 (see e.g. van der Laan and Robins (2002)), it is still of practical interest to also make the estimator of [the nuisance parameter] targeted towards estimation of $D(p_0)$.

Section 2.5 recommends cross-validation to select between two targeted (unbiased) estimators:

> The log likelihood loss function provides a sensible criteria for comparing densities which perform equally well with respect to the parameter of interest, ψ_0.

Paraphrasing Section 2.5: Consider two targeted estimators based on two different initial estimates of Q. Both solve the efficient influence curve equation for the target parameter, and are therefore unbiased for the target parameter. But in finite samples we don't necessarily have that $\Psi(p_n^1) = \Psi(p_n^2) = \psi_0$. Instead, we know that $P_0(D_p) \approx \Psi(p_0) - \Psi(p_n) = \psi_0 - \psi_n$. Any deviation of $\psi_0 = \psi_n$ from 0 is due to nuisance parameters needed to identify D. A final quote provides the rationale for utilizing cross-validation (for theoretical justification see also van der Laan and Dudoit (2003)):

> For example, consider two possible increases in fit of the nuisance parameters needed to fit $D(p_0)$, but suppose that one of the fits results in a large gain of the log-likelihood during the targeted MLE algorithm, while for the other fit the targeted MLE algorithm yields only a small increase in log-likelihood. Then a comparison of the log-likelihood for the two targeted fits will select the increase in nuisance parameter fit which results in the subsequent maximal increase in log-likelihood during the targeted MLE algorithm. That is, ... the criteria rewards increases in fits of the density which are directly relevant for estimation of the parameter of interest.

Each of these key ideas is presented in a separate context in the 2006 technical report. The theory of collaborative double robustness ties the pieces together.

Chapter 2

Implications of Collaborative Double Robustness of the Efficient Influence Curve

2.1 Introduction

Under regularity conditions a double robust estimator is consistent and asymptotically normal if either the outcome regression model estimating $\bar{Q}_0$ or the censoring mechanism model for g_0 is correctly specified (Scharfstein et al., 1999; Neugebauer and van der Laan, 2005). The theory behind collaborative double robustness reveals that there can exist combinations $(\bar{Q}_n, g_s)$ that provide consistent parameter estimation even when neither is correctly specified (van der Laan and Gruber, 2010). This fundamental property of the efficient influence curve applies to all estimators that solve the efficient influence curve estimating equation, and has implications for effective nuisance parameter estimation when the goal is optimizing the bias/variance tradeoff for the parameter of interest.

The data consists of n independent and identically distributed copies of $O = (W, A, Y) \sim P_0$. The likelihood of the data factorizes as

$$\mathcal{L}(O) = \underbrace{P(Y \mid A, W)}_{Q_Y} \underbrace{P(A \mid W)}_{G} \underbrace{P(W)}_{Q_W}.$$

Consider estimation of the additive treatment effect parameter of a binary point treatment,A, on an outcome Y, adjusted for baseline covariates W, $\psi_0 = E_0(Y(1) - Y(0))$. The Let $Q = (Q_Y, Q_W)$. The efficient influence curve for this parameter is given by,

$$D(\psi_0, G, Q) = \left(\frac{2A-1}{g(A \mid W)}\right)(Y - \bar{Q}(A, W)) + \bar{Q}(1, W) - \bar{Q}(0, W) - \psi_0.$$

and $P_0 D(\psi_0, G, Q) = 0$ if an estimate of either $\bar{Q}_0$ or g_0 is consistent.

We next define the collaborative double robustness (C-DR) property of the efficient influence curve, referring to van der Laan and Gruber (2010) for technical details. Define $\mathcal{G}(Q, P_0)$ as the

set of all true conditional distributions of A given W^s, where W^s is a function of W such that $Q - Q_0$ depends on W through a reduction W^s. Then, for any $g \in \mathcal{G}(Q, P_0), P_0 D(\psi_0, g, Q) = 0$. The implication of this result for all estimators that rely on D as an estimating function (including IPTW and AIPTW) is that the estimator remains unbiased when and $g_s \in \mathcal{G}(Q, P_0)$ is substituted for g_0,

$$P_0 D(Q, g_s) = 0 \rightarrow \Psi(Q) = \Psi(Q_0).$$

For TMLEs, this property implies that further targeting an already targeted TMLE, Q_n^* using $g_s' \in \mathcal{G}(Q, P_0)$ will not bias the estimate.

We use the term "sufficient" to describe the conditional distributions $g_s \in \mathcal{G}(Q, P_0)$, because in conjunction with Q, each $g_s \in \mathcal{G}(Q, P_0)$ is sufficient for consistent estimation of ψ_0. $\mathcal{G}(Q, P_0)$ includes

1. g_0, the true conditional distribution of A given W,
2. any conditional distribution of A given W^s that conditions on at least $(Q - Q_0(0, W), Q - Q_0(1, W))$,
3. any conditional distribution that conditions on variables in addition to those included in (1) or (2).

In the next section we demonstrate the C-DR property for IPTW, AIPTW and TMLE in the particular case that the initial estimate of Q is non-informative. Subsequent chapters describe C-TMLE in more detail, and demonstrate the C-DR property under misspecified and correctly specified models for Q.

2.2 Demonstration of collaborative double robustness

The following examples illustrate this property when the inverse probability of treatment weighted estimator (IPTW), and TMLE are applied to estimate an additive treatment effect (Example 1), and for application of AIPTW (Example 2). The IPTW and TMLE estimators for the additive treatment effect are defined as

$$\psi_n^{IPTW} = \frac{1}{n} \sum_{i=1}^{n} \left[I(A_i = 1) - I(A_i = 0) \right] \frac{Y_i}{g_n(A_i, W_i)}$$

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

where $\bar{Q}_n^*(A, W)$ is a targeted estimate of $E_0(Y \mid A, W)$. IPTW estimation does not incorporate any estimate of $\bar{Q}_0$. TMLE can be made comparable to IPTW by setting $\bar{Q}_n^0(A, W) = 0$, thus forcing TMLE to rely entirely on g_n.

2.2.1 Example 1: TMLE and IPTW

The true data generating distribution is given by

$$\begin{aligned} W_1, W_2, W_3 &\sim Bernoulli(0.5) \\ A &= expit(-2 + W_1 + 2W_2 + 3W_3) \\ \bar{Q}_0(A, W) &= A + W_1. \end{aligned}$$

Note that g_0 depends on (W_1, W_2, W_3), however W_1 is the only confounder, and therefore an unbiased estimate of ψ_0 can be obtained by adjusting only for W_1. This indicates that $g_0(1 \mid W_1, W_2, W_3) \in \mathcal{G}(0, P_0)$ and $g_s(1 \mid W_1) = E_0(g_0(1 \mid W)|W_1) \in \mathcal{G}(0, P_0)$. Table 2.1 lists stratum-specific treatment assignment probabilities for each of these conditional distributions. We observe

Table 2.1: Example 1: Stratified treatment assignment probabilities.

W_1	W_2	W_3	g_0	g_s
0	0	0	0.12	0.575
0	0	1	0.73	0.575
0	1	0	0.50	0.575
0	1	1	0.95	0.575
1	0	0	0.27	0.715
1	0	1	0.88	0.715
1	1	0	0.73	0.715
1	1	1	0.98	0.715

that although stratum-specific treatment assignment probabilities vary, the conditional distribution of A given the lone confounder, W_1, is the same under g_0 and g_s, $P_{g_0}(A \mid W_1) = P_{g_s}(A \mid W_1)$. An imbalance of non-confounders between treatment and control groups does not bias an estimate of the treatment effect, thus the distribution of A conditional on non-confounders is irrelevant.

Effect estimates shown in Table 2.2 were obtained for 500 samples of size $n = 10^6$ to illustrate asymptotic estimator behavior. Adjusting for g_s removes bias as effectively as adjusting for g_0, and can be more efficient. (TMLE is not semi-parametric efficient in this setting because $\bar{Q}_0$ is misspecified.) For these data, estimating g_s is 3 to 25 times more efficient than estimating g_0.

There are more sufficient conditional distributions than just these two. For example, when $\bar{Q}_n^0 = 0$ it is necessary for g_s to condition on all confounders in g_0. Predicted probabilities based on g_s equal an expectation of g_0 over non-confounders that are not conditioned upon. Non-confounders in g_0 can either be conditioned upon or integrated over, which confirms the non-uniqueness of $g_s \in \mathcal{G}(Q, P_0)$. In addition, distributions that condition on all confounders plus additional covariates beyond the true confounders are sufficient, however finite sample performance suffers when their inclusion leads to violations of the positivity assumption. Table 2.3

Table 2.2: IPTW and TMLE estimates using different models for g, setting $Q_n^0 = 0$, varying sample sizes.

	$g_0(A\|W1, W2, W3)$		$g_s(A\|W1)$	
	IPTW	TMLE	IPTW	TMLE
bias	-9.61e-05	-1.14e-03	-9.76e-05	-9.76e-05
var	1.71e-05	1.44e-04	5.11e-06	5.11e-06
MSE	1.70e-05	1.45e-04	5.11e-06	5.11e-06
relMSE*	1	1	0.30	0.04

*relative to g_0 for that estimator

Table 2.3: Example 1: IPTW and TMLE results using g_s and g_s'.

	g_s		g_s'	
	IPTW	TMLE	IPTW	TMLE
$n = 10^6$				
bias	-9.8e-05	-9.8e-05	-4.1e-05	-4.1e-05
var	5.1e-06	5.1e-06	6.0e-06	6.0e-06
MSE	5.1e-06	5.1e-06	5.9e-06	5.9e-06
relMSE*	0.30	0.04	0.35	0.04
$n = 500$				
bias	-3.9e-04	-3.9e-04	2.2e-03	2.2e-03
var	9.3e-03	9.3e-03	1.2e-02	1.2e-02
MSE	9.3e-03	9.3e-03	1.2e-02	1.2e-02
relMSE*	0.26	0.01	0.33	0.01

*relative to g_0 for that estimator at the same sample size

provides a comparison of results for g_s defined above , and an additional conditional distribution, $g_s' \in \mathcal{G}(Q, P_0) \equiv P(A = 1 \mid W_1, W_2)$. We see that bias reduction and improved efficiency in finite samples ($n = 500$) mirrors asymptotic behavior.

Table 2.4: Example 2: AIPTW, $\bar{Q}_n^0$ correctly specified and misspecified.

$(n = 500)$	Bias	Var	MSE
Q_{cor}			
OLS	-0.0019	0.0091	0.0091
AIPTW, g_0	0.0004	0.0107	0.0106
AIPTW, g_s	-0.0021	0.0092	0.0092
Q_{mis}			
OLS	-0.2354	0.0200	0.0753
AIPTW, g_0	0.0013	0.0116	0.0116
AIPTW, g_s	-0.0020	0.0094	0.0093

2.2.2 Example 2: AIPTW

AIPTW was applied to estimate ψ_0 from 500 datasets drawn from a data generating distribution defined as follows:

$$\begin{aligned} W_1, W_2, W_3 &\sim Bernoulli(0.5) \\ A &= expit(0.2 - 0.5W_1 + 0.4W_2 + 1.2W_3) \\ \bar{Q}_0(A, W) &= A + 2W_1 + W_2 \end{aligned}$$

The AIPTW estimator of the additive treatment effect is given by

$$\psi^{AIPTW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i = 1)}{g(1 \mid W_i)} - \frac{I(A_i = 0)}{g(0 \mid W_i)}\right)(Y_i - \bar{Q}(A_i, W_i)) + \bar{Q}(1, W_i) - \bar{Q}(0, W_i).$$

TMLE and AIPTW are both DR, which indicates there is no residual bias in an estimated ψ_n based on a correctly specified fit for $\bar{Q}_0(A, W)$. Consider a specific misspecified estimate, $\bar{Q}_{n,mis}(A, W) = E(Y \mid A, W_2)$. Residual bias is due to confounding by (W_1, W_2), therefore $\mathcal{G}(\bar{Q}_{n,mis}, P_0)$ includes $g_0 = P(A = 1 \mid W_1, W_2, W_3)$, $g_s = P(A = 1 \mid W_1, W_2)$, and more.

Table 2.4 shows results of applying AIPTW to 500 samples of size $n = 500$. Estimates were obtained using a correctly specified regression model to estimate $\bar{Q}_0$, and also using the misspecified model for $\bar{Q}_0$. g_0 and g_s defined above are both in $\mathcal{G}(\bar{Q}_{n,mis}, P_0)$. Notice that when the initial fit for $\bar{Q}_0$ is correctly specified $(\bar{Q}_{n,cor} - \bar{Q}_0) = 0$, thus $\mathcal{G}(Q_{n,cor})$ includes all conditional distributions of A given W, and in particular, includes all the conditional distributions that are in $\mathcal{G}(\bar{Q}_{n,mis}, P_0)$. This is not just a small finite sample result. Sample size was increased to $n = 10^6$, and estimates were obtained for 500 samples (Table 2.5).

Results confirm that g_s that adjusts only for $(\bar{Q}_0 - \bar{Q})$ is sufficient to remove bias. Also note AIPTW using g_s is super-efficient. That is to say, when the estimate of $\bar{Q}_0$ is correctly specified, AIPTW variance is less than the semi-parametric efficiency bound in the semi-parametric model,

Table 2.5: Example 2: AIPTW, $\bar{Q}_n^0$ correctly specified and misspecified.

$(n = 10^6)$	Bias	Var	MSE	rel Eff
Q_{cor}				
OLS	$7.41e-05$	$4.66e-06$	$4.66e-06$	0.89
AIPTW, g_0	$2.34e-05$	$5.27e-06$	$5.26e-06$	1
AIPTW, g_s	$8.02e-05$	$4.71e-06$	$4.70e-06$	0.89
Q_{mis}				
OLS	$-2.30e-01$	$8.97e-06$	$5.30e-02$	10^4
AIPTW, g_0	$-1.47e-05$	$5.26e-06$	$5.25e-06$	1.00
AIPTW, g_s	$1.33e-05$	$4.69e-06$	$4.69e-06$	0.89

$\mathcal{M}$ where it is not known that W_3 is a non-confounder. In an alternate, restricted, model space where such knowledge is known, g_s defined above would be considered g_0, and AIPTW estimates based on g_s would be considered efficient, but not super-efficient. This distinction explains the super-efficient behavior of C-TMLE in simulations in subsequent chapters.

In summary, understanding that any $g_s \in \mathcal{G}(Q, P_0)$ provides sufficient bias reduction allows us to tailor nuisance parameter estimation to improve the bias/variance tradeoff for ψ_n. A known g_0 that does not lead to violations of the positivity assumption will always asymptotically remove residual bias (e.g., treatment assignment probabilities in randomized controlled trials). At other times, the use of g_s may improve identifiability of ψ_0, g_s may be easier to estimate than g_0, and effective nuisance parameter estimation procedure should exploit the knowledge that g can be fit based on $(Q_0 - Q)$. The likelihood for g is not the best guide, nor is an ad hoc procedure without good asymptotic properties ideal. The C-TMLE algorithm presented in Chapter 4 is a principled approach with a strong theoretical foundation.

2.3 Fluctuating an already targeted Q_n^*

Consider Q_n^*, a TMLE that was obtained by fluctuating some initial Q_n^0 with g_s known to be a member of $\mathcal{G}(Q_n^0, P_0)$, such that $P_0 D(Q_n^*, g_s) = 0$, and thus $\Psi(Q_n^*)$ is unbiased for ψ_0. Attempting to further target Q_n^* with a parametric fluctuation based on any $g \in \mathcal{G}(Q_n^0, P_0)$ will not bias the estimator. To see why, consider an ϵ fluctuation of $Q_n^*(A, W)$, where ϵ is fit by maximum likelihood such that $f(\epsilon) = argmin_\epsilon \sim P_0 L(Q_n^*(\epsilon))$, and assuming the existence a local maximum.

$$Q_n^{**}(A, W) = Q_n^*(A, W) + \epsilon_1 H_{g_s'}^*(A, W)$$

Recall Q_n^* is already targeted, and therefore $f'(\epsilon = 0) = P_0 D(Q_n^*, g_s) = 0$. By the assumption of a local maximum, $\epsilon = 0$ is therefore a unique solution, such that $Q_n^{**} = Q_n^*$, and $\Psi(Q_n^{**}) = \Psi(Q_n^*)$, provided $P_n = P_0$, i.e., n is sufficiently large. Example 3 demonstrates this behavior both at large

sample size ($n = 10^6$) and in smaller samples ($n = 500$) for estimating the additive treatment effect parameter using data generated as in Example 1 above,

$$\begin{aligned} W_1, W_2, W_3 &\sim Bernoulli(0.5) \\ A &= expit(-2 + W_1 + 2W_2 + 3W_3) \\ \bar{Q}_0(A, W) &= A + W_1. \end{aligned}$$

We set $\bar{Q}_n^0 = 0$, and define two sufficient conditional distributions of A given W, $g_{s1} = P(A = 1 \mid W_1)$, $g_{s2} = P(A = 1 \mid W_1, W_2)$. A parameter estimate $\Psi(\bar{Q}_n^*)$ is obtained by applying the mapping, Ψ, to a targeted estimate of $\bar{Q}_0$, $\bar{Q}_n^* = \bar{Q}_n^0 + \epsilon_1 H^*(g_{s1})$, with ϵ_1 fit by maximum likelihood. A second parameter estimate, $\Psi(\bar{Q}_n^{**})$ is obtained by apply the mapping Ψ to a further targeted estimate of $\bar{Q}_0$, $\bar{Q}_n^{**} = \bar{Q}_n^* + \epsilon_2 H^*(g_{s2})$, ϵ_2 also fit by maximum likelihood. The column labeled $\bar{\epsilon}_1$ in Table 2.6 shows the mean value of ϵ_1 over 500 samples at each sample size, $\bar{\psi}_{n1}$ is the mean estimate of ψ_0, ($\psi_0 = 1$). The column labeled $\bar{\epsilon}_2$ lists the mean value of ϵ_2 over the 500 replicates, and $\bar{\psi}_{n2}$ is the mean parameter estimate. As expected, ψ_{n1} is unbiased, $\epsilon_2 \approx 0$, even at relatively small sample size, and ψ_{n2} remains unbiased.

Table 2.6: Example 3: Further targeting a TMLE.

	$\mathbf{g_{s1}}$		$\mathbf{g_{s2}}$	
n	$\bar{\epsilon}_1$	$\bar{\psi}_{n1}$	$\bar{\epsilon}_2$	$\bar{\psi}_{n2}$
10^6	0.222	1.000	0.000011	1.000
500	0.221	0.998	0.000036	0.999

Chapter 3

Targeted Maximum Likelihood Estimation for Bounded Continuous Outcomes

3.1 Introduction

This chapter describes a recently developed targeted maximum likelihood estimator for estimating a causal effect of a binary treatment on a continuous outcome. This estimator is more robust than a previously presented TMLE procedure when there is sparsity in the data that decreases the identifiability of the parameter of interest.

Sparsity is defined as low information in the dataset 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: $I = n/var(D^*(O))$, 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 times the variance of an asymptotically efficient estimator converges to 1 when the sample size converges to infinity, 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.

Section 3.2 provides background on the application of TMLE methodology in the context of sparsity, and its power relative to other semi-parametric efficient estimators by being a substitution estimator respecting global constraints of the semi-parametric 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 semi-parametric efficient *substitution* estimators, and for a careful choice of fluctuation function for the targeted MLE step that fully respects these global constraints. A rigorous demonstration of the proposed targeted MLE of the causal effect of a binary treatment on a bounded continuous

outcome follows, and it is contrasted to a targeted MLE that makes use a fluctuation function that does not respect the bounds.

Simulation studies described in Section 3.3 compare the new TMLE estimator of the causal effect, which relies on a logistic fluctuation of an initial density estimate, with the traditional TMLE estimator, with and without sparsity in the data. Results for other commonly applied estimators, the inverse-probability-of-treatment weighted estimator (IPTW) (Hernan et al., 2000b; Robins, 2000b), a double robust augmented IPTW estimator (AIPTW) (Robins and Rotnitzky, 2001; Robins et al., 2000; Robins, 2000a) that is efficient but not a substitution estimator, and the maximum likelihood substitution estimator according to a parametric model (MLE) (Robins, 1986) are also presented.

3.2 TMLE for causal effect estimation on a continuous outcome

The targeted MLE is a semi-parametric efficient substitution estimator of a target parameter $\Psi(P_0)$ of a true distribution $P_0 \in \mathcal{M}$, based on sampling n i.i.d. $O_1, \ldots, O_n$ from P_0. Here P_0 is known to be an element of a semi-parametric statistical model $\mathcal{M}$. We will start with providing a succinct summary of how TMLE works, see van der Laan et al. (2009) for details.

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 $L(Q)(O)$ so that $Q_0 = \arg\min_{Q \in \mathcal{Q}} E_0 L(Q)(O)$, where $\mathcal{Q} = \{Q(P) : P \in \mathcal{M}\}$. Thirdly, one uses minimum loss-based learning, such as super learning (van der Laan et al., 2007), 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 $Q_{ng}^0(\epsilon)$, possibly indexed by nuisance parameter $g_0 = g(P_0)$, so that

$$\frac{d}{d\epsilon} L(Q_{ng}^0(\epsilon))(O)\Big|_{\epsilon=0} = D^*(Q_n^0, g)(O), \tag{3.1}$$

where $D^*(Q_0, g_0)$ is the canonical gradient/efficient influence curve of $\Psi : \mathcal{M} \to \mathbb{R}$ at P_0. Fifthly, one computes the amount of fluctuation

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

where g_n is an estimator of the unknown nuisance parameter g_0. This yields an update $Q_n^1 = Q_{ng_n}^0(\epsilon_n)$. This updating of an initial estimator Q_n^0 into a next Q_n^1 is iterated till convergence resulting in a Q_n^*. Since at the last step the amount of fluctuation $\epsilon_n \approx 0$, 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 asymptotic efficiency of $\Psi(Q_n^*)$: recall that an estimator is efficient if and only if it is asymptotically linear with influence curve equal to the efficient influence curve $D^*(Q_0, g_0)$. Finally, the targeted MLE of ψ_0 is the substitution estimator $\Psi(Q_n^*)$.

Thus we see that the targeted MLE involves constructing a parametric model $Q_n^0(\epsilon)$ through the initial estimator Q_n^0 with parameter ϵ representing an amount of fluctuation of the initial estimator, where the score of this fluctuation model at $\epsilon = 0$ equals the efficient influence curve. The latter constraint can be satisfied by many parametric models, since it represents only a local constraint of its behavior at zero fluctuation. However, it is very important that the fluctuations stay within the model for the observed data distribution, even if the parameter can be defined on fluctuations that fall outside the assumed observed data model. In particular, in the context of sparse data, a violation of this property can heavily affect the performance of the estimator.

One important strength of the semi-parametric efficient targeted MLE relative to the alternative semi-parametric efficient estimating equation methodology (van der Laan and Robins, 2003) is that it does respect the global constraints of the observed data model since 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 thereby often ignores important global constraints of the observed data model, though Tan (2008) introduces a non-parametric likelihood based approach to constructing a double robust estimator that is not a substitution estimator, and offers a comparison with other estimators, including TMLE that is not constrained to remain within the bounds of the observed data model. Ignoring constraints comes at a price in the context of sparsity. Indeed, simulations have confirmed this gain of targeted MLE relative to the efficient estimating equation method in the context of sparsity (Stitelman and van der Laan, 2010), and it is again demonstrated in the simulations described below. However, if the targeted MLE starts violating this 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 a fluctuation model is used that is guaranteed to stay within the desired observed data model.

To demonstrate this important consideration of selecting a valid fluctuation model in the construction of targeted MLE, we consider the problem of estimating a 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. Under non-parametric structural equation model (NPSEM) $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(W, A, U_Y)$ with a structure on the exogenous variables $U = (U_W, U_A, U_Y)$ satisfying the no unmeasured confounders assumption ($A \perp Y(a) \mid W$ for the counterfactuals $Y(a)$ defined by this NPSEM), the additive causal effect $E(Y(1) - Y(0))$ can be identified from the observed data distribution through the following 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 conditional means of unbounded, or very heavy tailed, outcomes requires very large data sets.

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

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

We note that

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

An estimate, limit distribution, and confidence interval for $\Psi^*(P_0)$ is now immediately mapped into an estimate, limit distribution, and confidence interval for $\Psi(P_0)$, by simple multiplication by $(b - a)$. As a consequence, without loss of generality, we can assume $a = 0$ and $b = 1$ so that $Y \in [0, 1]$.

The efficient influence curve of the statistical parameter $\Psi : \mathcal{M} \to \mathbb{R}$, defined on a non-parametric statistical model $\mathcal{M}$ for P_0, at the true distribution P_0, is given by

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

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

We will now define a targeted MLE of $\Psi(Q_0)$ as follows. Let $\bar{Q}_n^0$ be an initial estimator of $\bar{Q}_0(W, A) = E(Y \mid A, W)$ with predicted values in $(0, 1)$. In addition, we estimate P_W with the empirical distribution of W. Let Q_n^0 denote the resulting initial estimator of Q_0. The targeted MLE step will also require an estimator g_n of $g_0 = P_{A|W}$. Only the conditional mean $\bar{Q}_n^0$ will be modified by the targeted MLE procedure defined below: this makes sense since the empirical distribution of W is already a non-parametric maximum likelihood estimator so that no bias gain with respect to the target parameter will be obtained by modifying it.

We can represent the estimator $\bar{Q}_n^0$ 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 fluctuation model

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

with parameter ϵ, indexed by a function

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

Equivalently, we can write this as $\text{logit}\bar{Q}_n^0(\epsilon) = \text{logit}\bar{Q}_n^0 + \epsilon h(g_n)$.

Consider now the following loss function for $\bar{Q}_0$:

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

Note that this is the log-likelihood of the conditional distribution of a binary outcome Y, but now extended to continuous outcomes in $[0,1]$. (See also Wedderburn (1974), McCullagh (1983) for earlier use of logistic regression for continuous outcomes.) It is thus known that this loss function is a valid loss function for the conditional distribution of a binary Y, but we need that it is a valid loss function for a conditional mean of a continuous $Y \in [0,1]$. We have the following lemma establishing this result about this loss function.

Lemma 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 (W, A) *which map into* $(0,1)$. *In addition, define the fluctuation function*

$$logit\bar{Q}(\epsilon) = logit\bar{Q} + \epsilon h.$$

For any function h *we have*

$$\frac{d}{d\epsilon} L(\bar{Q}(\epsilon))\bigg|_{\epsilon=0} = h(W,A)(Y - \bar{Q}(W,A)).$$

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

$$-\frac{d}{d\epsilon} L(\bar{Q}_1(\epsilon))\bigg|_{\epsilon=0} = h(W,A)(Y - \bar{Q}_1(W,A)).$$

Thus, it follows that

$$E_0 h(W,A)(Y - \bar{Q}_1(W,A)) = E_0 h(W,A)(\bar{Q}_0 - \bar{Q}_1)(W,A).$$

But this needs to hold for any function $h(W,A)$, which proves that $\bar{Q}_1 = \bar{Q}_0$ a.e. □

This proves that $L(\bar{Q})$ is a valid loss function for the conditional mean $\bar{Q}_0$. Indeed, we can use $L(\bar{Q})$ as loss function to construct an initial estimator of $\bar{Q}_0$, and or use cross-validation to select among candidate targeted maximum likelihood estimators, such as in the collaborative targeted MLE procedure. For the purpose of construction of an initial estimator one could also use a minimum loss-based super learner based on the squared error loss function $L_2(\bar{Q}) = (Y - \bar{Q}(W,A))^2$, possibly with weights.

Given an initial estimator $\bar{Q}_n^0$, and our proposed fluctuation function $\bar{Q}_n^0(\epsilon)$, we have

$$\frac{d}{d\epsilon} L(\bar{Q}_n^0(\epsilon))\bigg|_{\epsilon=0} = h(g)(W,A)(Y - \bar{Q}_n^0(W,A)),$$

giving us the desired first component D_1^* of the efficient influence curve $D^* = D_1^* + D_2^*$.

Let's use the log-likelihood loss function, $-logQ_W$, as loss function for the marginal distribution of W, so that our combined loss function is given by $L(Q) = -logQ_W + L(\bar{Q})$. In addition,

we use as fluctuation of the empirical distribution Q_{Wn}, $Q_{Wn}(\epsilon_1) = (1 + \epsilon_1 D_2^*(Q))Q_{Wn}$, where $D_2^*(Q) = \bar{Q}(W,1) - \bar{Q}(W,0) - \Psi(Q)$ is the remaining component of the efficient influence curve. With these choices we indeed now have that

$$\frac{d}{d\epsilon} L(Q(\epsilon))\bigg|_{\epsilon=0} = D^*(Q,g).$$

This shows that we succeeded in defining a loss function for $Q_0 = (Q_W, \bar{Q}_0)$ and fluctuation function so that the desired derivative (3.1) indeed yields the efficient influence curve.

The MLE of ϵ_1 equals zero, so that the update of Q_{Wn} equals Q_{Wn} itself. The empirical mean of the component $D_2^* = \bar{Q}(W,1) - \bar{Q}(W,0) - \Psi(Q)$ of the efficient influence curve is always equal to zero, due to the fact that we estimate the marginal distribution of W with the empirical distribution of W.

The amount of fluctuation of ϵ for fluctuating $\bar{Q}_n^0$ is given by

$$\epsilon_n^0 = \underset{\epsilon}{\operatorname{argmin}}\ P_n L(\bar{Q}_n^0(\epsilon)).$$

This "maximum likelihood" estimator of ϵ can be computed with generalized linear regression using the binomial link, i.e. the logistic regression MLE procedure, simply ignoring that the outcome is not binary, which also corresponds with iterative re-weighted least squares estimation using weights $1/\bar{Q}(1-\bar{Q})$.

This provides us with the targeted MLE update $Q_n^1 = Q_n^0(\epsilon_n^0)$, where the empirical distribution of W did not get updated, and $\bar{Q}_n^0$ did get updated as $\bar{Q}_n^0(\epsilon_n^0)$. Iterating this procedure now defines the targeted MLE Q_n^*, but as in the binary outcome case, we have that $Q_n^2 = Q_n^1(\epsilon_n^1) = Q_n^1$ since the next MLE $\epsilon_n^1 = 0$. Thus convergence occurs in one step, so that $Q_n^* = Q_n^1$. The targeted MLE of ψ_0 is thus given by $\Psi(Q_n^*) = \Psi(Q_n^1)$. As predicted, we have that the targeted MLE Q_n^* solves the efficient influence curve estimating equation $P_n D^*(Q_n^*, g_n, \Psi(Q_n^*)) = 0$.

An inspection of this efficient influence curve,

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

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 targeted MLE $\bar{Q}_n^*$ remains bounded in $(0,1)$, so that the targeted MLE $\Psi(Q_n^*)$ fully respects the global constraints of the observed data model. On the other hand, the augmented IPTW estimator obtained by solving $P_n D^*(Q_n^0, g_n, \psi) = 0$ in ψ yields the estimator

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

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

Contrasting with targeted MLE using linear fluctuation function. Alternatively, we would employ the targeted MLE using the $L_2(\bar{Q}) = (Y - \bar{Q}(W, A))^2$ loss function, and fluctuation function $\bar{Q}^0(\epsilon) = \bar{Q}^0 + \epsilon h(g)$, so that (3.1) is still satisfied. In this case, large values of $h(g)$ will result in predicted values of $\bar{Q}^0(\epsilon_n)$ that are out of the bounds $[a, b]$. Therefore, this version of targeted MLE is not respecting the global constraints of the model, i.e., the knowledge that $Y \in [a, b]$. A comparison based on simulated data of the targeted MLE using the logistic fluctuation function and the targeted MLE using this linear fluctuation function is provided in the next section.

3.3 Simulation studies for the additive effect of a binary point treatment on a continuous outcome

Two simulation studies illustrate the effects of employing a logistic vs. linear fluctuation on TMLE estimator 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 marginal effect of a binary point treatment on the outcome, $\psi_0 = E_W(E(Y \mid A = 1, W) - E(Y \mid A = 0, W))$.

The "traditional" targeted maximum likelihood approach to estimating an additive treatment effect when the outcome is continuous is to fluctuate the initial density estimate on a linear scale. Given $\bar{Q}_n^0(W, A)$, an initial estimate of the conditional mean of Y given (W, A), the fluctuation function is defined as $\bar{Q}_n^0(\epsilon) = \bar{Q}_n^0 + \epsilon(h_{g_n})$ and the loss function $L(\bar{Q})$ is chosen to be the squared error loss function, so that we still have the required constraint (3.1). The estimate ϵ_n can be obtained by estimating ϵ with a linear regression of Y on h_{g_n}, using the initial fit, $\bar{Q}_n^0(W, A)$, as offset.

A second TMLE using the logistic fluctuation method described in Section 3.2 is also obtained. Y is transformed into $Y^* \in [0, 1]$ by shifting and scaling the values. 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, we set $a = \min(Y), b = \max(Y)$, and

$$Y^* = \frac{Y - a}{b - a}.$$

An initial estimate, $\bar{Q}_n^{0,Y^*}(W, A) = E(Y^*|W, A)$, is obtained, and then 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 $\bar{Q}_n^{0,Y^*}(W, A)$ is bounded away from 0 and 1 by truncating it at $(\alpha, (1 - \alpha))$. We used $\alpha = 0.005$ in these simulation studies, which did not yield appreciably different results than setting $\alpha = 0.001$ or $\alpha = 0.01$. The function $\bar{Q}_n^{0,Y*}$ is fluctuated on the logit scale with $\text{logit}\bar{Q}_n^{0,Y^*}(\epsilon) = \text{logit}\bar{Q}_n^{0,Y^*} + \epsilon h(g_n)$, using the same clever covariate, $h_{g_n}(W, A)$, employed in the linear fluctuation described above. Fitting ϵ is again carried out using standard software, but this time using logistic regression of Y^* on $h_{g_n}(W, A)$ with offset $logit(Q_n^{0,Y^*}(W, A))$. This results in the updated $\bar{Q}_n^{1,Y^*}$.

Fitted values for $\bar{Q}_n^{1,Y^*}(W,A)$ are mapped back to the original scale: $\bar{Q}_n^{1,Y} = \bar{Q}_n^{1,Y^*}(W,A) * (b-a) + a$. The marginal distribution is estimated with the empirical distribution of W, giving the $Q_n^* = Q_n^1 = (Q_{W,n}, \bar{Q}_n^{1,Y})$ of $(Q_W, \bar{Q}_0)$. The estimate

$$\psi_n = \Psi(Q_n^*) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^{1,Y}(W_i,1) - \bar{Q}_n^{1,Y}(W_i,0)$$

is the targeted MLE of the desired additive causal effect ψ_0.

Parameter estimates were also obtained using the augmented inverse probability of treatment weighed estimator (AIPTW),

$$\begin{aligned}\psi_n^{AIPTW} &= \frac{1}{n}\sum_{i=1}^{n}\frac{2A-1}{g_n(A_i \mid W_i)}(Y_i - \bar{Q}_n^0(W_i,A_i)) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^0(W_i,1) - \bar{Q}_n^0(W_i,0)).\end{aligned}$$

Both the targeted MLE and the augmented 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. Both the targeted MLE and the augmented IPTW estimator are asymptotically efficient if both g_n and $\bar{Q}_n^0$ are consistent.

Although the utilization of super learning is recommended in practice, in this simulation study simple parametric MLEs are used as initial estimators $\bar{Q}_n^0$ and g_n, . The purpose 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 MLE fits.

Results from two estimation methods that are not double robust and semi-parametric efficient are included as well. The maximum likelihood estimator according to a parametric model for $\bar{Q}_0$ (MLE), used as initial estimator in the targeted MLE and augmented IPTW, is included for the sake of evaluating the bias reduction step carried out by these two semi-parametric efficient procedures. Inverse probability of treatment weighted (IPTW) estimators are consistent when $g_n(A \mid W)$ is a consistent estimator of the treatment mechanism $g_0(A \mid W) = P(A=1|W)$, but are known to be inefficient. These two estimators are defined as

$$\begin{aligned}\psi_n^{MLE} &= \frac{1}{n}\sum_{i-1}^{n}\bar{Q}_n^0(W_i,1) - \bar{Q}_n^0(W_i,0), \\ \psi_n^{IPTW} &= \frac{1}{n}\sum_{i=1}^{n}(2A-1)\frac{Y_i}{g_n(A_i \mid W_i)}.\end{aligned}$$

3.3.1 Data generation

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:

$$g_0(1 \mid W) = P(A = 1 \mid W) \quad = \quad \text{logit}^{-1}(aW_1 + bW_2 + cW_3).$$

We consider two settings:

$$a_1 = 0.5, b_1 = 1.5, c_1 = -1 \text{ and } a_2 = 1.5, b_2 = 4.5, c_2 = -3.$$

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

$$\begin{aligned} Y &= \bar{Q}_0(W, A) + e,\ e \sim N(0, 1), \\ \bar{Q}_0(W, A) &= A_j + 2W_1 + 3W_2 - 4W_3. \end{aligned}$$

For both simulations the true additive causal effect equals one: $\psi_0 = 1$. In both simulations predicted values for $g_n(A \mid W)$ are bounded away from 0 and 1 by truncating at $(p, (1 - p))$, with $p = 0.01$. 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, simulation 2 poses a challenging estimation problem in the context of sparse data. Treatment assignment probabilities based on mechanism $g_{0,2}$ range from 0.047 to 0.998. These extreme values are nevertheless not uncommon for data from observational studies (see for example Dehejia and Wahba (2002); Stukel et al. (2007)).

Estimates were obtained for 1000 samples of size $n = 1000$ from each data generating distribution. Treatment assignment probabilities, $g_0(A \mid W)$, were estimated using a correctly specified logistic regression model. A correctly specified main terms regression model was used to obtain $\bar{Q}^0_{cor}(W, A)$. In addition, a misspecified initial estimate, $\bar{Q}^0_{mis}(W, A)$, was obtained by regressing Y on A.

We expect MLE estimates based on $\bar{Q}^0_{cor}$ to be unbiased and efficient, while those based on $\bar{Q}^0_{mis}$ will be biased. IPTW estimates only depend on consistent estimation of g_0, thus are identical regardless of how $\bar{Q}_0$ is estimated. For both simulations g_n is a consistent estimator, thus it is reasonable to expect unbiased IPTW estimates, with more variation in simulation 2 estimates. The targeted MLE and the augmented IPTW are known to be unbiased if g_n is consistent, and asymptotically efficient when both $\bar{Q}_0$ and g_0 are consistently estimated. Though correctly estimating g_0 will asymptotically correct for any bias due to misspecification of $\bar{Q}^0_n$, this is not guaranteed in finite samples, especially when there is sparsity. For simulation 2 we expect TMLE$_{log}$, using the logistic fluctuation, to outperform TMLE$_{lin}$, using the linear fluctuation.

3.3.2 Results

Table 3.1 reports the average estimate, bias, empirical variance, and mean squared error (MSE) for each estimator, under different specifications of the initial estimator $\bar{Q}^0_n$. In all cases g_n is consistent, and bounded at (0.01, 0.99). In simulation 1, when $\bar{Q}_0$ is correctly estimated all estimators

Table 3.1: Estimator performance for simulations 1 and 2 when the initial estimator of $\bar{Q}_0$ is correct and misspecified. Results are based on 1000 samples of size $n = 1000$.

	$\bar{Q}_0$ **correctly estimated**				$\bar{Q}_0$ **incorrectly estimated**			
	ave	bias	var	MSE	ave	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
AIPTW	1.003	0.003	0.005	0.005	1.005	0.005	0.010	0.010
TMLE$_{log}$	0.993	−0.007	0.005	0.005	0.993	−0.007	0.006	0.006
TMLE$_{lin}$	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
AIPTW	0.999	−0.001	0.023	0.023	1.708	0.708	0.298	0.798
TMLE$_{log}$	0.989	−0.011	0.037	0.037	0.722	−0.278	0.214	0.291
TMLE$_{lin}$	0.986	−0.014	0.042	0.042	−0.263	−1.263	2.581	4.173

perform quite well, though as expected, IPTW is the least efficient. However, when $\bar{Q}_0$ is incorrectly estimated, the MLE estimator is biased and has high variance relative to the other estimators. Because $g_n(A \mid W)$ is correctly specified, IPTW and AIPTW provide unbiased estimates, as do both TMLEs. TMLE$_{log}$ is on a par with TMLE$_{lin}$, as there is no sparsity in the data, and both are more efficient than any of the other estimators.

In simulation 2 all estimators except IPTW are unbiased when $\bar{Q}_0$ is correctly estimated. In this case, both TMLE estimators have higher variance than AIPTW, and all three are more efficient than IPTW, but less efficient than the parametric MLE estimator. 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, The MLE estimator 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$_{log}$, using the logistic fluctuation, has the lowest MSE of all estimators. Its superiority over TMLE$_{lin}$ in terms of bias and variance is clear. TMLE$_{log}$ also outperforms AIPTW with respect to both bias and variance, and performs much better than IPTW or MLE.

3.4 Discussion

When an estimation procedure incorporates weights, observations with large weights can heavily influence the point estimate and inflate the variance. Truncating these weights is a common approach to reducing the variance, but it generally introduces bias. The presented TMLE of an additive causal effect of a point treatment intervention, incorporating a logistic fluctuation of the initial conditional mean estimate, dampens the effect of these heavily weighted observations, thereby heavily reducing the reliance on truncation. As a substitution estimator, the proposed TMLE of the additive causal effect respects the global constraints of the observed data model. Simulation study results indicate that this approach is on a par with, and in the context of sparsity often superior to, fluctuating on the linear scale. In particular it is more robust when there is sparsity in the data, outperforming MLE, IPTW, and AIPTW.

For the sake of demonstration we considered estimation of the additive causal effect. However, the same targeted MLE, using the logistic fluctuation, can be used to estimate other point-treatment causal effects, including parameters of a marginal structural model. The newly proposed loss function also has applications in prediction of a bounded outcome, and for targeted MLE of the causal effect of a multiple time point intervention in which the final outcome is bounded and continuous. We also pointed out that the proposed fluctuation function and loss function, and corresponding targeted MLE, should also be used for continuous outcomes for which no a priori bounds are known, by simply using the minimal and maximal observed outcome values. In this way, these choices naturally robustify the targeted MLE by enforcing that the updated initial estimator will not predict outcomes outside the observed range.

The TMLE approach presented here using a logistic fluctuation of an initial estimate of the conditional mean of the continuous outcome retains all properties of targeted maximum likelihood estimators, including influence curve-based inference. The method presented here extends to collaborative targeted maximum likelihood estimation without modification.

Chapter 4

Collaborative Targeted Maximum Likelihood Estimation Methodology

4.1 Overview

Collaborative targeted maximum likelihood estimation (C-TMLE) is an extension of targeted maximum likelihood estimation (TMLE) that pursues an optimal strategy for nuisance parameter estimation. Observed data are viewed as realizations of random variables arising from some true underlying data-generating distribution, P_0. An association or causal effect corresponds with some particular parameter of P_0, and can be specified as a mapping $\Psi(P_0)$. The TMLE/C-TMLE methodology can be applied in a variety of settings, including survival analysis, gene association studies, and longitudinal data structures with time-dependent covariates. This chapter focuses on the application of C-TMLE to estimating the marginal additive effect of a dichotomous point treatment on an outcome, adjusting for pre-treatment covariates measured at baseline. This estimation problem is sufficiently rich to convey the essential elements of the procedure, and can be extended to more complex data structures.

The parametric approach to estimating this causal effect is to focus on estimating a coefficient in front of the treatment term in a regression model, where the type of regression (e.g. linear, logistic, poisson) is typically chosen for convenience. An alternative, theoretically sound, approach is to first define the parameter of interest non-parametrically, in terms of the underlying distribution, and then apply an efficient data-adaptive procedure to consistently estimate the target parameter.

The density, p_0, factorizes as $Q_0 * g_0$, where $Q_0 = P_0(Y \mid A, W)P_0(W)$ is the relevant portion of the likelihood for parameter estimation, and $g_0 = P_0(A \mid W)$, the treatment assignment mechanism, is a nuisance parameter that is vital for targeting an initial estimate of Q to reduce bias, but not required to evaluate the parameter. Recall, however, that the double robustness property of targeted maximum likelihood estimators guarantees consistent estimation of ψ_0 if either Q or g is correctly

specified, and that TMLEs achieve the semi-parametric efficiency bound when both Q and g are correctly specified. Nuisance parameter estimation is therefore an important, integral part of the targeted maximum likelihood estimation procedure.

In common with other estimators in the literature, TMLE relies on external estimation of g. The likelihood for g is the only available guide, yet not all predictors of treatment are necessarily also predictive of the outcome. This approach is inherently limited, therefore, by its inability to identify true confounders. In contrast, the C-TMLE procedure incorporates estimation of g based on the likelihood for Q, i.e., targeted towards the parameter of interest. Theory advanced in van der Laan and Gruber (2010) provides the key insight that only the portion of the nuisance parameter that is not adequately accounted for in the first stage needs to be incorporated into the second stage fluctuation. This collaborative double robustness result indicates that full bias reduction can be achieved using a sufficient nuisance parameter estimate, g_n, that targets a true conditional distribution of treatment that conditions on a reduction of the complete covariate set, representing only the covariates that can reduce residual bias. The idea is that the space for the joint distribution of (Q, g) with $g \in \mathcal{G}$ is reduced to a space for Q and a space for g that is shrunk down by the initial fit of Q. This dimension reduction yields gains in both bias and variance. The term "collaborative" refers to the fact that what constitutes the relevant portion of g_0 depends upon the bias with respect to the target parameter that remains after initial estimation of Q_0. This theoretical result has important implications for finite sample behavior:

- The C-TMLE estimator targets the residual bias, and thus can be more efficient in finite samples than the standard TMLE estimator that utilizes an estimate of the entire nuisance parameter in the targeting step. This finite sample gain is particularly striking in the context of sparse data.

- Selecting the best adjustment set is difficult when a large number of correlated baseline covariates have been measured for each subject. The C-TMLE procedure grows the adjustment set judiciously by, for example, preferring covariates most strongly associated with residual confounding over those that are highly predictive of treatment but unrelated to the outcome.

The data-adaptive C-TMLE approach to nuisance parameter estimation rests on a strong theoretical foundation, and simulation studies provide empirical evidence of its utility in practice. Given $\bar{Q}_n^0(A, W)$, an initial estimate of the conditional mean of Y given A and W, stage two of the C-TMLE procedure creates a series of candidate TMLE estimators, each obtained by fluctuating the initial $\bar{Q}_n^0$. The candidate TMLE estimators are based on a sequence of nuisance parameter estimates that grow increasingly larger, i.e., more and more non-parametric. Construction of the nuisance parameter estimates is guided data-adaptively based on the goodness-of-fit of the candidate targeted MLE possibly penalized by an estimate of the mean squared error of the estimate of the target parameter. In other words, each nuisance parameter estimate in the sequence is carefully constructed to provide the next in a series of fluctuations of the initial density estimate, and each fluctuation is carried out to create a series of candidate TMLE estimators. The C-TMLE estimator is defined as the best among the set of candidate TMLE estimators as chosen by likelihood-based

cross-validation. The procedure can be carried out for multiple stage one estimators, and cross-validation can choose among TMLE candidates indexed by stage one and stage two estimators.

For simplicity the remainder of the chapter describes the procedure for a single stage one estimator, and the algorithm is first described for a continuous outcome, using a linear fluctuation. The extension to binary outcomes is straightforward, with logistic regression replacing linear regression to obtain $\bar{Q}_n^0$, and the fluctuation carried out on the logit scale. The logistic fluctuation is also recommended for bounded continuous outcomes. Chapter 3 describes how this choice of fluctuation constrains the parameter estimate to remain within the bounds of the semi-parametric model. C-TMLE should also respect these global constraints. Simulation studies at the end of this chapter demonstrate that C-TMLE's parsimonious approach to nuisance parameter estimation in conjunction with the logistic fluctuation is particularly advantageous when analyzing sparse data.

4.2 Collaborative targeted maximum likelihood estimation

We are interested in estimating the marginal additive treatment effect of a point treatment on an outcome, given a data set containing n independent and identically distributed observations, $O_1, \ldots, O_n$, of a random variable $O = (W, A, Y)$, where W is a set of baseline covariates, A is a treatment variable, and Y is the outcome variable. For simplicity we initially focus on a continuous Y and binary A, $A = 1$ denotes treatment, and $A = 0$ denotes control. The parameter of interest of the probability distribution P_0 of O is defined non-parametrically as $\psi_0 = E_W\{E(Y \mid A = 1, W) - E(Y \mid A = 0, W)\}$. Under the appropriate causal graph assumptions ψ_0 corresponds with the G-computation formula for the marginal additive causal effect.

The probability distribution/density of O can be factored as $P_0(O) = Q_0(O)g_0(A \mid W)$, where $Q_0(O) = Q_{Y0}(Y \mid A, W)Q_{W0}(W)$ and $g_0(1 \mid W) = P_0(A = 1 \mid W)$. We used the notation Q_Y for a conditional distribution of Y, given A, W, and Q_W for the marginal distribution of W. For notational convenience, let $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ be the true conditional mean of Y, given A, W, which is thus a parameter of Q_{Y0}. We note that $\psi_0 = \Psi(Q_0)$ only depends on the data generating distribution P_0 through its Q_0-factor. The targeted maximum likelihood estimator of ψ_0 is a particular substitution estimator

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

where $\bar{Q}_n(A, W)$ is an estimated conditional mean of Y given A, W, and the marginal distribution Q_{W0} is estimated with its empirical probability distribution.

Targeted maximum likelihood estimation involves obtaining an initial estimate of the true conditional mean of Y given A and W, and subsequently fluctuating this estimate in a manner designed to reduce bias in the estimate of the parameter of interest. Let $\bar{Q}_n^0(A, W)$ be the initial estimate of the true conditional mean $\bar{Q}_0(A, W)$. For example, if Y is binary, then one constructs a parametric (least favorable) model $\text{logit}(\bar{Q}_n^0(\epsilon)(A, W)) = \text{logit}(\bar{Q}_n^0(A, W)) + \epsilon H^*(A, W)$, fluctuating the

initial estimate $\bar{Q}_n^0$, where ϵ is the fluctuation parameter. The function $H^*(A, W)$, known as the "clever covariate", depends on the treatment assignment mechanism g_0, and is given by

$$H^*(A, W) = \frac{I(A = 1)}{g_0(1 \mid W)} - \frac{I(A = 0)}{g_0(0 \mid W)}. \tag{4.1}$$

The theoretical basis for this choice of clever covariate is given in van der Laan and Rubin (2006a). In particular, it has the bias-reduction property that if one estimates ϵ with the parametric maximum likelihood estimator, and one sets $\bar{Q}_n^1$ equal to the resulting update, then the resulting substitution estimator $\Psi(Q_n^1)$ is asymptotically unbiased, even if the initial estimator $\bar{Q}_n^0$ is inconsistent. These results indicate that estimating g_0 is crucial for reducing bias. However, the choice of an estimator g_n should be evaluated by how it affects the mean squared error of the resulting targeted maximum likelihood estimator $\Psi(Q_n^1)$, making it a harder and different problem than estimating g_0 itself.

Collaborative double robustness implies that if the initial estimator converges to a possibly misspecified Q, then g_n needs to only converge to a conditional distribution of A that properly adjusts for a covariate that is a function of $Q_0 - Q$. This result is intuitively a natural consequence of the fact that the clever covariate can only reduce bias if it is predictive of the outcome after taking into account the initial estimator.

A particular method for construction of a collaborative estimator g_n involves building candidate treatment mechanism estimators (propensity scores) that grow towards an asymptotically unbiased estimator of g_0. In a departure from current practice, the construction of these candidates is guided by the log-likelihood loss function for Q_0 at targeted maximum likelihood estimators implied by an initial estimator of Q_0 and a choice of estimator of g_0, thus not by the log-likelihood loss function for the conditional distribution of A given W.

Clever covariates based on these candidates give rise to a sequence of updated estimates, $\bar{Q}_n^1(\bar{Q}_n^0, g_n^1), \ldots, \bar{Q}_n^K(\bar{Q}_n^0, g_n^1, \ldots, g_n^K)$, each of which provides a candidate TMLE estimate of ψ_0. The C-TMLE estimate is the best among these candidates, as determined by V-fold Q_0-log-likelihood-based cross-validation. The estimator is defined as

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

where $\bar{Q}_n^*(A, W)$ is the targeted estimate selected using cross-validation.

4.3 The C-TMLE procedure

One particular implementation of the C-TMLE procedure uses a collaborative targeted forward selection algorithm to build nested candidate estimators $g_n^1, \ldots, g_n^K$.

Algorithm 6.1 *C-TMLE Algorithm*

Step 1. Estimate $\bar{Q}_n^0 = \hat{E}(Y \mid A, W)$

Step 2. Create candidate TMLE estimators $\bar{Q}_1^*(g_n^1), \ldots \bar{Q}_K^*(g_n^K)$ using collaborative
targeted forward selection to build candidate estimates, $g_n^1, \ldots, g_n^K$

Step 3. Select the best candidate, $\bar{Q}_n^* = \bar{Q}_k^*(g_n^k)$, using likelihood-based cross validation

Step 4. Evaluate parameter: $\psi_n = \Psi(Q_n^*)$, based on substitution of $\bar{Q}_n^*$ and the empirical distribution as estimator of the marginal distribution of W.

Step 1: Obtain an estimate $\bar{Q}_n^0$ of $\bar{Q}_0$.

A data-adaptive machine learning approach to obtaining this initial estimate is recommended. The super learner (SL) is a prediction algorithm that creates a weighted combination of predictions of many individual prediction algorithms, with weights selected using V-fold cross-validation (van der Laan et al., 2007). In practice, it is important to include algorithms in the SL library of predictors that cover different model spaces, e.g. support vector machines, splines, neural nets, etc., since the true best estimation method is unknown.

Step 2: Generate candidate second stage estimators $\bar{Q}_n^k$.

Theory requires that the sequence of g estimators grows towards and arrives at a consistent estimator of the true g_0. Building nested candidate g estimators satisfies this requirement, and ensures that for all $m < k$, g_n^k is a better empirical fit for the treatment mechanism than g_n^m. Each move in the collaborative targeted forward selection procedure incorporates a single covariate, W_k, that minimizes a loss function for Q, into the model for g. Each move improves the fit for g in a way that maximally increases the fit for the Q portion of the likelihood, the relevant portion for the parameter of interest.

One begins with the intercept model for g to construct a first clever covariate, H_1^*, used to create the first targeted maximum likelihood candidate, $\bar{Q}_n^1$.

$$\begin{aligned} g_n^1(1 \mid W) &= P(A = 1) \\ g_n^1(0 \mid W) &= P(A = 0) \\ H_1^* &= \left(\frac{I[A = 1]}{g_n^1(1 \mid W)} - \frac{I[A = 0]}{g_n^1(0 \mid W)} \right). \end{aligned}$$

Let $\bar{Q}_n^1 = \bar{Q}_n^0 + \epsilon_1 H_1^*$, where ϵ_1 is fitted by least-squares regression of Y on H_1^* with offset $\bar{Q}_n^0$, be the first targeted candidate TMLE estimator.

The second candidate TMLE estimator will be based on an updated model for g that contains the intercept and one term. The best main term is selected based on a penalized log-likelihood criterion for the targeted MLE fit. This loss function, defined as the empirical sum of squared residuals at the resulting Q_0-fit plus a penalty term proportional to the estimated variance of substitution estimator of this Q_0-fit of the target parameter, is asymptotically minimized at the true Q_0, and thereby represents a valid loss function. The variance of the substitution estimator of the target parameter is estimated using the empirical variance of efficient influence curve D^*, at the resulting Q_0-fit and the candidate g-fit.

Consider the example presented in Figure 4.1, illustrating the process of choosing the best term to add to the intercept model for g given $W = (W_1, W_2, W_3)$. Each model for g of size one gives rise to a "tentative" targeted estimate, $\bar{Q}_n^{2a}, \bar{Q}_n^{2b}, \bar{Q}_n^{2c}$.

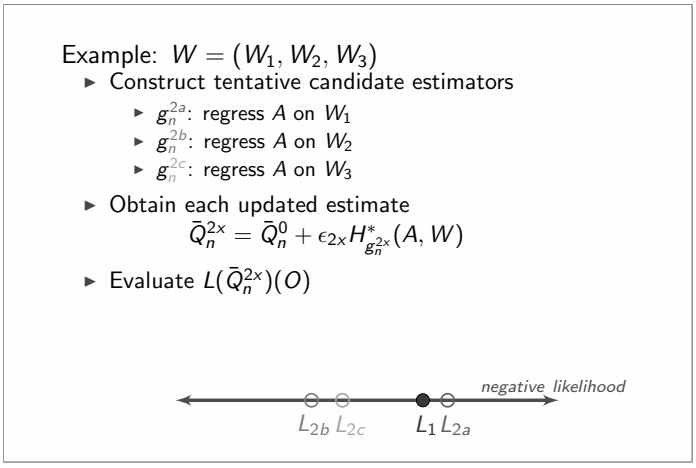

Figure 4.1: Construction of candidate TMLE estimator $\bar{Q}_n^2$.

The corresponding values of the loss functions are plotted on the number line, along with L_1, the value of the loss function for $\bar{Q}_n^1$. The addition of covariate W_2 (i.e., 2_b) to the current model for g minimizes the loss function, so the first move is the selection of W_2. This choice defines our second candidate TMLE estimator:

$$\bar{Q}_n^2 = \bar{Q}_n^0 + \epsilon_2 H^*_{g_n^2}.$$

This process continues as long as the addition of a term to the model for g increases the overall penalized log-likelihood for the resulting $\bar{Q}_0$-targeted MLE fit. In the event that no terms in the model for g increase the penalized likelihood of the resulting $\bar{Q}_0$-fit, the targeted MLE update is carried out using the most recent clever covariate, and then the process continues by fluctuating this updated estimator with a new clever covariate based on the next (larger) model for g in the series. This approach guarantees that the overall likelihood continues to increase, though the penalized likelihood may not improve.

Continuing the example from above, in Figure 4.2 we are trying to create the third candidate TMLE estimator, but find that neither the addition of W_1 nor W_3 minimizes the loss function in comparison with L_2, the value of the loss function $L(\bar{Q}_n^2)(O)$.

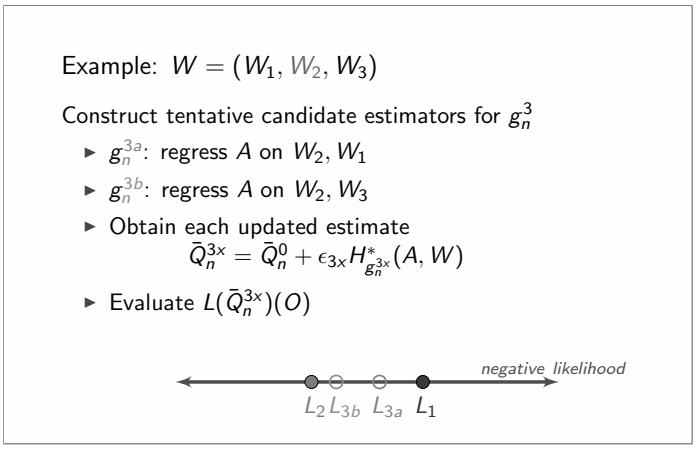

Figure 4.2: Construction of candidate TMLE estimator $\bar{Q}_n^3$, no term improves the likelihood.

Each estimator in the series of candidate TMLE estimators must represent an improvement over the previous one, yet in this example no single fluctuation of the initial estimate can achieve that goal. Therefore, the third candidate will instead be based on two fluctuations of $\bar{Q}_n^0$. Equivalently, we can say that g_n^3 is estimated in collaboration with $\bar{Q}_n^2$:

$$\begin{aligned} \bar{Q}_n^3 &= \bar{Q}_n^2 + \epsilon_3 H^*_{g_n^3}(A, W) \\ &= \bar{Q}_n^0 + \epsilon_2 H^*_{g_n^2}(A, W) + \epsilon_3 H^*_{g_n^3}(A, W) \end{aligned}$$

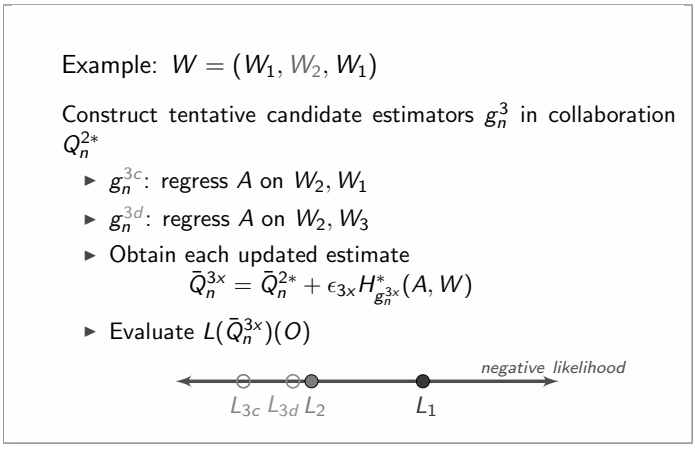

Figure 4.3: Construction of candidate TMLE estimator $\bar{Q}_n^3$ requires a second clever covariate.

Figure 4.3 shows that the addition of either covariate would further minimize the loss function. Estimator g_n^{3c} corresponding to a model containing the two terms, (W_1, W_2), is the best choice.

This targeted forward selection procedure continues as long as covariates remain that can be incorporated into the model for g. Each move gives rise to a candidate TMLE estimator, although as we've seen, each move does not result in the creation of a new clever covariate.

To state it more generally, suppose that in addition to the intercept term, m terms, ordered $1, \ldots, m$, are incorporated into the model for g, at which point no further increase of the penalized log-likelihood is possible. We define candidate estimators $\bar{Q}_n^2$ through $\bar{Q}_n^{m+1}$ as:

$$\begin{aligned} \bar{Q}_n^2 &= \bar{Q}_n^1 + \epsilon_2 H_2^* \\ \bar{Q}_n^3 &= \bar{Q}_n^1 + \epsilon_3 H_3^* \\ &\vdots \\ \bar{Q}_n^{m+1} &= \bar{Q}_n^1 + \epsilon_{m+1} H_{m+1}^* \end{aligned}$$

where the corresponding models g_n^{i+1} contains all the terms in the model for g_n^i plus one additional term, $i = 2, \ldots, m$. At this point $\bar{Q}_n^{m+1}$ is considered as a new "initial" estimate of the true regression, and the entire process starts over in order to build a second clever covariate augmenting the previous fit g_n^{m+1} used in H_{m+1}^*. To continue the example, $\bar{Q}_n^{m+2} = \bar{Q}_n^{m+1} + \epsilon_{m+2} H_{m+2}^*$. This process is iterated until all terms are incorporated into the final model for g. If the maximal number of terms that can be added is given by K, then this results in K candidate estimators $\bar{Q}_n^k$, $k = 1, \ldots, K$, corresponding with treatment mechanism estimators g_n^k, $k = 1, \ldots, K$. Note that the number of clever covariates in $\bar{Q}_n^k$ that are added to the initial estimator $\bar{Q}_n^0$ cannot be predicted, and depends on how many covariates can be added to the treatment mechanism estimator in each iteration before reaching the local maximum (not allowing a further increase of the penalized log-likelihood).

The number of clever covariates used to update the initial estimator $\bar{Q}_n^0$ depends entirely on the likelihood and cannot be pre-determined. Terms are incorporated into the model for g for a single clever covariate until there is a decrease in the likelihood. At that point the estimate is updated from $\bar{Q}_n^m \rightarrow \bar{Q}_n^{(m+1)}$ and the process iterates until all candidate TMLEs have been constructed.

These estimators $\bar{Q}_n^k$ and corresponding treatment mechanism estimators g_n^k can be represented as mappings $\hat{Q}^k$ and $\hat{g}^k$ applied to the empirical distribution P_n: $Q_n^k = \hat{Q}^k(P_n)$, $g_n^k = \hat{g}^k(P_n)$, $k = 1, \ldots, K$. These mappings $P_n \rightarrow \hat{Q}^k(P_n)$ represent our candidate estimators of the true regression $\bar{Q}_0$, and in the next step we use cross-validation to select among these candidate algorithms.

Step 3: Select the estimator that maximizes the V-fold cross-validated penalized likelihood.

The use of likelihood-based cross-validation to select the best candidate TMLE for the given stage one estimator avoids overfitting. Maximizing the penalized likelihood is equivalent to minimizing the residual sum of squares (RSS) plus a penalty term corresponding to the mean squared error (MSE), which can be decomposed into variance and bias terms.

$$k^* = \underset{k}{\operatorname{argmin}}\ cvRSS_k + cvVar_k + n * cvBias_k^2.$$

These terms are defined as follows:

$$\begin{aligned}
cvRSS_k &= \sum_{v=1}^{V} \sum_{i \in Val(v)} (Y_i - \bar{Q}_n^k(P_{nv}^0)(W_i, A_i))^2 \\
cvVar_k &= \sum_{v=1}^{V} \sum_{i \in Val(v)} D^{*2}(\bar{Q}_n^k(P_{nv}^0), g_n^k(P_n), \hat{\Psi}(\bar{Q}_n^k(P_{nv}^0)))(O_i) \\
cvBias_k &= \frac{1}{V} \sum_{v=1}^{V} \Psi(\hat{Q}^k(P_{nv}^0)) - \Psi(\hat{Q}^k(P_n)) \\
D^*(Q, g, \Psi(Q))(O) &= \frac{I[A=1] - I[A=0]}{g(A \mid W)}(Y - \bar{Q}(A, W)) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(Q)
\end{aligned}$$

where v ranging from 1 to V indexes the validation set $Val(v)$ for the vth fold, $\Psi(Q)$ is a mapping from Q to the parameter of interest, and $\bar{Q}_n^k(P_{nv}^0)$ denotes the k-th C-TMLE of $\bar{Q}_0$ applied to the corresponding training sample P_{nv}^0, containing $n(1-p)$ observations, with $p = 1/V$.

Note that the model for g is not restricted to main terms only. For example, variables can be created that correspond to higher-order terms. In addition, a categorical or continuous covariate can be split into many binary covariates, thereby allowing for more non-parametric modeling of the effect of a single covariate. When there are many covariates it might be desirable in practice to terminate the procedure before all covariates have been incorporated into the model for g, though care must be taken to ensure that none of the candidates thereby excluded from the subsequent selection process potentially maximize the penalized log-likelihood criterion. SL can be integrated into the second stage as well. A series of increasingly non-parametric propensity score SL estimates can be obtained based on different adjustment sets. These SL fits would then be used as the main terms for the stage two forward selection to build candidate g_n estimators.

4.3.1 Inference

Under appropriate conditions, C-TMLE is an asymptotically linear estimator with influence curve

$$IC(P_0) = D^*(Q_n^*, g_0, \psi_0) + IC_{Q^*} + IC_{g_0}.$$

The formula for the efficient influence curve/canonical gradient $D^*(Q, g_0, \psi_0)$ is parameter-specific (see Appendix B.2.6). For the additive treatment effect parameter this formula is given by:

$$D^* = \left(\frac{A}{g(1 \mid W)} - \frac{1-A}{g(0 \mid W)}\right)(Y - \bar{Q}(A, W)) + \bar{Q}(1, W) - \bar{Q}(0, W) - \psi_0.$$

The additional terms IC_{Q^*} and IC_{g_0} represent the contribution to the influence curve from estimating g_n in collaboration with Q_n^*. A proof and a formal statement of the conditions under which asymptotic linearity holds, are provided in van der Laan and Rose (2011). To summarize, first assume that Q_n converges to some Q^*, and g_n converges to some $g_{s_0} \in \mathcal{G}(Q^*, P_0)$, such that $P_0 D^*(Q^*, g_{s_0}) = 0$, which implies the estimator is consistent, $\Psi(Q^*) = \psi_0$. If either Q_n is a consistent estimator of Q_0, or g_n is a consistent estimator of g_0, $IC_{Q^*} = 0$. In addition, suppose the set of true confounders of the treatment effect is a subset, W^s, of W. In this case, if $\bar{Q}^* = E_0(Y \mid A, W^s)$, or if $g_{s_0} = P(A = 1 \mid W^s)$, then again $IC_{Q^*} = 0$. In other words, if estimation of either the relevant part of the Q or g portion of the likelihood independently guarantees consistency, the estimation of Q_0 does not contribute to the efficient influence curve equation. However, when this is not the case, e.g. $Q^* = E(Y \mid A, W'), W' \subset W^s$, and g_{s_0} adjusts for covariates in the difference $(W^s - W')$, $IC_{Q^*} \neq 0$. There is not a closed form solution for estimating IC_{Q^*} when $\bar{Q}_0$ is estimated semi-parametrically.

The formula for IC_{g_0} is given by:

$$IC_{g_0}(O) = -a_0 \cdot IC_\alpha(O)$$

where

$$\begin{aligned} a_0 &= P_0(Y - \bar{Q}(A,W))\overrightarrow{W}h_\alpha(A,W) \\ h_\alpha(A,W) &= \left[\frac{Ag_\alpha(0 \mid W)}{g_\alpha(1 \mid W)} + \frac{(1-A)g_\alpha(1 \mid W)}{g_\alpha(0 \mid W)}\right] \\ IC_\alpha(O) &= P_0\left[\overrightarrow{W}\overrightarrow{W}^T g_\alpha(1 \mid W)g_\alpha(0 \mid W)\right]^{-1}(A - g_\alpha(1 \mid W))\overrightarrow{W}. \end{aligned}$$

The notation $\overrightarrow{W}$ is used to denote the vector of main terms that is included in the logistic regression model g_{α_n}. Note that a_0 is a vector of the same dimension as $\overrightarrow{W}$. A derivation is provided in an Appendix.

This influence curve is estimated by its empirical analog, given by:

$$\widehat{IC}_{g_0}(O) = -a_n \cdot \widehat{IC}_\alpha(O)$$

where

$$\begin{aligned} a_n &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Q}_n(A_i,W_i))\overrightarrow{W}_i h_{\alpha_n}(A_i,W_i) \\ h_{\alpha_n}(A_i,W_i) &= \left[\frac{A_i g_{\alpha_n}(0 \mid W_i)}{g_{\alpha_n}(1 \mid W_i)} + \frac{(1-A_i)g_{\alpha_n}(1 \mid W_i)}{g_{\alpha_n}(0 \mid W_i)}\right] \\ \widehat{IC}_\alpha(O) &= \left[\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{W}_i\overrightarrow{W}_i^T g_{\alpha_n}(1 \mid W_i)g_{\alpha_n}(0 \mid W_i)\right]^{-1}(A - g_{\alpha_n}(1 \mid W))\overrightarrow{W}. \end{aligned}$$

In the absence of knowledge of IC_{Q^*}, the standard error of the C-TMLE can be estimated as $\text{SE}(\psi_n) = \sqrt{var(IC)/n}$, where $\text{var}(IC) = 1/n\sum_i \widehat{IC}_i^2$ is the sample variance of the estimated influence curve. A 95% confidence interval (CI) is constructed as $\psi_n \pm 1.96\text{SE}(\psi_n)$. In practice confidence intervals constructed by ignoring the contribution from IC_{Q^*} have achieved good coverage in practice across a range of simulated datasets. The bootstrap is an alternative valid method for asymptotically valid inference.

4.3.2 Further remarks

There are many variations for obtaining ψ_n^{C-TMLE}. For example, given an a priori set of candidate nuisance parameter estimators, g_n^j, that includes highly non-parametric candidates we could construct clever covariates $H_j^*(g)$, and then use forward selection with this set of clever covariates, using the initial estimator as offset, to build (second stage) model-fits for $\bar{Q}_0$ of increasing size, where each term in the model corresponds to one of the clever covariates. The number of clever covariates that are added in this forward-selection algorithm can be selected using likelihood-based cross-validation.

Note that in contrast with the algorithm described above, in which previous coefficients are used as fixed offsets in the regression, coefficients in front of each term are estimated by least squares, thereby solving the efficient influence equation corresponding to each g_n^j, in particular the most non-parametric of these. Because these covariates are highly correlated, refitting all coefficients in front of clever covariates at each step in the forward selection algorithm is likely to result in highly variable coefficient estimates, and therefore less stability in the estimate of the parameter of interest.

Another alternative approach is to define $\psi_n^{C-TMLE} = \Psi(Q_n^1)$, where $\bar{Q}_n^1 = \bar{Q}_n^0 + \epsilon_n H^*(g_n^{k*})$ is the targeted MLE updating the initial estimator with the final selected clever covariate defined by carrying out the k^* moves in the above forward selection algorithm to obtain a g-fit, where k^* is the optimal number of moves selected by likelihood-based cross-validation (exactly as above). Though this variation did not improve performance in simulation studies, these alternatives are mentioned to underscore the fact that C-TMLE methodology can be implemented in a variety of ways, and is not limited to the specific implementation presented here.

4.4 Simulation studies

Three simulation studies illustrate the performance of the C-TMLE estimator under different data-generating scenarios. The simulations are designed to demonstrate estimator performance in the face of confounding of the relationship between treatment and outcome, complex underlying data-generating distributions, and practical violations of the Experimental Treatment Assignment (ETA) Assumption, i.e., $P(A = a \mid W) < \alpha$, for some small α, implying that there is very little possibility of observing both treated and untreated subjects for some combination of covariates present in the data. Other estimators commonly used to assess causal effects are also used to analyze the data. A comparison of these estimators highlights the differences in their behavior, illustrates the importance of double robustness, and underscores the need to use an estimator that works well across a broad range of underlying probability distributions when analyzing real data from an unknown distribution. For example,

- if a correct model for the underlying data generating distribution is known, a parametric regression approach would be optimal.
- For rare outcomes we would not expect the initial fit, $\bar{Q}_n^0$, to have much predictive power. In this case, the fully adjusted g_0 is very likely needed for full bias reduction, so creating and evaluating intermediate candidates with C-TMLE may be needlessly computationally expensive. Standard TMLE might be a better approach.
- Adjusting for many confounders may lead to violations of the ETA assumption when n is small relative to the number of confounders or if the confounders are very strongly predictive of treatment. Parametric estimators rely strongly on model-based extrapolation in this case.

C-TMLE allows extrapolation through the initial first stage estimator, but not in the second stage, where confounders are selected based on the penalized log-likelihood.

A common misconception is that C-TMLE does not target the fully-adjusted parameter of interest. In fact, except in a case of severe sparsity, C-TMLE delivers the same bias reduction as TMLE, often with smaller variance.

4.4.1 Estimator review

Marginal treatment effect estimates were calculated based on the unadjusted regression of Y on A, maximum likelihood estimation (MLE) using the G-computation formula (Robins, 1986), inverse probability of treatment weighted (IPTW) estimation (Hernan et al., 2000a; Robins, 2000b), augmented IPTW (AIPTW) estimation, a double robust method (Robins and Rotnitzky, 2001; Robins et al., 2000; Robins, 2000a), a propensity score estimator (pscore), (Rosenbaum and Rubin, 1983) that calculates the marginal treatment effect as the mean across strata defined by the conditional probability of receiving treatment, and an extension to propensity score estimators implemented in `matching`, a publicly available R package (Sekhon, 2008) that matches observations in treatment and control groups based on minimizing a distance between the user supplied covariates W. In this matching procedure, each set of matched observations indexed by m results in a corresponding mean regression $\bar{Q}_n^0(a, m)$ representing an estimate of $E(Y \mid A = a, M = m)$ and its contrast $E(Y \mid A = 1, M = m) - E(Y \mid A = 0, M = m)$. The creation of the partitioning in sets of matched observations is only a function of the data (W_i, A_i), $i = 1, \ldots, n$, thus ignoring the outcome data.

The unadjusted estimator is defined as

$$\psi_n^{Unadj} = \frac{1}{n} \sum_{i=1}^{n} (2A_i - 1)Y_i.$$

When covariates confound the relationship between treatment and outcome, the unadjusted estimator will be biased.

The MLE estimator

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

performs well when the model for $\bar{Q}_0$ is correctly specified. $\bar{Q}_n^0$ refers to an initial estimate of $\bar{Q}_0(A, W)$.

The IPTW estimator is defined as

$$\psi_n^{IPTW} = \frac{1}{n} \sum_{i=1}^{n} \left[I(A_i = 1) - I(A_i = 0) \right] \frac{Y_i}{g_n(A_i, W_i)}.$$

Large weights on a small subset of observations is known to bias the IPTW estimator. This arises in the context of sparsity, and can actually increase bias, even when the treatment mechanism is correctly specified (Freedman and Berk, 2008).

The AIPTW estimator is defined as

$$\begin{aligned}\psi_n^{AIPTW} &= \frac{1}{n}\sum_{i=1}^{n}\frac{[I(A_i=1)-I(A_i=0)]}{g_n(A_i \mid W_i)}(Y_i-\bar{Q}_n^0(A_i,W_i)) \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^0(1,W_i)-\bar{Q}_n^0(0,W_i)).\end{aligned}$$

AIPTW estimates are unbiased and are asymptotically efficient when both g_0 and the functional form for $\bar{Q}_0$ is correctly modeled, and they remain unbiased if at least one is correctly specified. Unlike C-TMLE, AIPTW relies on external estimation of g, and may therefore include covariates that are predictive only of treatment, tending to increase both bias and variance.

The pscore estimator is given by

$$\psi_n^{pscore} = \frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^0(1,s_i)-\bar{Q}_n^0(0,s_i)).$$

$\bar{Q}_n^0(a,s)$ is an estimator of the true conditional mean $E(Y \mid A=a, S=s)$, s_i indicates a stratum of the propensity score of covariate vector W_i. Propensity score methods are especially effective when the propensity score is a function of true confounders. Estimates can suffer even when overall match quality based on the propensity score is high if a small subset of covariates responsible for introducing the most bias into the estimate is unevenly distributed between treatment and control groups. Like most estimators, these estimators are known to perform poorly when there are ETA violations (Sekhon, 2008). A practical violation of the experimental treatment assignment assumption, also called the positivity assumption, is known to reduce the quality of the match and introduce bias into the estimate, and can be detected once the matches have been specified. However, without using information about the outcome the matching quality is not guided by the potential bias reduction for the parameter of interest. Because matches are made without knowledge of the outcome, these methods do not exploit all information available in the data and cannot achieve the semi-parametric efficiency bound (Abadie and Imbens, 2006).

The matching estimator is defined as

$$\psi_n^{matching} = \frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^0(1,m_i)-\bar{Q}_n^0(0,m_i)).$$

$\bar{Q}_n^0(a,m)$ is an estimate of true conditional mean $E(Y \mid A=a, M=m)$, m_i indicates a set of matched observations to which subject i is assigned. The matching algorithm estimator generalizes the pscore approach by carefully matching observations in the treatment and control groups in an

effort to evenly distributed potential confounders. The matching procedure relies on the genetic algorithm (Holland and Reitman (1977)) to achieve this goal. This is a non-parametric approach for selecting weights on covariates that are in turn are used to determine which observations are matched. Candidate sets of matches are evaluated based on a loss function and a distance metric specified at run-time, and are used to generate successive sets of candidates that achieve good balance (Sekhon, 2008). The marginal treatment effect is the average effect across strata defined by the matches. This estimator also ignores the outcome, and is less than fully efficient.

The C-TMLE estimator is defined as

$$\psi_n^{C-TMLE} = \frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i))$$

where $\bar{Q}_n^*$ refers to an updated targeted estimate of $\bar{Q}_0(A, W)$.

4.4.2 Data generation

Covariates $W_1, \ldots W_5$ are generated as independent normal random variables, while W_6 is a binary variable. Specifically,

$$\begin{aligned} W_1, W_2, W_3, W_4, W_5 &\sim N(0, 1) \\ P(W_6 = 1 \mid W_1, W_2, W_3, W_4, W_5) &= expit(0.3W_1 + 0.2W_2 - 3W_3). \end{aligned}$$

Two treatment mechanisms are defined:

$$\begin{aligned} g_{1,0} &= P(A = 1 \mid W) = expit(0.3W_1 + 0.2W_2 - 3W_3) \\ g_{2,0} &= P(A = 1 \mid W) = expit(0.15(0.3W_1 + 0.2W_2 - 3W_3)). \end{aligned}$$

The observed outcome Y is generated as

$$Y = \bar{Q}_{i,0}(A, W) + \epsilon,\ \epsilon \sim N(0, 1)$$

with corresponding true outcome regressions

$$\begin{aligned} \bar{Q}_{1,0}(A, W) &= A + 0.5W_1 - 8W_2 + 9W_3 - 2W_5 \\ \bar{Q}_{2,0}(A, W) &= A + 0.5W_1 - 8W_2 + W_3 + 8W_3^2 - 2W_5. \end{aligned}$$

We consider three different data-generating distributions, $(\bar{Q}_{1,0}, g_{1,0})$ in simulation 1, $(\bar{Q}_{2,0}, g_{1,0})$ in simulation 2, and $(\bar{Q}_{2,0}, g_{2,0})$ in simulation 3. Note that W_6 is strongly correlated with treatment mechanism A in simulations 1 and 2 (corr=0.54), but is not an actual confounder of the relationship between A and Y. W_1, W_2, and W_3 are confounders. The linear nature of the confounding due to W_3 in simulation 1 differs from that in simulations 2 and 3, where the true functional form is quadratic. In this way simulations 2 and 3 mimic realistic data analysis scenarios

in which the unknown underlying functional form is seldom entirely captured by the regression model used in the analysis. Finally, the treatment mechanism in simulations 1 and 2 leads to ETA violations ($p(A = a \mid W)$ ranges between 9×10^{-7} and 0.9999978, approximately one-third of the probabilities are outside the range (0.05, 0.95)). In simulation 3 there are no ETA violations ($0.11 < P(A = a \mid W) < 0.88$). In each simulation the true value of the parameter of interest is the same, $\psi_0 = 1$.

4.4.3 Description

One thousand samples of size $n = 1000$ were drawn from each data generating distribution. A main-effects model for $\bar{Q}_n^0$ used for the MLE and AIPTW estimators was obtained using the DSA algorithm (Sinisi and van der Laan, 2004). This data-adaptive algorithm searches over a large space of polynomial models by adding, deleting, or substituting terms, starting with a base user-specified regression model. The final model is selected by cross-validation using an $L2$ loss function.

A model for the treatment mechanism g_n used in IPTW, AIPTW, propensity score, and matching estimation was also selected by DSA, restricted to main terms. The propensity score method was implemented by dividing observations into strata based on the quintiles of the predicted conditional treatment probabilities. Regression of Y on A and strata indicator variables using the full model enabled the calculation of stratum-specific treatment effects, which were averaged to obtain the marginal effect.

We expect to see that the estimators that rely on consistent estimation of $\bar{Q}_0$ are unbiased in simulation 1, (MLE, AIPTW, C-TMLE), while estimators that are consistent given consistent estimation of g_0 are unbiased in simulation 3 (IPTW, AIPTW, pscore, matching, C-TMLE). Sparsity in simulations 1 and 2 poses a challenge for estimators that rely on g.

4.4.4 Results

Mean estimates of the treatment effect and standard errors for each simulation are shown in Table 4.1. Figure 4.4 illustrates each estimator's behavior. As expected, estimators relying on consistent estimation of $\bar{Q}_0$ are unbiased in simulation 1, those relying on consistent estimation of g_0 are unbiased in simulation 3.

The unadjusted estimator yields biased results in all three simulations due to its failure to adjust for confounders.

The MLE estimator performs well in simulation 1 when the model is correctly specified. We understand that misspecification (simulations 2 and 3) will often, though not always, lead to bias in the estimates. However the plots highlight another phenomenon that is easy to overlook. the inability of the misspecified model to adequately account for the variance in the outcome often leads to large residual variance of the estimator, and in practice would have low power to reject a null hypothesis.

Table 4.1: Mean estimate and standard error (SE) for each estimator based on 1000 iterations with sample size $n = 1000$. $\psi_0 = 1$.

	Simulation 1		**Simulation 2**		**Simulation 3**	
	$\overline{\psi}_n$	SE	$\overline{\psi}_n$	SE	$\overline{\psi}_n$	SE
Unadj	-11.97	0.64	-0.98	0.91	0.29	0.86
MLE	0.99	0.09	0.76	1.22	0.95	0.68
IPTW	-4.36	0.72	0.03	0.76	0.83	0.90
AIPTW	0.99	0.09	0.94	0.62	1.03	0.80
pscore	-1.09	1.27	0.42	1.38	0.93	0.59
matching	-1.22	0.82	0.54	0.73	0.96	0.25
C-TMLE	0.99	0.09	1.00	0.10	1.00	0.07

Truncation bias due to ETA violations causes the IPTW estimator using truncated weights to fail in simulations 1 and 2. The estimate is not biased in simulation 3, but the variance is so large that even in this setting where we'd expect IPTW to be reliable it would fail to produce a significant result.

AIPTW estimates are unbiased and have low variance when the functional form is correctly modeled by the regression equation (simulation 1). However, the variance of the AIPTW estimator is large in simulations 2 and 3 because W_6, a strong predictor of A that is not a confounder, is always included in the estimate of the treatment mechanism, thus needlessly increasing the variance.

Though we see little bias in the other two simulations, the variance is large due to misspecification of the treatment mechanism. Because W_6 is a strong predictor of A and is indistinguishable from a true confounder of the relationship between Y and A it is always included in the treatment mechanism, behavior that does not help achieve an accurate estimate of the true treatment effect.

Researchers constructing the propensity score could observe the poor performance in simulations 1 and 2 when there are ETA violations and choose an alternate propensity score model, but without using information about the outcome this choice could be based on the predictive power of the model, but not the potential bias reduction. The propensity score method does a reasonable job in simulation 3.

The matching estimator performs well, however, its theoretical inability to achieve the semiparametric efficiency bound is confirmed in the simulations, since the confidence interval are not as tight as that of the collaborative targeted maximum likelihood estimator.

4.4.5 Summary

These simulation studies demonstrate the collaborative double robustness and efficiency of C-TMLE methodology, which allows for consistent efficient estimation in situations when other

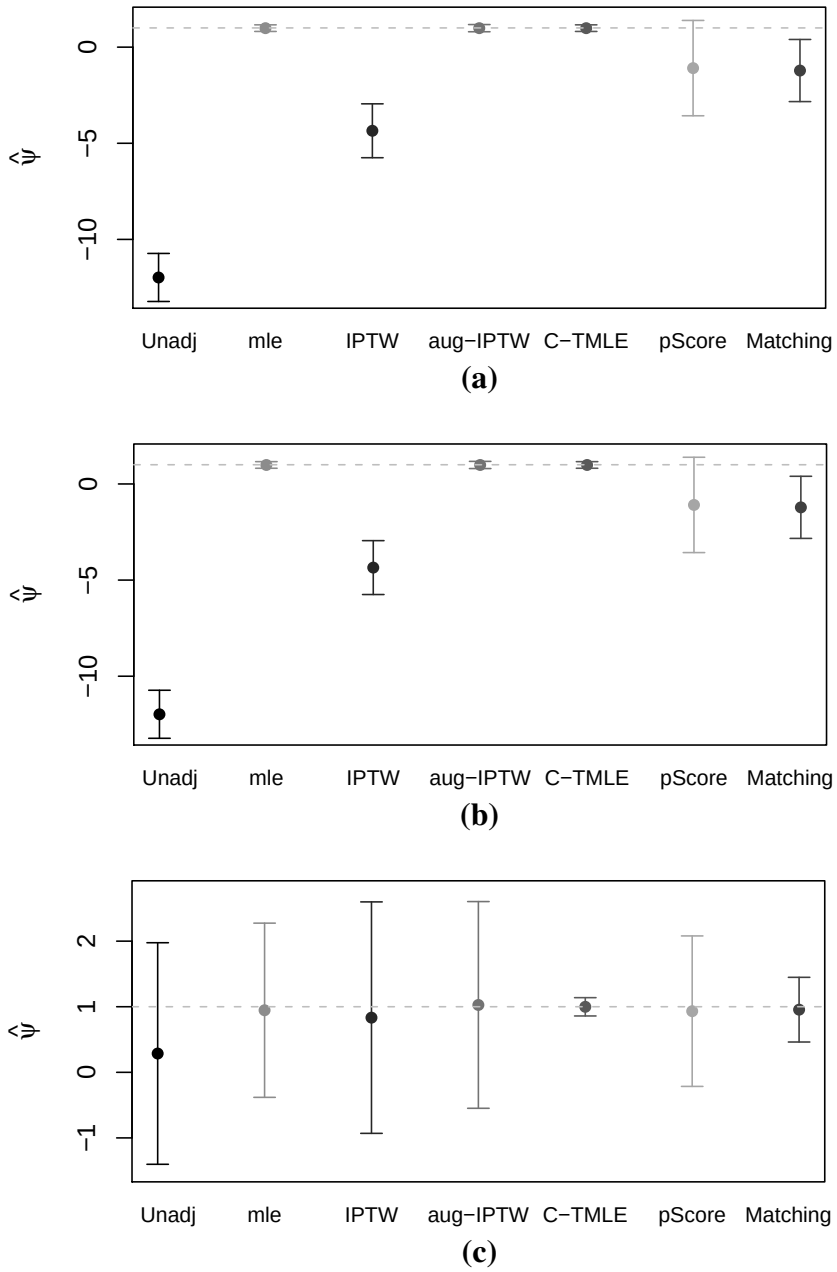

Figure 4.4: Mean estimates and (0.025, 0.975) quantiles for each estimation method, (a) simulation 1, (b) simulation 2, (c) simulation 3. Dashed line in each plot is at true parameter value.

estimators can fail to perform adequately. In practice these failures may lead to biased estimates and to confidence intervals that fail to attain the correct coverage, as suggested by the IPTW results

in simulations 1 and 2, where weights depend on a variable highly predictive of treatment that is not a true confounder of the relationship between Y and A. It is worth noting that the unadjusted estimator applied to data from a randomized controlled trial in which randomization fails to evenly distribute confounders across treatment arms will also yield (finite sample) biased results, as we saw in simulations 1, 2, and 3.

As simulations 2 and 3 demonstrate, a misspecified parametric model not only results in biased estimates, but can also easily fail to adequately explain the variance in the outcome. Therefore estimates of the parameter of interest will have a larger variance than the semi-parametric information bound achieved by an efficient estimator, such as C-TMLE. Such misspecified parametric models can easily result in the construction of a confidence interval that contains 0, and therefore a failure to reject a false null hypothesis, even when the point estimate is close to the true value of the parameter of interest. Since misspecified parametric models are the rule rather than the exception, in the analysis of data from an unknown data-generating distribution, using C-TMLE combined with super learning for the initial estimator, is a prudent course of action, and provides sound influence curve-based inference.

Estimators that rely on nuisance parameter estimation (IPTW, AIPTW, TMLE, propensity score-based estimation) break down when there are ETA violations, failing to reduce bias, or even increasing bias, while incurring high variance that renders estimates meaningless (no statistical significance). An effort to reduce variance through truncation introduces bias into the estimate, and requires a careful trade-off. C-TMLE addresses these issues, in the sense that it is able to utilize the covariates for effective bias reduction, avoiding harmful bias reduction efforts. As a targeted-MLE, the bias-variance tradeoff is targeted towards the estimation of the parameter of interest, not the estimate of the entire density.

The collaborative nature of the estimation of the treatment mechanism in the C-TMLE confers three advantages:

1. The treatment mechanism model will exclude covariates that are highly predictive of treatment but do not truly confound the relationship between treatment and the outcome.

2. The treatment mechanism model will include only covariates that help adjust for residual bias remaining after stage one adjustment.

3. Cross-validation based on a penalized log-likelihood will not select a treatment mechanism model that includes a term that leads to violations of the ETA assumption and thereby large variance of the corresponding targeted MLE without the benefit of a meaningful bias reduction.

4.5 Comparison of C-TMLE and TMLE

The double robust property of the targeted maximum likelihood estimator minimizes the need for accurate estimation of both $\bar{Q}_0$ and g_0 since correct specification of either one leads to consistent

estimates of the parameter of interest. However, accurate estimates of both are needed to achieve the Cramer-Rao efficiency bound. Implementations of the standard targeted maximum likelihood estimator (TMLE) therefore strive for ideal estimates of both $\bar{Q}_0$ and g_0.

In contrast, the collaborative nature of the second stage of the C-TMLE estimation algorithm leads to selection of an estimator, g_n, that targets that portion of the treatment mechanism needed to reduce bias not already adequately addressed by the first stage estimator for $\bar{Q}_0$. For example, covariates included in the model for $\bar{Q}_n^0$ might not be selected into the model for g because they do not increase the penalized log-likelihood. At the same time, confounders that are not adequately adjusted for in the initial density estimate are quickly added to model for g unless the gain in bias reduction is offset by too great an increase in variance. When the initial estimate of the density is a very good fit for the true underlying density, TMLE and C-TMLE have similar performance with respect to bias, but the C-TMLE will have smaller variance by selecting a g_n that targets non fully adjusted g_0, resulting in a possibly super efficient estimator. When the initial fit is less good, C-TMLE makes informed choices regarding inclusion of covariates in the treatment mechanism. As predicted by theory, again, this might lead to lower variances when no covariates cause ETA violations. When inclusion of all confounding covariates does violate the ETA assumption, the C-TMLE estimator, in essence, targets a less ambitious data adaptively selected parameter that is identifiable. Simulation 4 illustrates these phenomena.

4.5.1 Data generation

In simulation 4 the covariates W_1, W_2, and W_3 are generated as independent random uniform variables over the interval $[0, 1]$, while W_4 and W_5 are independent normally distributed random variables. Specifically,

$$\begin{aligned} W_1, W_2, W_3 &\sim U(0,1) \\ W_4, W_5 &\sim N(0,1). \end{aligned}$$

Treatment mechanism g_0 is designed so that W_3 is highly predictive of treatment:

$$g_0 = P(A = 1 \mid W)) \quad = \quad expit(2W_1 + W_2 - 5W_3 + W_5).$$

The observed outcome Y is generated as

$$Y \quad = \quad \bar{Q}_0(A, W) + \epsilon,\ \epsilon \sim N(0,1)$$

with corresponding true outcome regression

$$\bar{Q}_0(A, W) \quad = \quad A + 4W_1 - 5W_2 + 5W_4W_5.$$

4.5.2 Description

C-TMLE and TMLE estimates of the parameter of interest, again defined as $\psi = E_W\{E(Y \mid A = 1, W) - E(Y \mid A = 0, W)\}$, were obtained for 1000 samples of size $n = 1000$ drawn from data generating distribution implied by (Q_0, g_0). For this study we deliberately select a misspecified main-terms only model for $\bar{Q}_0$ by running the DSA algorithm on 100,000 observations drawn from that same distribution. $P(A = a \mid W)$ for these observations ranges from 0.004 to 0.996. Approximately 17% of the observations have covariates indicating that the probability of receiving treatment is less than 0.05, indicating that practical ETA violations in finite samples will cause unstable TMLE estimates.

For each iteration an initial regression, $\bar{Q}_n^0$, was obtained by fitting the DSA-selected model, $Y = A + W_1 + W_2$, on n observations in the sample. We expect that any estimate of ψ_0 based solely on this model is likely to be incorrect because the model fails to take into account the effect on the outcome of the missing interaction term, and also fails to adjust for the confounding effect of W_5. The targeting step for both targeted maximum likelihood estimators reduces this bias.

In order to construct the covariate used to target the parameter of interest in the updating step of the TMLE algorithm we obtain an estimate g_n of g_0 by running the DSA algorithm, allowing quadratic terms and two-way interaction terms to enter the model. This model was not fixed over the 1000 iterations; the model selection process was carried out each time a sample was drawn from the population. Similarly, covariates that were candidates for inclusion in the model for g_n in the second stage of the C-TMLE estimation algorithm include $(W_1, \ldots, W_5, W_1^2, \ldots, W_5^2)$, and all two-way interaction terms $(W_i W_j)$, where $i \neq j$.

4.5.3 Results

Results of the simulation are shown in Table 4.2. A small number of aberrant TMLE estimates were major contributers to the variance of that estimator. The three highest TMLE estimates of the treatment effect were (771.91, 37.22, 9.52). It is likely that these high values arise from atypical samples containing observations that presented unusually strong ETA issues. In contrast, all C-TMLE estimates calculated from those same samples range between 0.307 and 1.698. Both estimators' average treatment effect estimates are not far from the true value, $\psi_0 = 1$. As expected, the variance of the TMLE estimator is many times larger than that of the C-TMLE estimator.

Not surprisingly, W_3, the strong predictor of treatment that is not a true confounder of the relationship between treatment and outcome, is included in every one of the 1000 models for g_n selected by the DSA algorithm, but it is included in only 35 of the models constructed in the second stage of the C-TMLE algorithm. At the same time, the interaction term $W_4 W_5$ is included in only two out of 1000 models for g_0 selected by DSA, but is present in 576, more than half, of the collaborative models.

This clearly demonstrates the differences between TMLE's reliance on an external estimate of g_0 and the collaborative approach to estimating the treatment mechanism used by C-TMLE. However, we note that the degradation of TMLE performance under sparsity is due to the unbound-

Table 4.2: Simulation 4: Comparison of C-TMLE and TMLE estimators at different levels of truncation. Mean estimate and variance based on 1000 iterations.

	Truncation level	# Obs truncated	$\overline{\psi}_n$	Variance
C-TMLE	∞	0	0.98	0.04
TMLE	∞	0	1.73	597.52
	40	1	1.36	162.38
	10	2	0.94	1.99
	5	9	0.92	1.68

edness of the fluctuation function, and can be mitigated by employing an alternative fluctuation function that respects known bounds on the data model.

4.5.4 Confidence Intervals

The variance of the influence curve provides the basis for calculation of a 95% confidence interval for the C-TMLE estimate.

$$95\%\, CI = \psi^{C-TMLE} \pm 1.96\sqrt{var(IC)/n}$$

Two sets of confidence intervals were constructed for each of the 1000 iterations in simulation 4, with $\bar{Q}_n^0$ misspecified by a main-terms only regression model. As described above, one set of CIs is based on $D^*(Q,g)$, the first term of the IC. The second set is based on the variance of $D^*(Q,g) + IC_g$, which includes the contribution from the estimation of g_n. Table 4.3 shows that CIs based on D^* alone are conservative when the model for Q_n^0 is misspecified, as expected. In contrast, observed coverage closely approximates the nominal 95% coverage rate when the contribution from the IC_g term is taken into account.

Confidence intervals were also created for an additional 1000 samples from the same data generating distribution that were analyzed using a correct model for $\bar{Q}_n^0$. Coverage rates for these confidence intervals are given in Table 4.3. When $\bar{Q}_n^0$ is correctly specified we observe little difference in the coverage rate whether or not we take the contribution from IC_g into account, indicating zero contribution to the variance from the estimate of g_n. Attaining the nominal rate indicates that inference is reliable even when the estimator is super efficient.

Table 4.3: Empirical coverage of 1000 confidence intervals constructed at a nominal 95% level. SE calculated as $\sqrt{var(IC)/n}$, where the IC was estimated with and without IC_g.

	Coverage	
	$D^*(Q, g_0)$	$D^*(Q, g_0) + IC_g$
$\bar{Q}_n^0$ misspecified	.979	.943
$\bar{Q}_n^0$ correct	.932	.933

4.6 HIV Mutation Data Analysis

We apply the C-TMLE estimator to an observational dataset previously analyzed with the goal of identifying HIV mutations that affect response to the antiretroviral drug lopinavir. (Bembom et al., 2009, 2008) The data includes observations on $O = (W, A, Y)$, where the outcome, Y, is the change in $\log_{10}$ viral load measured at baseline and at follow-up after treatment has been initiated. If follow-up viral load was beneath the limit of detection Y was set to the maximal change seen in the population. $A \in \{0, 1\}$ is an indicator of the presence or absence of a mutation of interest, taking on the appropriate value for each of the 26 candidate mutations in 26 separate analyses. W consists of 51 covariates including treatment history, baseline characteristics, and indicators of the presence of additional HIV mutations. Practical ETA violations stemming from high correlations among some of the covariates and/or low probability of observing a given mutation of interest make it difficult to obtain stable low variance estimates of the association between A and Y. Bembom used a targeted maximum likelihood estimation approach incorporating data-adaptive selection of an adjustment set that relies on setting a limit on the maximum allowable truncation bias introduced by truncating treatment probabilities less than α to some specified lower limit. Covariates whose inclusion in the adjustment set introduces an unacceptable amount of bias are not selected. That study's findings showed good greement with Stanford HIVdb mutation scores, values on a scale of 0 to 20 (http://hivdb.stanford.edu, as of September, 2007, subsequently modified), where 20 indicates evidence exists that the mutation strongly inhibits response to drug treatment and 0 signifies that the mutation confers no resistance. Because the C-TMLE method includes covariates in the treatment mechanism only if they improve the targeting of the parameter of interest without having too adverse an effect on the MSE, we expect similar performance without having to specify truncation levels or an acceptable maximum amount of bias.

4.6.1 Analysis description

The dataset consists of 401 observations on 372 subjects. Correlations due to the few subjects contributing more than one observation were ignored. Separate analyses was carried out for each mutation. In each, an initial density estimate, $\bar{Q}_n^0$, was obtained using DSA restricted to addition moves only to select a main-terms model containing at most 20 terms, where candidate terms in

W include pre-computed interactions detailed in Bembom et al. A was forced into the model. An estimate of the effect on change in viral load was recorded for each mutation. Influence curve-based variance estimates incorporating the contribution from estimating g given by the IC_g term, was used to construct 95% confidence intervals.

4.6.2 Results

Table 4.4 lists the Stanford mutation score associated with each of the HIV mutations under consideration, as well as the C-TMLE estimate of the adjusted effect of mutation on lopinavir resistance. 95% confidence intervals were constructed based on the variance of the IC. Confidence intervals entirely above zero indicate a mutation increases resistance to lopinavir. Eight of the twelve mutations having a mutation score of 10 or greater fall into this category. Point estimates for the remaining four mutations were positive, but the variance was too large to produce a significant result. Only one of the six mutations thought to confer slight resistance to lopinavir was flagged by the procedure, though with the exception of p10FIRVY point estimates were positive. Stanford mutation scores of 0 for four of the five mutations found to have a significantly negative effect on drug resistance support the conclusion that these mutations do not increase resistance, but are not designed to offer confirmation that a mutation can decrease drug resistance. However, Bembom et al. report that there is some clinical evidence that two of these mutations, 30N and 88S, do indeed decrease lopinavir resistance.

These findings are quite consistent with the Stanford mutation scores and with the results from the previous analysis using the data-adaptively selected adjustment set targeted maximum likelihood estimation approach. The C-TMLE method was able to achieve these results without relying on ad hoc or user-specified tuning parameters.

Table 4.4: Stanford score (2007), C-TMLE estimate and 95% confidence interval for each mutation. Starred confidence intervals do not include 0.

Mutation	Score	Estimate	95% CI
p50V	20	1.703	(0.760, 2.645)*
p82AFST	20	0.389	(0.091, 0.688)*
p54VA	11	0.505	(0.241, 0.770)*
p54LMST	11	0.369	(0.002, 0.735)*
p84AV	11	0.099	(-0.139, 0.337)
p46ILV	11	0.046	(-0.222, 0.315)
p82MLC	10	1.610	(1.377, 1.843)*
p47V	10	0.805	(0.282, 1.328)*
p84C	10	0.602	(0.471, 0.734)*
p32I	10	0.544	(0.325, 0.763)*
p48VM	10	0.306	(-0.162, 0.774)
p90M	10	0.209	(-0.063, 0.481)
p33F	5	0.300	(-0.070, 0.669)
p53LY	3	0.214	(-0.266, 0.695)
p73CSTA	2	0.635	(0.278, 0.992)*
p24IF	2	0.229	(-0.215, 0.674)
p10FIRVY	2	−0.266	(-0.545, 0.012)
p71TVI	2	0.019	(-0.243, 0.281)
p23I	0	0.822	(-0.014, 1.658)
p36ILVTA	0	0.272	(-0.001, 0.544)
p16E	0	0.239	(-0.156, 0.633)
p20IMRTVL	0	0.178	(-0.111, 0.467)
p63P	0	−0.131	(-0.417, 0.156)
p88DTG	0	−0.426	(-0.842,-0.010)*
p30N	0	−0.440	(-0.853,-0.028)*
p88S	0	−0.474	(-0.781,-0.167)*

4.7 C-TMLE for bounded continuous outcomes

Chapter 3 described the importance of respecting global constraints on the estimation problem for bounded continuous outcomes, and introduced a logistic fluctuation procedure that ensures that TMLE estimates of $\bar{Q}_0(A, W)$ remain within the bounds of the semi-parametric model. This is especially relevant in sparse data situations, when outlying values for Y or $\bar{Q}_0(A, W)$, or extreme conditional treatment assignment probabilities inflate the variance of the efficient influence curve of the parameter of interest. C-TMLE's parsimonious approach to nuisance parameter estimation addresses both these sources of variance, and increase the practical identifiability of the parameter. Selecting the smallest sufficient model for g often yields less extreme predicted values for $g_n(1 \mid W)$ than those based on the fully adjusted g_0, and the penalized targeted forward selection approach to building candidate estimators $\bar{Q}^*_n(g^1_n), \ldots, \bar{Q}^*_n(g^K_n)$ tends to keep the variance of the early candidates in check, while ensuring that the likelihood for Q increases monotonically. The logistic fluctuation procedure also addresses both sources of variance by producing targeted estimates that always remain within the known bounds. An analysis of simulated data illustrates that employing a logistic fluctuation of $\bar{Q}^0_n$ in the targeting step of the C-TMLE procedure further robustifies the C-TMLE estimator with respect to sparsity.

4.7.1 The logistic fluctuation procedure

The targeting step of the TMLE procedure for a binary outcome uses logistic regression of Y on $H^*(A, W)$ with offset $logit(\bar{Q}^0_n)$ to fit ϵ, a parameter that dictates the magnitude of the fluctuation of the initial estimate. This naturally constrains the updated estimate, $\bar{Q}^1_n(A, W) = expit(logit(\bar{Q}^0_n(A, W)) + \epsilon H^*(A, W))$, to be between 0 and 1. If instead Y represents a continuous outcome known to be bounded between $(0, 1)$, for example, a proportion, then it is equally true that any observed or fitted value for Y should fall between 0 and 1. Fluctuating an initial estimate $\bar{Q}^0_n \in (0, 1)$ for the conditional mean of this continuous Y on the logit scale will yield a targeted estimate, $\bar{Q}^1_n$ that is guaranteed to fall between 0 and 1.

Now suppose there is instead a continuous outcome Y known to be bounded by (a, b), with $a < b$. Ideally, an estimate of the conditional mean of Y given A and W should remain within $[a, b]$. We've just seen that this is easily arranged when $(a, b) = (0, 1)$. Chapter 3 showed that for arbitrary (a, b), $Y \in [a, b]$ can be mapped to $Y^* \in [0, 1], Y^* = (Y - a)/(b - a)$. We define the causal effect of treatment on the bounded outcome Y^* as $\Psi^*(P_0) = E_0\{E_0(Y^* \mid A = 1, W) - E_0(Y^* \mid A = 0, W)\}$. The same C-TMLE procedure outlined in Algorithm 6.1 is applied to $O^* = (W, A, Y^*)$ to obtain an estimate ψ^*_n that immediately maps to a ψ_n of the causal effect on the original scale, using the relation $\Psi(P_0) = (b - a)\Psi^*(P_0)$. A confidence interval for ψ_0 can be obtained by multiplying the bounds on the confidence interval for $\Psi^*(P_0)$ by $(b - a)$. Similarly, the estimated variance $\hat{\sigma}^2$ of ψ_n is obtained by multiplying the estimated variance $\hat{\sigma}^{2*}$ of ψ^*_n with$(b - a)^2$.

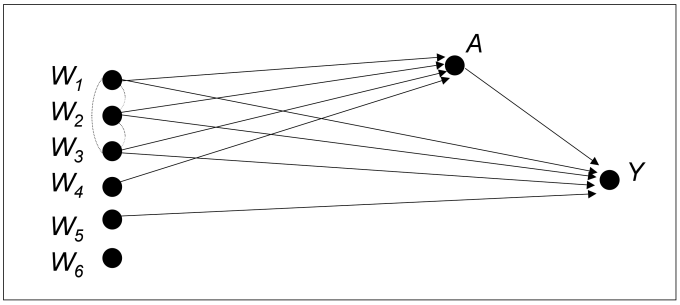

Figure 4.5: Simulation 5 DAG shows the relationship between covariates collected at baseline W, treatment, A, and outcome Y. Solid lines represent causal relationships, dashed lines represent non-causal correlations.

4.7.2 Simulation study

Data generation

Simulation 5 was designed to highlight different aspects of estimator performance in the context of sparsity. The directed acyclic graph (DAG) in Figure 4.5 shows the relationship between covariates $W = (W_1, W_2, W_3, W_4, W_5, W_6)$, binary treament, A, and continuous outcome Y. Data were generated as follows: covariates W_1, W_2, W_3 are trivariate normal, W_4, W_5, W_6 are independent binary variables, and specifically,

$$\begin{aligned} W_1, W_2, W_3 &\sim N(\mu_1, \mu_2, \mu_3, \Sigma),\ \mu_1 = \mu_2 = \mu_3 = 0,\ \Sigma = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{bmatrix} \\ W_4 &\sim Bernoulli(0.2) \\ W_5 &\sim Bernoulli(0.6) \\ W_6 &\sim Bernoulli(0.7). \end{aligned}$$

The treatment mechanism g_0 is given by

$$g_0 = P(A = 1 \mid W)) \quad = \quad expit(2W_1 + 0.25W_2 - 0.5W_3 + W_4).$$

The observed outcome Y is generated as

$$Y \quad = \quad \bar{Q}_0(A, W) + \epsilon,\ \epsilon \sim N(0, 1)$$

with corresponding true outcome regression

$$\bar{Q}_0(A, W) \quad = \quad A + 2AW_5 + W_1 + W_2 - W_3W_5.$$

Notice that covariates (W_1, W_2, W_3, W_4) are causally associated with treatment assignment. The covariates W_1, W_2, and W_3 are also causally related to treatment, and therefore confound the relationship between treatment and the outcome. The lack of an arrow from W_4 to Y in Figure 4.5 indicates that W_4 is not causally related to Y, and is thus not a confounder of the relationship between treatment and outcome. Covariate W_6 was measured at baseline but has no association with either the treatment or the outcome. Covariate W_5 is an effect modifier. The effect of treatment is larger for subjects having $W_5 = 1$ than for subjects having $W_5 = 0$. The marginal treatment effect depends on the joint distribution of covariates W_1, W_2, W_3, W_5 in the population, which is estimated as the empirical distribution in the observed data. Though approximately one half of subjects receive treatment ($P(A = 1) = 0.53$ marginally), true treatment assignment probabilities are between (0.0002, 0.9999), and for approximately 9% of observations, the conditional probability of receiving treatment given the measured covariates is outside (0.05, 0.95).

1000 samples of size $n = 1000$ were drawn from this data generating distribution. Observed values for Y and initial estimated conditional mean values were truncated at the (0.01, 0.99) quantiles of the full data (-5.83, 8.48). The true value of the marginal additive treatment effect is $\psi_0 = 2.192$.

Comparison of C-TMLE using logistic fluctuation and C-TMLE using a linear fluctuation

Two C-TMLE estimators were applied to estimate the additive causal effect, C-TMLE$_{log}$, using a logistic fluctuation, and C-TMLE$_{lin}$ using a linear fluctuation. In order to demonstrate the impact the targeting step has on reducing bias, instead of data-adaptively estimating $\bar{Q}_n^0$, the first stage estimate of $\bar{Q}_n^0$ was obtained in two ways: 1) using the correct parametric regression model for $\bar{Q}_n$, and 2) using a misspecified model for $\bar{Q}_n$, i.e., the unadjusted regression of Y on A.

Results in Table 4.5 illustrate that, as expected, when the model for $\bar{Q}_n$ is correctly specified there is little difference between fluctuating on the logistic or linear scale. The C-TMLE procedure tends to select the intercept model for g_0 when the model for stage one estimation of $\bar{Q}_n^0$ is correctly specified. When this correct model for $\bar{Q}_n^0$ already contains the treatment variable, A, and $H_{g_n}^*(A, W)$ is based on a $g_n(A, W)$ modeled as the marginal treatment assignment probability, the MLE for ϵ is 0. Thus, in this case the scale of the fluctuation is irrelevant. Values for ϵ will also be close to 0 when $\bar{Q}_n^0$ is correctly specified and $H_{g_n}^*(A, W)$ is based on some larger model for g_n, as there will typically be little to no residual confounding after stage one estimation of $\bar{Q}_n^0$.

Differences emerge when the model for $\bar{Q}_n^0$ is deliberately misspecified. At each level of bound on g_n, the linear fluctuation yields estimates that are much more biased and have higher variance than the logistic fluctuation-based estimates. We see that increasing the bound on g_n from 0 to 0.025 reduces both bias and variance for the linear fluctuation estimates, but imposes a bias/variance tradeoff on the logistic fluctuation estimates. In this simulation the MSE is smaller when g_n is bounded at (0.025, 0.975) than when the bounds are closer to $(0, 1)$, but this is not always the case.

Table 4.5: Simulation 5: Comparison of C-TMLE$_{log}$ with C-TMLE$_{lin}$ at different bounds on g_n.

	Q **correctly specified**				Q **misspecified**			
	Ave	Bias	Var	MSE	Ave	Bias	Var	MSE
g_n **bound = 0**								
C-TMLE$_{log}$	2.222	0.030	0.008	0.009	2.154	−0.038	0.033	0.034
C-TMLE$_{lin}$	2.221	0.029	0.008	0.009	1.992	−0.200	0.349	0.389
g_n **bound=0.01**								
C-TMLE$_{log}$	2.222	0.030	0.008	0.009	2.151	−0.041	0.032	0.034
C-TMLE$_{lin}$	2.221	0.029	0.008	0.009	2.057	−0.135	0.297	0.315
g_n **bound=0.025**								
C-TMLE$_{log}$	2.222	0.030	0.008	0.009	2.146	−0.046	0.027	0.029
C-TMLE$_{lin}$	2.221	0.029	0.008	0.009	2.116	−0.076	0.054	0.060

Multiple estimator comparison

As is often true when study data are collected, not all covariates generated for simulation 5 are related to both treatment and the outcome. Though domain knowledge can be useful for identifying potential confounders, and can be incorporated into model selection techniques, treatment assignment may in fact partly depend on covariates that are not related to the outcome. In situations like these, incorporating information about Y in a systematic, prescribed manner that does not introduce bias can be beneficial. Simulation 5 is designed to provide the opportunity for this potential gain in relative efficiency.

Table 4.6 shows results of applying the estimators defined in Section 4.4.1 above, and the TMLE, of the additive treatment effect under the data generating distribution scheme for simulation 5. Estimators that make use of $g_n(A, W)$ were given values estimated using the correct regression formula, and subsequently bounded at $(p, 1-p)$, for $p = (0, 0.01, 0.025)$. Observed values for Y were truncated at the values of the (0.01, 0.99) quantiles. TMLE and C-TMLE results presented were obtained using the logistic fluctuation.

These results indicate that when the parametric model for $\bar{Q}_0$ is correctly specified, estimators that rely on consistent estimation of Q_0 perform very well. However, estimators that rely only on consistent estimation of g_0 and fail to exploit the information from estimation of Q_0, (IPTW, pscore, and matching), are less efficient, in spite of being given the correct model for g_0. Misspecifying the model for $\bar{Q}_0$ does not harm these estimators, but in situations like the one in this simulation, they are still less efficient than TMLE and C-TMLE.

The unadjusted estimate is biased due to confounding by covariates W_1, W_2, W_3. MLE has the smallest mean squared error when the model for $\bar{Q}_0$ is correctly specified, but is not robust to misspecification of this model. IPTW, AIPTW, matching TMLE, and C-TMLE estimators, all of which rely on g_n, show improvements in MSE as the bounds on g increase from 0 to 0.025 due to

decreases in the variance at the cost of increasing bias. The IPTW estimator provides consistent parameter estimation, but is not maximally efficient. AIPTW has lower bias than IPTW, but pays a high price in variance when $\bar{Q}_n$ is inconsistent. The pscore estimator is quite stable across all truncation levels for g_n, however its lack of data-adaptiveness yields an estimate that is quite biased in comparison with the other methods. The matching estimator is less biased than pscore, and also quite stable with respect to changes in the bounds on g_n. The MSE of the matching estimator is slightly smaller than the MSE of TMLE when $\bar{Q}_n$ is inconsistent, and approximately the same as C-TMLE, but the matching estimate is more biased than either TMLE or C-TMLE. TMLE and C-TMLE are able to exploit information that is unavailable to the matching algorithm when $\bar{Q}_n$ is consistent, and thus have lower bias and variance than the matching estimator. These results also indicate that C-TMLE may trade off a small increase in bias for a larger reduction in variance, relative to TMLE, thus minimizing overall MSE.

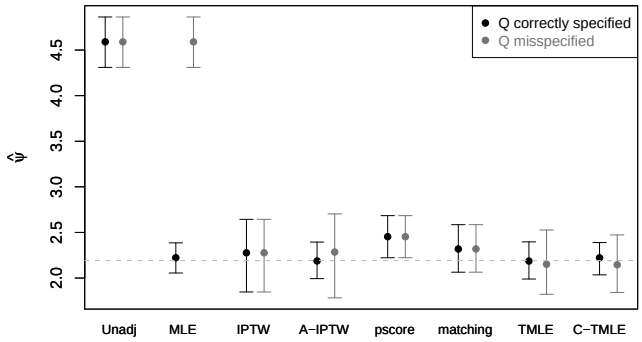

Figure 4.6: Mean estimates and (0.025, 0.975) quantiles, $g_n(1 \mid W)$ bounded at (0.025, 0.975), Q correctly specified (l) and misspecified (r). Dashed line at true parameter value.

MSE provides only one of several points of comparison for estimators. Minimizing MSE is an important goal, and as we've just seen, C-TMLE can make a beneficial data-adaptive trade-off, but Figure 4.6 illustrates that relying on a biased, low-variance, low MSE estimator such as the pscore estimator, can be problematic. The plot in Figure 4.6 shows the mean and (0.025, 0.975) quantiles of the estimates obtained from the 1000 generated samples. 91% of the pscore estimates were larger than ψ_0. This suggests that though an estimate far from the null with a tight confidence interval may look convincing, it might in fact be misleading, and that confidence intervals for the pscore estimator might fail to achieve the nominal coverage rate under circumstances resembling those found in this simulation. This is in marked contrast to TMLE and C-TMLE, double-robust, unbiased, locally efficient substitution estimators that have good properties across a range of data-generating distributions.

4.7.3 Summary

Simulation 5 posed a challenging estimation problem that serves to clarify the distinctions between double robust and non-double robust estimation, and between substitution estimators that remain within the bounds of the semi-parametric model and those that do not. TMLE coupled with the logistic fluctuation for binary and bounded, continuous outcomes is doubly robust to violation of assumptions on Q_0 and g_0, and also robust to outliers. C-TMLE's collaborative, targeted approach to nuisance parameter estimation strengthens that robustness, and is especially valuable for sparse data.

Table 4.6: Simulation 5, comparison of estimators at different bounds on $g_n(1 \mid W)$, $\psi_0 = 2.192$.

	Q **correctly specified**				Q **misspecified**			
	Bias	Var	MSE	Rel MSE*	Bias	Var	MSE	Rel MSE*
g_n **bound=(0, 1)**								
Unadj	2.398	0.021	5.771	1.000	2.398	0.021	5.771	1.000
MLE	0.031	0.007	0.008	0.001	2.398	0.021	5.771	1.000
IPTW	0.018	0.090	0.090	0.016	0.018	0.090	0.090	0.016
AIPTW	−0.006	0.011	0.011	0.002	0.001	0.157	0.157	0.027
pscore	0.262	0.014	0.083	0.014	0.262	0.014	0.083	0.014
matching	0.124	0.018	0.033	0.006	0.124	0.018	0.033	0.006
TMLE	−0.007	0.011	0.011	0.002	−0.018	0.049	0.049	0.008
C-TMLE	0.030	0.008	0.009	0.002	−0.038	0.033	0.034	0.006
g_n **bound=(0.01, 0.99)**								
Unadj	2.398	0.021	5.771	1.000	2.398	0.021	5.771	1.000
MLE	0.031	0.007	0.008	0.001	2.398	0.021	5.771	1.000
IPTW	0.033	0.063	0.064	0.011	0.033	0.063	0.064	0.011
AIPTW	−0.005	0.011	0.011	0.002	0.024	0.092	0.093	0.016
pscore	0.262	0.014	0.083	0.014	0.262	0.014	0.083	0.014
matching	0.125	0.018	0.033	0.006	0.125	0.018	0.033	0.006
TMLE	−0.006	0.011	0.011	0.002	−0.024	0.044	0.044	0.008
C-TMLE	0.030	0.008	0.009	0.002	−0.041	0.032	0.034	0.006
g_n **bound=(0.025, 0.975)**								
Unadj	2.398	0.021	5.771	1.000	2.398	0.021	5.771	1.000
MLE	0.031	0.007	0.008	0.001	2.398	0.021	5.771	1.000
IPTW	0.085	0.041	0.049	0.008	0.085	0.041	0.049	0.008
AIPTW	−0.004	0.010	0.010	0.002	0.093	0.055	0.064	0.011
pscore	0.262	0.014	0.083	0.014	0.262	0.014	0.083	0.014
matching	0.127	0.018	0.034	0.006	0.127	0.018	0.034	0.006
TMLE	−0.005	0.010	0.010	0.002	−0.040	0.031	0.032	0.006
C-TMLE	0.030	0.008	0.009	0.002	−0.046	0.027	0.029	0.005

*Relative to unadjusted estimator

Chapter 5

Relative Performance of Targeted Maximum Likelihood Estimators Under Sparsity

5.1 Introduction

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 wishes 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. This chapter presents a number of different targeted maximum likelihood estimators (TMLEs). We demonstrate that TMLEs, particularly those 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 by Freedman and Berk (2008), and in modified simulations with even greater estimation challenges.

5.2 Kang and Schafer Simulations

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, 2000a); Neugebauer and van der Laan (2005); Petersen et al. (2010). A particular class of estimators are so called double robust estimators (see, 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 (Robins (1986, 1987, 2000a); Robins and Wang (2000); van der Laan and Robins (2003)), showing simulations demonstrating the extreme sparsity bias of IPCW-estimators (e.g., Neugebauer and van der Laan (2005)), diagnosing violations of the positivity assumptions in response to this concern (Petersen et al. (2010); Wang et al. (2006a); Moore et al. (2009); Cole and Hernan (2008); Kish (1992); Bembom and van der Laan (2008)), data adaptive selection of the truncation constant to control the influence of weighting (Bembom and van der Laan (2008), and selecting

parameters that are relying on realistic assumptions (see van der Laan and Petersen (2007), and Petersen et al. (2010)).

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 chapter. 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 (2006a); 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 Section 5.2.5.

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, we add targeted maximum likelihood estimators (TMLEs), or more generally, targeted minimum loss based estimators (van der Laan and Rubin (2006a)) 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 bias reduction of 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 chapter, 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 chapter is organized as follows. Section 5.2.1 presents notation, which deviates from that presented in KS, for the data structure and parameter of interest. Section 5.2.2 formally defines the positivity assumption and gives an overview of causes, diagnostics and responses to violations. Section 5.2.3 defines the estimators on which we focus in this paper, including a sample of estimators in the literature and TMLEs. Section 5.2.4 compares estimator performance in the original and modified KS simulations. Section 5.2.5 then looks at coupling TMLEs with machine learning. Section 5.2.6 concludes with a discussion of the findings.

5.2.1 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:

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

5.2.2 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:

$$g_0(\Delta = 1 \mid W) > 0 \; P_0\text{-almost everywhere.} \tag{5.1}$$

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.

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

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 (^0_n), or as we introduce below, a refit ($^{'}_n$) or a fluctuated fit ($^*_n$).) 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 AIPTW 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 TMLEs as defined in this paper 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. 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 $h_n = (h^T_{n,1}, h^T_{n,2})$,

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

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 $\bar{Q}^0_n$. TanRV relies on the

empirical efficiency maximization estimator of Rubin and van der Laan (Rubin and van der Laan, 2008),

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

TMLEs

We compare the above estimators with several versions of TMLEs. The targeted maximum likelihood procedure was first introduced in van der Laan and Rubin (2006a). 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 assumptions.

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:

$$\mu_{n,TMLE} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(W_i), \tag{5.2}$$

where we use the logistic regression submodel

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

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:

$$-L(\bar{Q})(O) = \Delta\left\{Y \log \bar{Q}(W) + (1-Y)\log(1 - \bar{Q}(W))\right\}. \tag{5.3}$$

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

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

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 (5.2). That is, we use the same logistic regression submodel, and the same loss function (5.3), 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 $\bar{Q}_0$ (see Chapter 3, and Wedderburn (1974)):

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

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 (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 chapter. We now compute the above TMLE of $\mu^*(P_0)$, 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 (2006a) 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 Chapter 3, 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 presented in Chapter 3, where it was demonstrated that the previously observed sensitivity of these two estimators to the positivity assumption was due to those specific choices.

Our TMLE for continuous outcomes that uses a squared error loss and linear fluctuation function, uses the same clever covariate as introduced by Scharfstein et al. (1999). However, as also discussed in an addendum to Rosenblum and van der Laan (2010), the Scharfstein et al. (1999) it is a special type TMLE due to using a clever parametric regression as initial estimator, thereby removing the need for the TMLE-update, but also restricting the estimator to parametric regression models. Both of these TMLEs (squared error loss and linear fluctuation) suffer from the same sensitivity to lack of positivity.

Finally, a natural extension of all of the above TMLEs is to make a more sophisticated estimate of g_0. Therefore, estimator $\mu_{n,C-TMLE}$ is defined by (5.2) 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 Chapter 3, 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 w.r.t. 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).

5.2.4 Simulation Studies

In this section, we compare the performance of TMLEs to the estimating equation-based DR estimators (WLS, AIPTW, 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 and C-TMLE estimators based on the squared error loss function and the linear regression submodel, as well as TMLE, TMLEY*, C-TMLE and C-TMLEY* 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.

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:

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

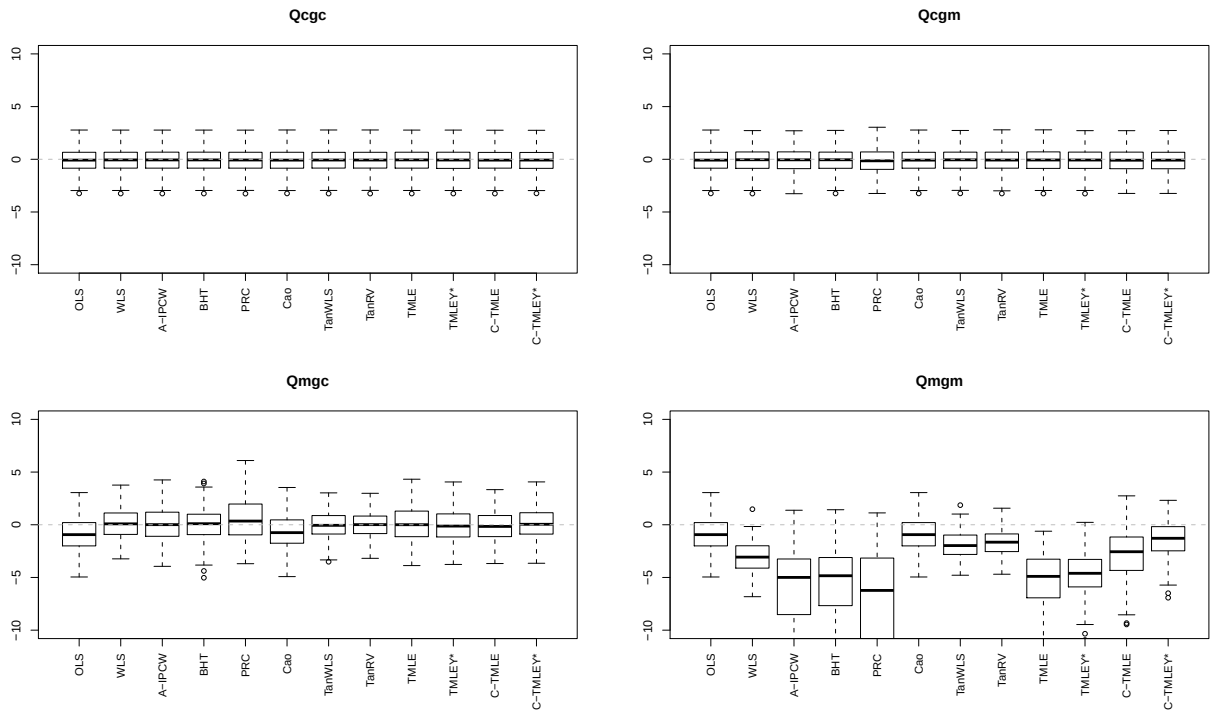

Figure 5.1: Sampling distribution of $(\mu_n - \mu_0)$ with no bounding of g_n, Kang and Schafer simulation.

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 5.1 and Table 5.1 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-TMLEY* are the 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-TMLEY*,

Table 5.1: 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	−0.09	1.40	1.41	−0.09	1.39	1.39	0.10	2.52	2.52	−15.26	1.4e+4	1.4e+4
TMLEY*	−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	−0.10	1.40	1.41	−0.11	1.39	1.40	−0.19	1.90	1.93	−2.84	5.80	13.81
C-TMLEY*	−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 5.2: 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	−0.09	1.40	1.41	−0.09	1.41	1.41	0.09	2.48	2.48	−3.60	2.56	15.48
TMLEY*	−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	−0.10	1.40	1.41	−0.10	1.40	1.41	−0.25	2.23	2.28	−2.96	3.15	11.89
C-TMLEY*	−0.10	1.40	1.41	−0.10	1.40	1.41	0.11	1.74	1.74	−1.37	2.30	4.16

WLS have the smallest MSE, and TanRV, TanWLS are least biased. The superior performance of both Tan estimators can in part be attributed to their internal bounding of g_n.

Figure 5.2 and Table 5.2 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 and TMLE 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.

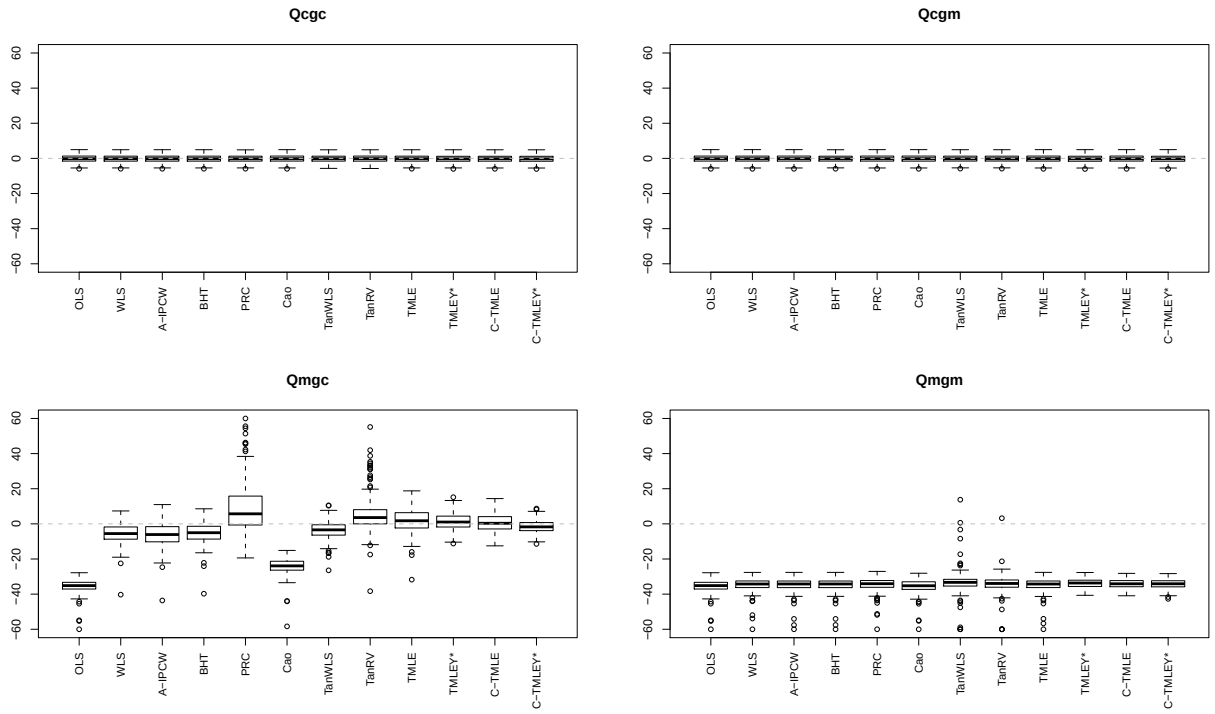

Figure 5.2: Sampling distribution of $(\mu_n - \mu_0)$ with g_n bounded at 0.025, Modification 1 of Kang and Schafer simulation.

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 $[2.2 \times 10^{-16}, 0.87]$.

Table 5.3 present 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, TMLE, and C-TMLE. The variance and MSE of C-TMLEY* is less than half of the other non-TMLE estimators. These results also demonstrate the effect of implementing a logistic fluctuation for TMLE and C-TMLE on bias and variance. TMLE is nearly twice as biased as TMLEY*, and twice as variable. The slight increase in bias of C-TMLEY* relative to the bias of C-TMLE is mitigated by the increase in efficiency. 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 $\bar{Q}_n^0$.

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. This unnecessary inclusion could unnecessarily introduce positivity violations. Moreover, this unnecessary inclusion

Table 5.3: 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	0	42.07	2.4e+3	4.2e+3	-5.4e+9	5.1e+22	5.1e+22
	0.025	1.87	46.81	50.12	-34.75	16.93	1.2e+3
TMLEY*	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	0	4.93	89.20	1.1e+2	-34.08	6.29	1.2e+3
	0.025	0.45	23.20	23.31	-34.10	6.58	1.2e+3
C-TMLEY*	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

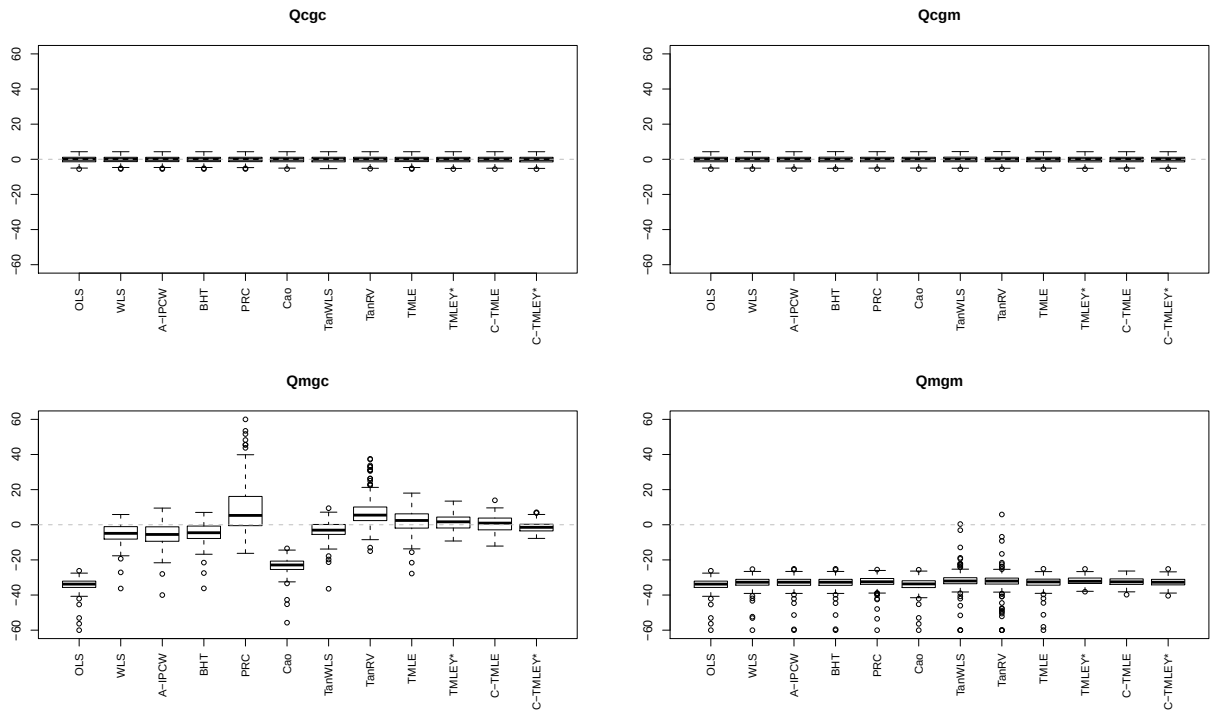

Figure 5.3: Sampling distribution of $(\mu_n - \mu_0)$ with g_n bounded at 0.025, Modification 2 of Kang and Schafer simulation.

can itself introduce substantial bias and inflated variance, sometimes referred to as Z-bias. If the relationships between the variables are linear, the inclusion on non-confounders in the censoring mechanism will always increase bias (Bhattacharya and Vogt, 2007; Wooldridge, 2009). In the non-parametric case, the direction of the bias is less straightforward, but increasing bias is a real possibility (Pearl, 2010). While this problem is not presented in the Kang and Schafer paper nor the responses, it is highlighted in the literature, including Bhattacharya and Vogt (2007); Wooldridge (2009) and Pearl (2010).

Figure 5.3 and Table 5.4 show that C-TMLE estimators have similar or 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."

Table 5.4: 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	0	41.04	2.1e+3	3.8e+3	1.5e+10	9.5e+22	9.5e+22
	0.025	2.13	41.48	45.84	-33.02	16.10	1.1e+3
TMLEY*	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	0	4.60	65.60	86.47	-32.48	6.21	1.1e+3
	0.025	0.72	19.79	20.23	-32.43	6.09	1.1e+3
C-TMLEY*	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

Modification 3 of Kang and Schafer Simulation

In some rare cases, C-TMLEs can be a super efficient estimator because they use a collaborative estimator g_n that takes into account the fit of the initial estimator $\bar{Q}_n^0$ (van der Laan and Gruber, 2009). As a consequence, it is of particular interest to investigate the behavior of C-TMLEs 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. We report the results for $C \in \{10, 20, 50\}$. Table 5.5 provides the results for all estimators when $\bar{Q}_n^0$ is misspecified, g_n bounded at 0.025 for each level of C. We note that neither C-TMLE nor C-TMLEY* break down, even under these particularly challenging conditions.

Table 5.5: 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
				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	2.05	41.84	45.85	2.04	41.85	45.86	2.04	41.91	45.89
TMLEY*	1.17	18.03	19.34	1.17	18.02	19.32	1.16	18.02	19.29
C-TMLE	0.51	20.18	20.35	0.40	20.54	20.61	0.39	20.74	20.81
C-TMLEY*	−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	−33.05	16.06	1.1e+3	−33.07	16.05	1.1e+3	−33.14	16.03	1.1e+3
TMLEY*	−32.03	5.57	1.0e+3	−32.05	5.56	1.0e+3	−32.12	5.54	1.0e+3
C-TMLE	−32.41	5.75	1.1e+3	−32.45	5.64	1.1e+3	−32.59	6.08	1.1e+3
C-TMLEY*	−32.64	5.82	1.1e+3	−32.74	5.94	1.1e+3	−32.75	6.22	1.1e+3

5.2.5 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 TMLE. The C-TMLE (C-TMLEY*) 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 fluctuated on the logit scale (TMLEY*) is larger than the MSE of C-TMLEY*, with increased bias and variance. We ask how the estimation process would be affected if we assume that 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 to incorporate data-adaptive machine learning.

We coupled super learning with TMLE and C-TMLE to estimate both $\bar{Q}_0$ and g_0. For C-TMLEY*, 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-TMLEY* used to build a series of candidate nested logistic regression estimators of the missingness mechanism and corresponding candidate TMLEs. The C-TMLEY* 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, 2002b; 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, 2002a; Friedman, 1994)

* only for binary outcomes, added to library for estimating g

Results

Table 5.6 reports the results when super learning is incorporated into TMLEY* and C-TMLEY* 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. TMLEY* efficiency improved by a factor of 8.5, and C-TMLEY* 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 5.6: Results with and without incorporating super learning into TMLE and C-TMLE, $Qmgm$,
g_n truncated at 0.025.

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

5.2.6 Discussion

By mapping continuous outcomes into [0,1] and using a logistic fluctuation, we show that the TMLEs (both TMLEY* and C-TMLEY*) 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 TMLEY* is bounded by $1/n$ while many of the other estimators do not have such a robustness property. We also show that C-TMLEs have 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 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-TMLEs still perform well.

Related work is also being done with respect to other parameters of interest. 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, as is discussed in the next section.

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 contribute 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 w.r.t. 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 loss function and submodel can affect the finite sample robustness, as observed in the current simulations. 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.

5.3 Freedman and Berk Simulations

We have seen that sparse data poses a challenge to efficient unbiased estimation of statistical parameters. Estimates obtained from parametric regression procedures rely on model-based extrapolation, and can be biased under model misspecification. Propensity score-based estimates typically have high variance under sparsity due to large weights on rare observations. Efforts to control the variance by using stabilized or truncated weights can introduce bias into the estimate. The performance of double robust (DR) estimation procedures that do not respect known bounds on the model is similarly impaired. TMLE and C-TMLE, DR methods that exploit this knowledge, are more robust in situations where sparsity affects the identifiability of the parameter of interest.

Freedman and Berk (2008) (FB) compares weighted and unweighted regression approaches to estimating coefficients in parametric causal models, with the intention of demonstrating that in many situations propensity score weighting can increase bias and/or variance of the estimates relative to unweighted regression, even when the true propensity score model is known. We apply targeted maximum likelihood estimation (TMLE), collaborative targeted maximum likelihood estimation (C-TMLE), and and augmented IPTW (AIPTW) estimator, three DR estimators, to estimation problems proposed in FB, estimation of the marginal additive treatment effect of a binary treatment on a continuous outcome.

The additive treatment effect is defined non-parametrically as $\psi_0 = E_W\{E(Y|A=1,W) - E(Y|A=0,W)\}$, where n i.i.d. copies of $O = (W, A, Y) \sim P_0$ is observed data, with outcome Y, binary treatment assignment A, and covariates W. This parameter is numerically equal to the coefficient in front of the treatment variable in a linear regression model where no terms involve interactions with treatment.

FB simulation 1 presents weighted and unweighted linear regression results based on the correct model and two misspecified parametric models, using a data-generating distribution that violates the positivity assumption that the conditional probability of treatment given the covariates is bounded away from 0 and 1. Actual treatment assignment probabilities are between (0.03 and 0.99995). We first formally define each estimator, then present results from applying each estimator to FB simulation 1, and additional results using modified data-generating distributions that provide insight into estimator performance.

5.3.1 FB Simulations

Three data-generating distributions are defined. For each one, 250 samples of size $n = 1000$ are drawn from the given data generating distribution. A propensity score $g_0(A \mid W)$ is estimated using the correct probit model for treatment. The correct linear regression model for the outcome and two increasingly misspecified models are defined as

correct model: $Y \sim A + W_1 + W_2$,

misspecified model 1: $Y \sim A + W_1$,

misspecified model 2: $Y \sim A$.

Estimates of the marginal additive treatment effect are obtained based on each of these parametric models paired with $g_n(A \mid W)$, a maximum likelihood estimator according to a correct parametric model for g_0.

Data generation

Simulation 1 replicates FB simulation 1. Both covariates, W_1 and W_2, confound the relationship between treatment and the outcome, so we expect OLS to be biased when the regression model for Q is misspecified. Incorporating estimated propensity scores should allow the remaining estimators to be unbiased, at the cost of higher variance.

$$\begin{aligned} Y &= a + bX + c_1W_1 + c_2W_2 + U,\ U \sim N(0,1), \\ g_{0,1} = P(A=1 \mid W) &= \Phi(e + f_1W_1 + f_2W_2), \end{aligned}$$

$$(W_1, W_2) \text{ is bivariate normal, } N(\mu, \Sigma), \text{ with} \mu_1 = 0.5,\ \ \mu_2 = 1,\ \ \Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

with $a = b = c_1 = d = 1, c_2 = 2, e = 0.5, f_1 = 0.25, f_2 = 0.75$. Φ is the CDF of the standard normal distribution, thus the treatment mechanism conforms to a probit model.

These settings for d, f_1, and f_2 lead to practical violations of the positivity assumption, also known as the experimental treatment assignment (ETA) assumption, that the probability of receiving treatment at all levels given baseline covariates is bounded away from 0 and 1. Conditional treatment probabilities $g(1, W) = P(A = 1|W)$ range from .03 to 0.99995.

Simulation 2 allows us to examine the effect of including a covariate in the treatment mechanism model that is predictive of treatment but not of the outcome. The only difference between simulation 1 and simulation 2 is the treatment assignment mechanism.

$$g_{0,2} = P(A = 1 \mid W) = \Phi(e + f_1 W_1 + f_2 W_2 + f_3 W_3).$$

$W_3 \sim N(0, 1), e = 0.15, f_1 = .075, f_2 = .225, f_3 = -0.9$. Both W_1 and W_2 are less predictive of treatment than in simulation 1. W_3 is a strong predictor of treatment, but not associated with the outcome. This treatment assignment mechanism lead to sparsity. The true conditional probability of receiving treatment is between (0.004 and 0.99995). We expect the C-TMLE estimator to give strong preference to covariates W_1, W_2 when building the candidate treatment mechanism estimators, and because of a tendency to exclude W_3 from the model for g_n have smaller variance in the parameter estimates across all samples.

Simulation 3 demonstrates that weighting can introduce bias in the estimate of the additive treatment effect under sparsity, even when the the correct propensity score model is known. In this simulation,$P(A = 1 \mid W)$ is between (0.0003 and 0.9997). The linear form of the relationships between the covariates and the outcome is unchanged, but the strengths of those relationships are altered to weaken the association between W_1 and W_2, and between W_2 and A, but strengthen the relationships between W_1 and Y and W_2 and Y. As in simulation 2, W_3 is associated with A, but not with the outcome Y.

$$\begin{aligned} Y &= a + bA + c_1 W_1 + c_2 W_2 + U,\ U \sim N(0, 1), \\ P(A = 1 \mid W) &= \Phi(f_1 W_1 + f_2 W_2 + f_3 W_3), \end{aligned}$$

$$(W_1, W_2) \text{ is bivariate normal, } N(\mu, \Sigma), \text{ with } \mu_1 = 0.5, \quad \mu_2 = 1, \quad \Sigma = \begin{bmatrix} 2 & 0.1 \\ 0.1 & 1 \end{bmatrix},$$

with $a = b = d = 1, c_1 = 5, c_2 = 10, f_1 = 0.25, f_2 = 0.001, f_3 = 1$.

5.4 Results

OLS, WLS, AIPTW, TMLE, and C-TMLE estimators were applied to each simulated dataset. Two sets of C-TMLE results were obtained. For the first, labeled C-TMLE in Tables 5.7- 5.9, the

covariate set W used to create the series of treatment mechanism estimators is restricted to main term covariates. In the second set, labeled C-TMLE(augW), W is augmented with four terms corresponding to the propensity score estimate supplied to all other estimators and truncated propensity scores, truncated at level $(p, 1-p)$, with p set to (0.1, 0.25, 0.5).

In all simulations, when the model for Q is correctly specified OLS, the unweighted parametric estimator, has the smallest MSE, but when Q is misspecified all other estimators out-perform OLS with respect to both MSE and bias. Simulation 1 results suggest that TMLE and C-TMLE are more robust than WLS and AIPTW to sparsity. C-TMLE results in Simulation 2 demonstrate that performance improves under sparsity by using a procedure that estimates only the necessary portion of the treatment mechanism. C-TMLE's MSE and variance are superior to the other estimators that incorporate propensity score estimates. Under extreme misspecification (misspecified model 2) bias is almost entirely removed. In simulation 3 augmenting the covariate set improves performance of the C-TMLE estimator. Augmentation confers the greatest benefit when Q is most severely misspecified.

5.4.1 Discussion

FB cautions against blindly relying on weights to remove bias due to possible model misspecification. Results presented here indicate that even under extreme sparsity in the data, intelligently incorporating weights in double robust estimation procedures does little harm when the model is correctly specified, and can greatly reduce bias and MSE in the more common situations when the model is misspecified. In addition to being a double robust substitution estimator that respects global constraints C-TMLE's internal collaborative estimation of g proved especially robust in this setting.

Domain knowledge can be incorporated into both stages of the TMLE and C-TMLE estimation procedures. One example is the use of the augmented covariate set when the true treatment assignment mechanism is known. The strength of this approach is most clearly illustrated in simulation 3 with Q modeled as $Qmis2$, where the right thing to do is adjust for all covariates, yet that causes strong ETA violations. In this case, the inclusion of truncated propensity scores in W offered a more refined choice beyond simply including or excluding an entire covariate. These additional terms can be helpful in situations where including a particular covariate causes an ETA violation, but in fact, experimentation is lacking in only some portion of the covariate values.

Table 5.7: Simulation 1.

	g_n **unbounded**				g_n **bound = (0.025, 0.975)**			
	Bias	Var	MSE	RelMSE*	Bias	Var	MSE	RelMSE*
Unadj	4.061	0.046	16.538		4.061	0.046	16.538	
Correct Model								
OLS	0.010	0.009	0.010	1.000	0.010	0.009	0.010	1.000
WLS	0.012	0.039	0.039	4.144	0.016	0.024	0.024	2.526
AIPTW	0.019	0.058	0.059	6.153	0.014	0.017	0.017	1.766
TMLE	0.190	0.475	0.509	53.460	0.019	0.027	0.027	2.834
C-TMLE	0.004	0.014	0.014	1.449	0.013	0.013	0.013	1.410
C-TMLE (augW)	0.011	0.010	0.010	1.092	0.014	0.014	0.014	1.501
Misspecified Model 1								
OLS	1.138	0.020	1.314	1.000	1.138	0.020	1.314	1.000
WLS	0.133	0.115	0.133	0.101	0.295	0.040	0.127	0.096
AIPTW	0.120	0.344	0.357	0.272	0.433	0.033	0.220	0.167
TMLE	−0.588	0.380	0.724	0.551	−0.001	0.048	0.048	0.037
C-TMLE	0.262	1.516	1.579	1.202	−0.412	0.098	0.267	0.203
C-TMLE (augW)	−0.242	1.068	1.122	0.854	−0.077	0.054	0.060	0.046
Misspecified Model 2								
OLS	4.061	0.046	16.538	1.000	4.061	0.046	16.538	1.000
WLS	0.431	0.660	0.843	0.051	1.070	0.091	1.234	0.075
AIPTW	0.381	3.039	3.172	0.192	1.507	0.130	2.402	0.145
TMLE	−0.451	1.392	1.590	0.096	−0.132	0.120	0.137	0.008
C-TMLE	1.885	5.358	8.889	0.537	0.456	0.276	0.482	0.029
C-TMLE (augW)	−0.046	0.158	0.160	0.010	0.011	0.063	0.063	0.004

*relative to OLS estimator using the same model specification

Table 5.8: Simulation 2.

	g_n **unbounded**				g_n **bound = (0.025, 0.975)**			
	Bias	Var	MSE	RelMSE*	Bias	Var	MSE	RelMSE*
Unadj	1.136	0.043	1.333		1.136	0.043	1.333	
Correct Model								
OLS	0.001	0.004	0.004	1.000	0.001	0.004	0.004	1.000
WLS	0.004	0.012	0.012	2.921	0.003	0.008	0.008	1.977
AIPTW	0.005	0.016	0.016	3.810	0.003	0.008	0.008	1.962
TMLE	0.015	0.022	0.022	5.394	0.004	0.009	0.009	2.250
C-TMLE	0.000	0.004	0.004	1.068	0.001	0.004	0.004	1.056
C-TMLE (augW)	0.002	0.005	0.005	1.155	0.001	0.005	0.005	1.211
Misspecified Model 1								
OLS	0.275	0.010	0.086	1.000	0.275	0.010	0.086	1.000
WLS	−0.005	0.030	0.030	0.349	0.016	0.012	0.013	0.147
AIPTW	−0.009	0.041	0.041	0.476	0.020	0.012	0.013	0.149
TMLE	−0.102	0.035	0.045	0.529	−0.020	0.014	0.015	0.172
C-TMLE	−0.007	0.008	0.008	0.091	−0.013	0.008	0.008	0.090
C-TMLE (augW)	−0.017	0.008	0.008	0.094	−0.018	0.007	0.008	0.088
Misspecified Model 2								
OLS	1.136	0.043	1.333	1.000	1.136	0.043	1.333	1.000
WLS	0.003	0.090	0.090	0.067	0.076	0.031	0.037	0.027
AIPTW	−0.012	0.143	0.142	0.107	0.094	0.033	0.042	0.031
TMLE	−0.147	0.137	0.158	0.119	−0.074	0.041	0.046	0.034
C-TMLE	−0.001	0.005	0.005	0.004	−0.001	0.005	0.005	0.004
C-TMLE (augW)	0.008	0.010	0.010	0.007	0.009	0.009	0.009	0.007

*relative to OLS estimator using the same model specification

Table 5.9: Simulation 3.

	g_n **unbounded**				g_n **bound = (0.025, 0.975)**			
	Bias	Var	MSE	RelMSE*	Bias	Var	MSE	RelMSE*
Unadj	3.022	0.688	9.816		3.022	0.688	9.816	
Correct Model								
OLS	0.002	0.004	0.004	1.000	0.002	0.004	0.004	1.000
WLS	0.002	0.012	0.012	3.175	0.004	0.009	0.009	2.476
AIPTW	0.004	0.018	0.018	4.694	0.004	0.009	0.009	2.470
TMLE	0.001	0.067	0.067	17.676	0.002	0.011	0.011	3.003
C-TMLE	0.002	0.004	0.004	0.991	0.002	0.004	0.004	0.989
C-TMLE (augW)	0.001	0.004	0.004	1.044	0.001	0.004	0.004	1.059
Misspecified Model 1								
OLS	0.024	0.447	0.446	1.000	0.024	0.447	0.446	1.000
WLS	−0.108	0.500	0.510	1.143	−0.037	0.223	0.224	0.501
AIPTW	−0.144	0.830	0.847	1.898	−0.037	0.223	0.224	0.502
TMLE	−0.127	1.077	1.089	2.440	−0.053	0.291	0.293	0.656
C-TMLE	−0.077	0.050	0.056	0.125	−0.077	0.047	0.053	0.118
C-TMLE (augW)	−0.091	0.042	0.050	0.112	−0.094	0.045	0.054	0.120
Misspecified Model 2								
OLS	3.022	0.688	9.816	1.000	3.022	0.688	9.816	1.000
WLS	−0.077	1.686	1.685	0.172	0.186	0.392	0.425	0.043
AIPTW	−0.167	3.727	3.740	0.381	0.232	0.406	0.459	0.047
TMLE	−0.940	1.357	2.235	0.228	−0.294	0.555	0.639	0.065
C-TMLE	0.002	0.073	0.073	0.007	−0.005	0.021	0.021	0.002
C-TMLE (augW)	−0.049	0.073	0.075	0.008	−0.033	0.045	0.046	0.005

*relative to OLS estimator using the same model specification

Chapter 6

An Application of Targeted Maximum Likelihood Estimation to the Meta-Analysis of Safety Data

6.1 Introduction

Safety analysis poses a challenging statistical problem because a study that is powered for estimating the effect of treatment on an efficacy outcome, is typically under-powered with respect to estimating the effect of treatment on rare adverse events. In addition, the follow-up period for collecting safety data may extend beyond the primary endpoint, thus safety data are increasingly subject to censoring. An estimator that does not account for informative censoring will be biased, and in addition, even when censoring is non-informative, sacrifices efficiency. This chapter compares the performance of TMLE with that of several other estimators in the meta-analysis of real and simulated data.

Multi-center phase three randomized controlled trials (RCT) were conducted to study the effectiveness of a new antibiotic for a range of conditions with the potential to severely compromise the health of the patient. Four different comparator drugs were used in the trials' control arms, corresponding to four distinct indications for treatment among the trials. Though RCTs are designed to investigate the treatment effect on some primary outcome, the effect on the occurrence of adverse events is often also of interest. Several estimation methodologies described below were used to analyze safety data from seven RCTs to estimate the effect of treatment with the study drug versus a comparator drug on subsequent mortality.

Each study is summarized in Table 6.1. The raw numbers in the table indicate that with the exception of study 5 (indication C), in each study there were more deaths in the treatment arm than in the control arm. Overall, 4.0% of subjects given the study drug died, as compared to 3.2% of subjects randomized to a comparator drug. Estimates of risk differences (RD), risk ratios (RR), and odds ratios (OR) stratified by the indication for treatment were obtained. Additional pooled

Table 6.1: Summary of subjects in each study.

Indication	Study	Number of subjects		Number of deaths		
		Treatment	Control	Treatment	Control	Total
A	1	295	288	4	1	5
	2	275	271	1	0	1
B	3	417	416	16	10	26
	4	409	415	5	3	8
C	5	212	213	4	4	8
	6	220	214	6	4	10
D	7	473	471	55	52	107

parameter estimates were calculated as a weighted combination of the indication-specific effects.

Although each study is an RCT, because there is variation in the indication for treatment and the comparator drug, the treatment effect is likely to be heterogeneous. In addition, because differences in study populations might confound the effect of treatment on mortality when studies are combined, an unadjusted estimate has the potential to be biased. Targeted maximum likelihood estimation (TMLE), a methodology designed to exploit covariate information to reduce bias, was therefore applied to estimate the stratified and pooled effect of the study drug on mortality (van der Laan and Rubin, 2006a; van der Laan et al., 2009). TMLE results are compared with those of other causal effect estimators in the literature, the inverse-probability-of-treatment-weighted estimator (IPTW) (Hernan et al. (2000b), Robins (2000b)), the maximum likelihood based G-computation estimator (Gcomp) (Robins, 1986), and the augmented IPTW estimator (AIPTW) (Robins and Rotnitzky (2001); Robins et al. (2000); Robins (2000a)). These estimators are defined in Section 6.2.

Section 6.3 describes the meta-analysis of the RCT data. Estimates and 95% confidence intervals obtained for each estimator are quite similar to the unadjusted regression of mortality on treatment. Confidence intervals for all stratified parameter estimates include the null value. The 95% confidence interval for the pooled estimate of the risk difference just barely excludes the null. Agreement between the naive unadjusted estimator and the more sophisticated estimation procedures does not mean that the use of a sophisticated estimation procedure was unwarranted. On the contrary, the only assurance we have that the unadjusted estimate is not biased is that more advanced procedures confirm the results of the naive approach. This implies that estimators that perform equally well in both the presence or absence of confounding should be used routinely.

A simulation study was designed to illustrate the relative performance of each estimator when there is missingness in the outcome, (Section 6.4). Missing outcome data is not unusual when adverse events are self-reported, and/or compiled from post-followup or post-market data. When not adequately addressed, this reporting bias might lead to either false acceptance or false rejection

of the null hypothesis of no treatment effect. Simulation results indicate that efficient double-robust estimators (AIPTW and TMLE) work well in a variety of settings, with TMLE having the additional property that estimates are guaranteed to remain within the parameter space.

6.2 Statistical Methods

The dataset contains observations from seven studies involving $K = 4$ unique indications for treatment. Consider the data for indication k as n_k i.i.d. copies of $O_k = (W_k, A_k, Y_k) \sim P_{0,k}$, where W_k is a vector of baseline covariate information, A_k is a binary treatment indicator variable, ($A_k = 1$ for treatment, $A_k = 0$, for control), and Y_k is a binary outcome variable set to 1 if mortality occurs.

Observations $O_{1,k}, ..., O_{n_k}$ can be viewed as realizations of random variables sampled from an underlying data generating distribution, $P_{0,k}$. $P_{0,k}$ factorizes into $(Q_{0,k}, g_{0,k})$, where $Q_{0,k}$ is the true conditional distribution of outcome Y_k given A_k and W_k, and the marginal distribution of covariate W. $g_{0,k}$ is the treatment assignment mechanism, the conditional probability that $A_k = 1$ given W_k. In this analysis of RCT data treatment assignment probabilities are known to be 0.5 for all subjects. Though the true $g_{0,k}$ is known, efficiency may be improved when this is estimated from the data (van der Laan and Robins, 2003). $Q_{0,k}$ is unknown, and therefore must be estimated from the data.

A comparison of the effect of the study drug on subsequent mortality versus a comparator drug can be quantified as a parameter of $P_{0,k}$. Let $\mu_{1,k} = E_{W_k}(E(Y_k \mid A_k = 1, W_k)$ and $\mu_{0,k} = E_{W_k}(E(Y_k \mid A_k = 0, W_k)$, then the indication-specific marginal risk difference, risk ratio, and odds ratio parameters are defined non-parametrically as:

$$\begin{aligned}
\psi_{0,k}^{RD} &= \mu_{1,k} - \mu_{0,k} \\
\psi_{0,k}^{RR} &= \frac{\mu_{1,k}}{\mu_{0,k}} \\
\psi_{0,k}^{OR} &= \frac{\mu_{1,k}/(1-\mu_{1,k})}{\mu_{0,k}/(1-\mu_{0,k})}.
\end{aligned}$$

Note that these parameters are functions of only the Q portion of the likelihood, and in fact, aside from the marginal distribution of W, only require knowledge of the conditional mean of Y given (A, W), not the entire density.

Five estimation procedures were applied to estimate the stratified and pooled RD, RR, and OR: unadjusted estimation, inverse-probability-of-treatment-weighted (IPTW) estimation, augmented IPTW estimation (AIPTW), maximum likelihood G-computation (Gcomp), and targeted maximum likelihood estimation (TMLE). One (stratified) analysis was carried out for each subset of observations having the same indication for treatment.

In the following descriptions ψ_n denotes a parameter estimate, $\bar{Q}_n(A, W)$ corresponds to an estimate of the regression of Y on treatment assignment A and baseline covariates, W. Indexing

by indication k is suppressed in the notation. $\bar{Q}_n^*(A,W)$ is a targeted estimate of the conditional mean, an update of an initial $\bar{Q}_n^0(A,W)$ designed to reduce bias in the estimate of the parameter of interest, as described more fully below. $A=1$ for subjects in the treatment group, $A=0$ indicates the subject is in the control group, n is the number of observations.

unadj: The unadjusted estimates are functions of the mean mortality in treatment and control arms. The unadjusted estimators for the RD, RR, and OR are given by,

$$\begin{aligned}
\psi_{n,unadj}^{RD} &= \mu_{1,unadj} - \mu_{0,unadj} \\
\psi_{nunadj}^{RR} &= \frac{\mu_{1,unadj}}{\mu_{0,unadj}} \\
\psi_{n,unadj}^{OR} &= \frac{\mu_{1,unadj}/(1-\mu_{1,unadj})}{\mu_{0,unadj}/(1-\mu_{0,unadj})}
\end{aligned}$$

where

$$\begin{aligned}
\mu_{1,unadj} &= \frac{\sum_{i=1}^n I(A_i=1)Y_i}{\sum_{i=1}^n I(A_i=1)} \\
\mu_{0,unadj} &= \frac{\sum_{i=1}^n I(A_i=0)Y_i}{\sum_{i=1}^n I(A_i=0)}.
\end{aligned}$$

When there is no informative missingness and treatment is randomized, this estimator is unbiased.

IPTW: The inverse-probability-of-treament-weighted estimator was calculated as a weighted linear regression of outcome, Y, on treatment indicator, A. The weight for each observation was set to the inverse of the marginal probability of receiving the observed treatment assignment. The IPTW estimators are given by,

$$\begin{aligned}
\psi_{n,IPTW}^{RD} &= \mu_{1,IPTW} - \mu_{0,IPTW} \\
\psi_{n,IPTW}^{RR} &= \frac{\mu_{1,IPTW}}{\mu_{0,IPTW}} \\
\psi_{n,IPTW}^{OR} &= \frac{\mu_{1,IPTW}/(1-\mu_{1,IPTW})}{\mu_{0,IPTW}/(1-\mu_{0,IPTW})}
\end{aligned}$$

where

$$\begin{aligned}
\mu_{1,IPTW} &= \sum_{i=1}^n \frac{I(A_i=1)Y_i}{g_n(1 \mid W_i)} \\
\mu_{0,IPTW} &= \sum_{i=1}^n \frac{I(A_i=0)Y_i}{g_n(0 \mid W_i)}.
\end{aligned}$$

The IPTW estimator is consistent when g_0 is estimated correctly, however it is an inefficient estimator, yielding estimates with wider confidence intervals than those based on more efficient approaches.

AIPTW: The double-robust augmented IPTW estimator incorporates estimates of both the regression of Y on A, W and the propensity score.

$$\begin{aligned}
\psi^{RD}_{n,AIPTW} &= \mu_{1,AIPTW} - \mu_{0,AIPTW} \\
\psi^{RR}_{n,AIPTW} &= \frac{\mu_{1,AIPTW}}{\mu_{0,AIPTW}} \\
\psi^{OR}_{n,AIPTW} &= \frac{\mu_{1,AIPTW}/(1-\mu_{1,AIPTW})}{\mu_{0,AIPTW}/(1-\mu_{0,AIPTW})}
\end{aligned}$$

where

$$\begin{aligned}
\mu_{1,AIPTW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i=1)(Y_i-\bar{Q}_n(1,W_i))}{g_n(1 \mid W_i)} + \bar{Q}_n(1,W_i)\right) \\
\mu_{0,AIPTW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i=0)(Y_i-\bar{Q}_n(0,W_i))}{g_n(0 \mid W_i)} + \bar{Q}_n(0,W_i)\right).
\end{aligned}$$

Gcomp: G-computation estimates of the risk difference parameter are consistent when $\bar{Q}_n$ provides a consistent estimate of the conditional mean of Y given A and W.

$$\begin{aligned}
\psi^{RD}_{n,Gcomp} &= \mu_{1,Gcomp} - \mu_{0,Gcomp} \\
\psi^{RR}_{n,Gcomp} &= \frac{\mu_{1,Gcomp}}{\mu_{0,Gcomp}} \\
\psi^{OR}_{n,Gcomp} &= \frac{\mu_{1,Gcomp}/(1-\mu_{1,Gcomp})}{\mu_{0,Gcomp}/(1-\mu_{0,Gcomp})}
\end{aligned}$$

where

$$\begin{aligned}
\mu_{1,Gcomp} &= \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(1,W_i) \\
\mu_{0,Gcomp} &= \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(0,W_i).
\end{aligned}$$

TMLE: The targeted maximum likelihood estimator is a double-robust substitution estimator.

$$\begin{aligned}
\psi^{RD}_{n,TMLE} &= \mu_{1,TMLE} - \mu_{0,TMLE} \\
\psi^{RR}_{n,TMLE} &= \frac{\mu_{1,TMLE}}{\mu_{0,TMLE}} \\
\psi^{OR}_{n,TMLE} &= \frac{\mu_{1,TMLE}/(1-\mu_{1,TMLE})}{\mu_{0,TMLE}/(1-\mu_{0,TMLE})}
\end{aligned}$$

where

$$\begin{aligned}\mu_{1,TMLE} &= \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n^*(1,W_i)\\ \mu_{0,TMLE} &= \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n^*(0,W_i).\end{aligned}$$

$\bar{Q}_n^*(A,W)$ refers to a targeted estimate of the regression of Y on (A,W), obtained by fluctuating an initial estimate, $\bar{Q}_n^0(A,W)$, in a manner designed to reduce bias in the target parameter estimate. The direction of the fluctuation, H^*, is derived from the efficient influence curve of the target parameter mapping. Because the RD, RR, and OR are all functions of $\mu_1 = EY_1$ and $\mu_0 = EY_0$, these two conditional means are the target parameters. Though it is possible to directly target each of the parameters (RD, RR, OR) in three separate analyses, this leads to redundant calculations, so was not the approach taken here. Instead the two conditional means were targeted simultaneously:

$$\begin{aligned}H_1^*(A,W) &= \frac{I(A=1)}{g_n(1\mid W)}\\ H_0^*(A,W) &= \frac{I(A=0)}{g_n(0\mid W)}\\ \bar{Q}_n^*(A,W) &= \frac{1}{1+e^{-m}},\ m = log\left(\frac{\bar{Q}_n^0(A,W)}{1-\bar{Q}_n^0(A,W)}\right) + \epsilon_1 H_1^*(A,W) + \epsilon_0 H_0^*(A,W),\end{aligned}$$

and ϵ_1, ϵ_0 were fitted using logistic regression.

Targeted maximum likelihood estimation and AIPTW incorporate estimates of both $\bar{Q}_0$ and g_0. These estimators are double robust, meaning that consistent estimation of either $\bar{Q}_0$ or g_0 implies consistent estimation of the parameter, ψ_0. In contrast, the Gcomp estimator is consistent only when $\bar{Q}_n$ is a consistent estimator of $\bar{Q}_0$, and the IPTW estimator relies on consistent estimation of g_0.

Main terms linear regression was used to obtain an estimate of $\bar{Q}_0$ for the AIPTW and Gcomp procedures. The super learner (SL), a data-adaptive prediction algorithm, was used as initial estimator of $\bar{Q}_n^0$ in the TMLE procedure (Polley, 2010; van der Laan et al., 2007). Specifically, a discrete super learner was implemented to select the best prediction algorithm among the candidates in the super learner library and the super learner itself. Honest cross validation was then used to compare super learner performance with an unadjusted regression of Y on A. The unadjusted estimate was used as the initial estimate unless the super learner provided at least a 10% improvement. For indications where the number of events was small (< 25), super learner was not used. The SL library contained prediction algorithms that search over different portions of the model space.

- **glm** generalized linear models, main terms logistic regression (Team, 2010)
- **k-nearest-neighbors** classification using most common outcome among identified k nearest nodes (k set to 10 and 20) (Venables and Ripley, 2002a; Friedman, 1994)
- **gam** generalized additive models, tuning parameter set to 3 and 4,(Hastie, 2009)
- **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)

Although the true treatment assignment probability is known to be 0.5 for all subjects by design, estimating the treatment mechanism can help to adjust for any empirical confounding, and may improve efficiency (van der Laan and Robins, 2003). The treatment mechanism was estimated as the marginal probability of being assigned to treatment or control. The probability for an observation having $A_{ik} = a, a \in \{0, 1\}$ was set equal to the indication-specific empirical proportion of subjects assigned treatment at level a. Note that when g_n is equal to the empirical proportion of subjects in the treatment group the IPCW estimator is equivalent to the unadjusted estimator.

6.3 Data Analysis

6.3.1 Preprocessing steps

Missing data were imputed using covariate data collected in the same study using the *aregImpute* function for predictive mean matching (Harrell Jr, 2010). The imputation was stratified by study; only observations from the same study contributed data for the imputation procedure. A binary imputation indicator variable was created for each covariate that had more than 2% missing values. Tables 6.2 and 6.3 list all covariates that had missing values, and the number of missing values per study. Note that no APACHE scores were recorded for five studies, and were not imputed for these studies. Categorical APACHE covariates for which less than 2% of the observations were missing were assigned the most common study-specific value. Only two APACHE covariates had more than 2% missing, ARTPH in three studies, (3, 4, 7), and OXYG in study 7. A full description of the covariates is given in Table 6.13.

Indicators were created for levels of factor variables having a p-value of ≤ 0.1 for the univariate association with the outcome. Only ten out of forty-two countries were associated with one or more of the outcomes stratified by indication, so ten binary indicator variables were utilized in place the COUNTRY variable, implicitly creating an eleventh country category, "other." Covariates that were not measured for any subject in a particular study were set to 0 for all subjects in that study.

6.3.2 Stratification by indication

When an outcome is rare, as is the case for three of the four indications studied, there is insufficient power to fit large models. A is included in all adjustment sets, and we apply a rule of

Table 6.2: Number of missing values for each covariate, by study. Covariates not listed had no missing values. Study size shown in parentheses.

	Study 1 (583)	Study 2 (546)	Study 3 (833)	Study 4 (824)	Study 5 (425)	Study 6 (434)	Study 7 (944)
BMI	7	4	8	3	6	1	9
APACHE	583	546	5	1	425	434	7
ALBUMIN	47	43	31	116	36	26	42
ALKPHOS	27	27	21	58	28	19	27
ALT	22	5	15	19	26	8	17
AST	19	10	29	16	25	16	19
PROTTOT	43	11	39	50	36	18	30
CREATIN	12	4	10	9	12	1	11
POTASSIU	16	5	11	9	13	3	12
HEMOGLOB	26	3	9	9	10	3	11
WBC	16	3	9	8	9	2	9
PLATELET	38	4	12	11	11	4	24

thumb that sets the maximum number of additional covariates, W, to (2 * num events /15 - 1). For each indication-specific outcome, covariates were ordered by p-value based on each univariate logistic regression of Y on W_i, using as offset the fitted values from a regression of outcome Y on A, on the logit scale. At most w_k covariates having a p-value ≤ 0.1 were incorporated into the adjustment set for each indication (Table 6.5). Fewer covariates were retained for indication D due to a lack of association with the outcome.

6.3.3 Inference

Variances and 95% confidence intervals for the RD, RR, and OR parameters were obtained with the bootstrap and influence curves. Both methods are in good agreement. The large width of the bootstrap confidence intervals for RR and OR for indications 1 and 5 reflects the instability of these estimates under extreme sparsity in the data, and illustrates the challenge of constructing valid confidence intervals in these settings.

The bootstrap Parameter estimates were obtained for 1000 bootstrap samples for each indication and outcome. The (0.025, 0.975) quantiles of the bootstrap estimates provide the bounds on a 95% confidence interval reported in Table 6.6 below.

Influence curve-based inference Theory tells us that $(\psi_n - \psi_0)$, the difference between a parameter estimate obtained from a regular asymptotically linear estimator and the truth, converges

Table 6.3: Number of missing values for each APACHE covariate, by study. Covariates not listed had no missing values. Studies not listed had no APACHE values recorded. Study size shown in parentheses.

	Study 3 (833)	Study 4 (824)	Study 7 (944)
TEMP	6	1	7
ARTPRESS	5	3	12
OXYG	833	824	573
AGEPTS	5	1	7
APACHE	5	1	7
HEARTRAT	5	4	11
RESPRATE	833	824	13
ARTPH	238	244	323
SERSOD	8	3	11
SERPOT	9	3	11
SERCREA	8	3	10
HEMAT	6	6	11
WBCPTS	6	2	8

to a Normal limit distribution,

$$\sqrt(n)(\psi_n - \psi_0) \xrightarrow{D} N(0, \Sigma),$$

where Σ is the covariance matrix of the (possibly multi-dimensional) parameter. In practice, this provides a means for estimating the variance of the estimator as the variance of the empirical influence curve divided by the sample size, n. The parameter-specific influence curves are given below. Asymmetrical confidence intervals for the RR and OR parameters are constructed on the log scale, based on the influence curves for the log(RR) and log(OR), respectively. Estimates and 95% confidence intervals are reported in Table 6.7.

Table 6.4: Number of events and maximum size of adjustment set for each stratified analysis.

Indication	num events	w_k max adj set size
A	6	0
B	34	3
C	18	1
D	107	13

Table 6.5: Covariates included in adjustment set for each stratified analysis.

Indication	Covariates
A	*(none)*
B	AGEPTS, I.APACHE, CHFHIST
C	SVK
D	ACIN, AUS, BACTER, CHFHIST, COPDHIST, HRV, I.APACHE, IND, PER

note: covariate names beginning with I are imputation or study id indicator variables

$$IC^{RD}(O) = \left(\frac{A}{g_0(1 \mid W)} - \frac{1-A}{g_0(0 \mid W)}\right)(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) - \bar{Q}_0(A,W) - \psi_0$$

$$\begin{aligned} IC^{logRR}(O) &= \frac{1}{\mu_{10}}\left(\frac{A}{g_0(1 \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) - \mu_{10}\right) \\ &\quad -\frac{1}{\mu_{00}}\left(\frac{1-A}{g_0(0 \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(0,W) - \mu_{00})\right) \end{aligned}$$

$$\begin{aligned} IC^{logOR}(O) &= \frac{1}{\mu_{10}(1-\mu_{10})}\left(\frac{A}{g_0(1 \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W)\right) \\ &\quad -\frac{1}{\mu_{00}(1-\mu_{00})}\left(\frac{1-A}{g_0(0 \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(0,W)\right) \end{aligned}$$

Each of these influence curves can be estimated by substituting the appropriate estimates $g_{n,k}(A \mid W)$, $\bar{Q}_{n,k}(A,W)$. For the unadjusted and IPTW estimators with g_n estimated from

the data one obtains the wished influence curves by setting $\bar{Q}_0(A,W) = E_Y \mid A,W)$, and $\bar{Q}_0(1,W) = \bar{Q}_0(0,W) = 0$. For the Gcomp and AIPTW procedures these conditional means were estimated using a main terms logistic regression of Y on A and the covariates listed in Table 6.5. For the TMLE the conditional means were estimated with the targeted estimators $\bar{Q}^*_{n,k}(A,W)$, $\bar{Q}^*_{n,k}(1,W)$, $\bar{Q}^*_{n,k}(0,W)$. $g_{n,k}(1 \mid W)$ corresponds to the empirical proportion of subjects in the treatment group $(A = 1)$ for all estimators.

6.3.4 Results

For the Gcomp and AIPTW estimators $\bar{Q}_0$ was estimated with a main terms linear regression of Y_k on the covariates shown in Table 6.5. TMLE used a discrete SL to obtain an initial estimate of $\bar{Q}_0$; cross validation resulted in the selection of the unadjusted regression of Y on A to estimate the conditional means for all analyses.

Table 6.6 lists point estimates and 95% confidence intervals corresponding to the (0.025, 0.975) quantiles of 1000 bootstrap estimates are reported for all estimators. The super learner algorithm selected the unadjusted regression in each of our data analyses. Therefore, for the sake of reducing computer time, in this bootstrap we used the unadjusted regression as the initial estimator $\bar{Q}^0_n$. The bootstrap also does not include covariate selection by indication. Confidence intervals based on influence curve variance estimates are also reported. Table 6.7 lists these same point estimates and influence-curve based 95% confidence intervals.

Pooled estimates of the risk difference, risk ratio, and odds ratio were calculated as a weighted average of the indication-specific parameter estimates, with weights equal to the inverse of the estimated bootstrap variances. Odds ratio and risk ratio estimates and confidence intervals were constructed on the log scale.

$$\psi^{RD}_{n,pooled} = \frac{\sum_{k=1}^{K} w^{RD}_k \psi^{RD}_{n,k}}{\sum_{k=1}^{K} w^{RD}_k}, \qquad w^{RD}_k = \frac{1}{var(\psi^{RD}_{n,k})}$$

$$\psi^{RR}_{n,pooled} = exp\left(\frac{\sum_{k=1}^{K} w^{RR}_k log(\psi^{RR}_{n,k})}{\sum_{k=1}^{K} w^{RR}_k}\right), \quad w^{RR}_k = \frac{1}{var(log(\psi^{RR}_{n,k}))}$$

$$\psi^{OR}_{n,pooled} = exp\left(\frac{\sum_{k=1}^{K} w^{OR}_k log(\psi^{OR}_{n,k})}{\sum_{k=1}^{K} w^{OR}_k}\right), \quad w^{OR}_k = \frac{1}{var(log(\psi^{OR}_{n,k}))}$$

Pooled estimates of the risk difference, risk ratio, and odds ratio calculated for each estimation procedure are given in Table 6.8. Weights used to obtain the pooled estimates are given in Table 6.9. Pooled risk ratio and odds ratio results are not listed for the AIPTW estimator because some estimated risk ratios and odds ratios in the bootstrap estimates used to calculate the variance were negative for indication A, where there was only one event in the control arm. Though

influence curve estimated variances are available, this problem illustrates a known drawback of the estimating equation-based augmented IPTW methodology—predicted values do not respect known bounds on the estimation problem, so parameter estimates may fall outside the range of possible values.

Table 6.6: Risk difference, risk ratio, and odds ratio estimates. 95% CI corresponds to (0.025, 0.975) quantiles of 1000 bootstrap estimates.

	Risk Difference		**Risk Ratio**		**Odds Ratio**	
	est	95% CI	est	95% CI	est	95% CI
Indication A						
unadj	0.01	(-0.0003, 0.02)	4.90	(0.93, 774)	4.94	(0.93, 784)
IPTW	0.01	(-0.0003, 0.02)	4.90	(0.93, 778)	4.94	(0.93, 785)
MLE	0.01	(-0.0003, 0.02)	4.90	(0.93, 774)	4.94	(0.93, 784)
AIPTW	0.01	(-0.0003, 0.02)	4.90	(-1e+10, 1e+10)	4.94	(-1e+10, 1e+10)
TMLE	0.01	(-0.0003, 0.02)	4.90	(0.93, 774)	4.94	(0.93, 784)
Indication B						
unadj	0.01	(-0.003, 0.02)	1.63	(0.87, 3.85)	1.64	(0.87, 3.92)
IPTW	0.01	(-0.003, 0.02)	1.63	(0.87, 3.85)	1.64	(0.87, 3.92)
MLE	0.01	(-0.004, 0.02)	1.53	(0.80, 3.56)	1.55	(0.80, 3.63)
AIPTW	0.01	(-0.004, 0.02)	1.53	(0.80, 3.56)	1.55	(0.80, 3.63)
TMLE	0.01	(-0.003, 0.02)	1.63	(0.87, 3.85)	1.64	(0.87, 3.92)
Indication C						
unadj	0.00	(-0.01, 0.02)	1.24	(0.49, 3.43)	1.24	(0.48, 3.50)
IPTW	0.00	(-0.01, 0.02)	1.24	(0.49, 3.43)	1.24	(0.48, 3.50)
MLE	0.00	(-0.02, 0.02)	1.19	(0.47, 3.38)	1.19	(0.46, 3.45)
AIPTW	0.00	(-0.02, 0.02)	1.19	(0.47, 3.38)	1.19	(0.46, 3.45)
TMLE	0.00	(-0.01, 0.02)	1.24	(0.49, 3.43)	1.24	(0.48, 3.50)
Indication D						
unadj	0.01	(-0.03, 0.05)	1.05	(0.73, 1.55)	1.06	(0.70, 1.64)
IPTW	0.01	(-0.03, 0.05)	1.05	(0.73, 1.55)	1.06	(0.70, 1.64)
MLE	0.00	(-0.04, 0.05)	1.04	(0.73, 1.51)	1.05	(0.71, 1.59)
AIPTW	0.00	(-0.04, 0.05)	1.04	(0.73, 1.51)	1.05	(0.71, 1.59)
TMLE	0.01	(-0.03, 0.05)	1.05	(0.73, 1.55)	1.06	(0.70, 1.64)

Table 6.7: Risk difference, risk ratio, and odds ratio estimates, influence curve-based 95% CIs.

	Risk Difference		**Risk Ratio**		**Odds Ratio**	
	est	95% CI	est	95% CI	est	95% CI
Indication A						
unadj	0.01	(-0.001, 0.02)	4.90	(0.59, 40.59)	4.94	(0.59, 41.14)
IPTW	0.01	(-0.001, 0.02)	4.90	(0.57, 42.01)	4.94	(0.57, 42.58)
MLE	0.01	(-0.001, 0.02)	4.90	(0.59, 40.46)	4.94	(0.59, 41.00)
AIPTW	0.01	(-0.001, 0.02)	4.90	(0.57, 41.88)	4.94	(0.57, 42.44)
TMLE	0.01	(-0.001, 0.02)	4.90	(0.57, 41.88)	4.94	(0.57, 42.44)
Indication B						
unadj	0.01	(-0.004, 0.02)	1.63	(0.81, 3.25)	1.64	(0.81, 3.33)
IPTW	0.01	(-0.004, 0.02)	1.63	(0.81, 3.25)	1.64	(0.81, 3.32)
MLE	0.01	(-0.005, 0.02)	1.53	(0.78, 3.00)	1.55	(0.78, 3.07)
AIPTW	0.01	(-0.005, 0.02)	1.53	(0.79, 3.00)	1.55	(0.78, 3.06)
TMLE	0.01	(-0.004, 0.02)	1.63	(0.82, 3.22)	1.64	(0.82, 3.30)
Indication C						
unadj	0.00	(-0.01, 0.02)	1.24	(0.49, 3.11)	1.24	(0.48, 3.19)
IPTW	0.00	(-0.01, 0.02)	1.24	(0.49, 3.13)	1.24	(0.48, 3.21)
MLE	0.00	(-0.02, 0.02)	1.19	(0.48, 2.93)	1.19	(0.47, 3.00)
AIPTW	0.00	(-0.02, 0.02)	1.19	(0.48, 2.95)	1.19	(0.47, 3.02)
TMLE	0.00	(-0.01, 0.02)	1.24	(0.49, 3.10)	1.24	(0.48, 3.18)
Indication D						
unadj	0.01	(-0.04, 0.05)	1.05	(0.72, 1.54)	1.06	(0.69, 1.62)
IPTW	0.01	(-0.04, 0.05)	1.05	(0.72, 1.54)	1.06	(0.69, 1.63)
MLE	0.00	(-0.03, 0.04)	1.04	(0.74, 1.47)	1.05	(0.71, 1.55)
AIPTW	0.00	(-0.03, 0.04)	1.04	(0.74, 1.47)	1.05	(0.71, 1.55)
TMLE	0.01	(-0.03, 0.05)	1.05	(0.74, 1.51)	1.06	(0.71, 1.59)

Table 6.8: Pooled estimates and 95% confidence intervals for risk difference, risk ratio, and odds ratio parameters.

	Risk Difference		**Risk Ratio**		**Odds Ratio**	
	est	95% CI	est	95% CI	est	95% CI
unadj	0.01	(7e-04, 0.01)	1.16	(0.85, 1.59)	1.19	(0.84, 1.67)
IPTW	0.01	(7e-04, 0.01)	1.16	(0.85, 1.59)	1.19	(0.84, 1.67)
MLE	0.01	(3e-04, 0.01)	1.13	(0.83, 1.54)	1.16	(0.83, 1.62)
AIPTW	0.01	(3e-04, 0.01)				
TMLE	0.01	(7e-04, 0.01)	1.16	(0.85, 1.59)	1.19	(0.84, 1.67)

Table 6.9: Weights used for pooled estimates of risk difference, risk ratio, and odds ratio parameters by estimator.

	Risk Difference				**Risk Ratio**				**Odds Ratio**			
	A	B	C	D	A	B	C	D	A	B	C	D
unadj	0.62	0.24	0.12	0.03	< 0.01	0.19	0.1	0.72	< 0.01	0.22	0.11	0.67
IPTW	0.62	0.24	0.12	0.03	< 0.01	0.19	0.1	0.72	< 0.01	0.22	0.11	0.67
MLE	0.61	0.24	0.12	0.03	< 0.01	0.18	0.1	0.72	< 0.01	0.21	0.11	0.68
AIPTW	0.61	0.24	0.12	0.03								
TMLE	0.62	0.24	0.12	0.03	< 0.01	0.19	0.1	0.72	< 0.01	0.22	0.11	0.67

6.4 Simulation Studies

Estimation procedures that do not adequately account for missingness in the outcome can yield biased results when missingness is not independent of the outcome. In addition, p-values and confidence intervals around these biased estimates may lead to false acceptance or rejection of a null hypothesis of no treatment effect. A simulation study was designed to illustrate this phenomenon. Data are generated once, then three separate missingness mechanisms are applied to set approximately 26% of outcomes to missing: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) missing at random, with weak positivity violations, meaning that for a subset of observations there is sparsity in the data because the probability of observing the outcome is small (MAR.sp).

6.4.1 Estimators

In addition to the five estimators defined above, multiple imputation (MI) is also applied to estimate the risk difference under each of these missingness scenarios. The unadjusted estimator, IPTW, Gcomp, AIPTW, and TMLE estimators defined above are modified to accommodate missing data.

The unadjusted regression is simply carried out for subjects where the outcome is observed. The Gcomp estimate also ignores missingness, fitting a regression model on observations where the outcome is observed. The multiple imputation approach used was predictive mean matching as implemented in the *aregImpute* procedure in R. Missing values were imputed to create $m = 5$ complete datasets, that were then analyzed using Gcomp. The MI estimate is the mean of the m estimates from each complete dataset.

The model for $\bar{Q}_0$ used for MI, Gcomp and AIPTW estimators was set to a regression of Y on $A, AGE, ACIN, GRAMNEG, STAPH, BACTER$, and interaction term $A * AGE$, fitted using the observed data. For all estimators that rely on the g portion of the likelihood, this now factorizes into the treatment assignment and missingness mechanisms. The treatment assignment mechanism was set to the marginal probability of receiving the assigned treatment. Missingness was modeled as a regression of Δ on all main terms plus interactions with A for MI, IPTW, and AIPTW. TMLE fits for $\bar{Q}_0$ and $P(\Delta = 1 \mid A, W)$ were obtained data-adaptively using super learning, described further below.

The IPTW estimator is a weighted regression of Y on A using observations with $\Delta = 1$, with weights

$$wt = \frac{1}{\pi_n(\Delta \mid A, W) g_n(A \mid W)}.$$

Large weights are known to lead to unstable estimates, therefore the product of the probabilities in the denominator was bounded at $(0.01, 0.99)$.

Both the AIPTW and TMLE estimators of the risk difference parameter for the more general data structure $O = (W, A, \Delta, \Delta Y)$ are solutions of the efficient influence curve equation, and can

thereby be represented as

$$\begin{aligned}\psi_n^{RD} &= \frac{1}{n}\sum_i \frac{\Delta_i}{\pi_n(\Delta_i \mid A_i, W_i)}\left(\frac{I(A_i=1)}{g_n(A_i \mid W_i)} - \frac{I(A_i=0)}{g_n(A_i \mid W_i)}\right)(Y_i - \bar{Q}_n(A_i, W_i)) \\ &\quad + \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\end{aligned}$$

where the product of the probabilities $g(A \mid W)$ and $\pi(\Delta = 1 \mid A, W)$ was bounded at $(0.01, 0.99)$, and $\bar{Q}_0(A, W) = E_0(Y|A, W, \Delta = 1)$. In contrast with AIPTW, the TMLE is a substitution estimator $\psi_n^{RD} = \Psi(\bar{Q}_n^*) = 1/n \sum_i \bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)$, based on a targeted estimate $\bar{Q}_n^*$ of $\bar{Q}_0$.

Estimation of $\bar{Q}_0$ involves regressing outcome Y on treatment and covariates among the non-missing observations ($\Delta = 1$). Covariates used in the TMLE targeting step incorporate missingness. Coefficients ϵ_1 and ϵ_2 are fitted using only observations where the outcome is observed, with the given weights.

$$\begin{aligned}h_1(A, W) &= A \\ h_0(A, W) &= 1 - A \\ wt &= \frac{1}{\pi_n(\Delta \mid A, W) g_n(A \mid W)}.\end{aligned}$$

For TMLE the super learner was used to estimate $\bar{Q}_0$ and $P(\Delta = 1 \mid A, W)$. Prediction algorithms include *DSA, glm, knn*. Settings for *DSA* specified a search over a model space that includes polynomials of degree three. Linear regression models that included main terms only, all second order interactions, and the regression model used for the MI, Gcomp and AIPTW estimators were included in the super learner library. The k-nearest neighbors algorithm was run with neighborhoods of size 10, 20 and 40.

6.4.2 Data generation

Baseline covariates and treatment data were taken from n=934 observations in Study 7. The dataset consists of five baseline covariates, $ACIN, AGE, BACTER, GRAMNEG, STAPH$, treatment indicator $A = 1$ if the subject was treated with the study drug, 0 otherwise. A simulated outcome Y was generated as

$$P(Y = 1|A, W) = expit(\beta_0 + \beta_1 A + \beta_2 GRAMNEG + \beta_3 A * AGE + \beta_4 STAPH * GRAMNEG).$$

Coefficient values $\beta_0 = -4, \beta_1 = 1.4, \beta_2 = 1, \beta_3 = -0.05, \beta_4 = 3$ and $\beta_5 = -0.05$ give a marginal event proportion = 0.06. The true value of the risk difference parameter, $\psi_0 = -0.0429$ (-4.29%).

Three missingness mechanisms were defined such that marginally 26% of observations were

set to missing.

$$\begin{aligned} P(\Delta_{MCAR} = 1 \mid A, W) &= 0.74 \\ P(\Delta_{MAR} = 1 \mid A, W) &= expit(f(A, W)) \\ P(\Delta_{MAR.sp} = 1 \mid A, W) &= expit(-1.5 + 4 * f(A, W)) \\ f(A, W) &= 2 - 0.03A * AGE + 0.1GRAMNEG \\ &\quad -0.03A * GRAMNEG * BACTER + ACIN. \end{aligned}$$

6.4.3 Results

Estimates and 95% confidence intervals for each simulation are shown in Figure 6.1. Table 6.11 contains the estimates, bootstrapped standard errors (B=1000 bootstrap samples), and 95% confidence intervals constructed from the bootstrap SEs. When missingness is non-informative (MCAR) all estimators are unbiased, however the unadjusted, IPTW, and MI estimators are inefficient, and fail to establish statistical significance. When missingness is informative (MAR), the unadjusted, MI, and Gcomp estimates are more biased, but the IPTW, AIPTW, and TMLE estimation procedures are able to exploit covariate information to produce unbiased estimates. The IPTW again has a higher variance than either double-robust estimator, and TMLE's data-adaptive targeting provides greater bias reduction than AIPTW. When missingness is informative and there is sparsity in the data(MAR.sp) the unadjusted estimator fails completely. Multiple imputation performance suffers because there is no longer sufficient information in the data to produce imputed values that capture the true correlations. Gcomp fails to produce a significant result. IPTW, AIPTW, and TMLE estimates are all statistically significant. Again IPTW is least efficient, and AIPTW is slightly more biased and variable than TMLE.

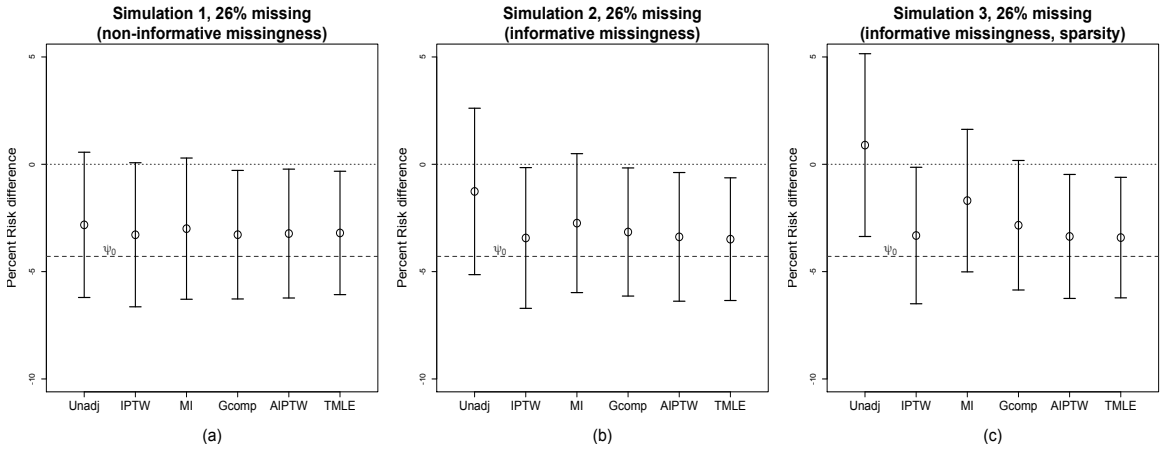

Figure 6.1: Percent risk difference estimates and 95% confidence intervals when data are missing completely at random (a), missing at random (b), and missing at random, with sparsity (c). Dashed lines are at the NULL (0) and the true parameter value (-4.29%).

Table 6.10: Summary of estimator performance on simulated data.

	95% CI Contains ψ_0			**Statistically Significant**			**Most Efficient**		
	MCAR	MAR	MAR.sp	MCAR	MAR	MAR.sp	MCAR	MAR	MAR.sp
Unadjusted	✓	✓							
IPTW	✓	✓	✓		✓	✓			
MI	✓	✓	✓						
Gcomp	✓	✓	✓	✓	✓				
AIPTW	✓	✓	✓	✓	✓	✓			
TMLE	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 6.11: Simulation study results, $\psi_0 = -4.29\%$.

			95% CI		
	est	SD	lb	ub	p-value
MCAR					
Unadj	-2.82	1.73	-6.21	0.56	0.102
IPTW	-3.28	1.71	-6.64	0.07	0.055
MI	-3.00	1.68	-6.29	0.29	0.074
Gcomp	-3.28	1.53	-6.27	-0.28	0.032
AIPTW	-3.23	1.53	-6.23	-0.22	0.035
TMLE	-3.20	1.47	-6.07	-0.32	0.029
MAR					
Unadj	-1.26	1.98	-5.14	2.61	0.523
IPTW	-3.43	1.67	-6.71	-0.15	0.040
MI	-2.74	1.65	-5.98	0.50	0.097
Gcomp	-3.15	1.52	-6.14	-0.17	0.038
AIPTW	-3.38	1.53	-6.38	-0.38	0.027
TMLE	-3.49	1.46	-6.35	-0.63	0.017
MAR.sp					
Unadj	0.89	2.17	-3.36	5.15	0.681
IPTW	-3.32	1.62	-6.50	-0.14	0.041
MI	-1.69	1.69	-5.01	1.63	0.318
Gcomp	-2.84	1.54	-5.86	0.18	0.065
AIPTW	-3.36	1.47	-6.25	-0.47	0.023
TMLE	-3.41	1.43	-6.22	-0.61	0.017

Results from 1000 bootstrap samples are shown in Table 6.12, including the average estimate, standard error, bias, mean squared error (MSE), the (0.025, 0.975) quantiles of the bootstrap estimates.

Table 6.12: Percent Risk Difference estimates: results from 1000 bootstrap samples.

					Quantiles	
	Ave	SD	Bias	MSE	0.025	0.975
MCAR						
Unadj	−2.91	1.73	1.39	4.90	−6.37	0.45
IPTW	−3.35	1.71	0.94	3.82	−6.79	0.06
MI	−3.10	1.68	1.19	4.22	−6.61	0.09
Gcomp	−3.39	1.53	0.90	3.15	−6.54	−0.32
AIPTW	−3.33	1.53	0.96	3.27	−6.27	−0.22
TMLE	−3.48	1.47	0.81	2.81	−6.40	−0.69
MAR						
Unadj	−1.36	1.98	2.93	12.51	−5.34	2.41
IPTW	−3.48	1.67	0.81	3.45	−6.89	−0.26
MI	−2.91	1.65	1.38	4.63	−6.29	0.26
Gcomp	−3.22	1.52	1.07	3.46	−6.26	−0.31
AIPTW	−3.44	1.53	0.85	3.07	−6.52	−0.52
TMLE	−3.69	1.46	0.60	2.48	−6.61	−0.95
MAR.sp						
Unadj	0.76	2.17	5.05	30.24	−3.57	4.94
IPTW	−3.33	1.62	0.96	3.55	−6.74	−0.33
MI	−1.85	1.69	2.44	8.81	−5.48	1.33
Gcomp	−2.90	1.54	1.40	4.31	−5.92	0.08
AIPTW	−3.40	1.47	0.89	2.96	−6.31	−0.62
TMLE	−3.58	1.43	0.72	2.56	−6.57	−0.91

6.5 Discussion

There is little opportunity to increase R-square when outcomes are rare. Data-adaptive estimation of $\bar{Q}_0$ does not improve the coefficient of determination (R-square) in comparison with the unadjusted regression of Y on A for any indications. Theory dictates that no gain in efficiency can be expected if there is no change in R-square. As illustrated by the simulation study, the motivation for using TMLE in these situations is therefore not to increase efficiency, but rather to correct for

possible bias introduced through failed randomization, informative censoring or informative missingness in the outcome. Missingness coupled with sparsity is not adequately addressed through multiple imputation without relying on unverifiable modeling assumptions. IPTW estimation can reduce bias, but is inefficient, and relies entirely on correct specification of unknown missingness and treatment models. The double robustness of augmented-IPTW is an improvement, but as the analysis of the original RCT data illustrates, AIPTW estimates can fall outside the known range of valid parameter values. TMLE estimation respect these global constraints, is asymptotically efficient and unbiased, and has good finite sample performance. Confidence intervals presented in Tables 6.11 and 6.12 demonstrate that influence curve-based inference is comparable with inference based on the bootstrap, without requiring additional computational resources. Meta-analysis and simulation results demonstrate that TMLE is a worthwhile alternative to current approaches for safety analysis.

Table 6.13: Covariate Descriptions.

Covariate	Description
ACIN	presence or absence of *Acinetobacter sp.*
AGE	age in years at baseline
AGEPTS	APACHE points derived from age
ALBUMIN	baseline albumin laboratory value
ALKPHOS	baseline alkaline phosphotase laboratory value
ALT	baseline alanine aminotransferase laboratory value
APACHE	APACHE II score
ARTPH	APACHE points derived from arterial pH or serum HCO3
ARTPRESS	APACHE points derived from arterial pressure
AST	baseline aspartate aminotransferase laboratory value
BACTER	presence or absence of bacteremia
BMI	body mass index
CHFHIST	history of congestive heart failure (yes/no)
COPDHIST	history of COPD (yes/no)
CREATIN	baseline Creatinine laboratory value
GRAMNEG	presence or absence of gram-negative bacteria
HEARTRAT	APACHE points derived from heart rate
HEMAT	APACHE points derived from hematocrit
HEMOGLOB	baseline hemoglobin laboratory value
HRV	Croatia, country of origin
HUN	Hungary, country of origin
IND	India, country of origin
OXYG	APACHE points derived from oxygenation
PER	Peru, country of origin
PLATELET	baseline platelet laboratory value
POTASSIU	baseline potassium laboratory value
PROTTOT	baseline total protein laboratory value
RESPRATE	APACHE points derived from respiratory reate
SERCREA	APACHE points derived from serum creatinine
SERPOT	APACHE points derived from serum potassium
SERSOD	APACHE points derived from serum sodium
STAPH	presence or absence of *Staphylococcus aureus*
SVK	Slovakia, country of origin
TEMP	APACHE points derived from temperature
WBC	baseline white blood cell laboratory value
WBCPTS	APACHE points derived from white blood cell count

Bibliography

A. Abadie and G.W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–67, 2006.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–72, 2005.

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 Biostatistics, University of California, Berkeley, 2008. URL `www.bepress.com/ucbbiostat/paper230/`.

O. Bembom, J.W. Fessel, R.W. Shafer, and M.J. van der Laan. Data-adaptive selection of the adjustment set in variable importance estimation. Technical report, DigitalCommons@Florida Atlantic University [http://digitalcommons.fau.edu/cgi/oai2.cgi] (United States), 2008. URL `http://www.bepress.com/ucbbiostat/paper231`.

O. Bembom, M.L. Petersen, S.-Y. Rhee, W. J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection. *Statistics in Medicine*, 28:152–72, 2009.

J. Bhattacharya and W Vogt. Do instrumental variables belong in propensity scores? NBER Technical Working Paper 343, National Bureau of Economic Research, MA., 2007.

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

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.

Hugh Chipman and Robert McCulloch. *BayesTree: Bayesian Methods for Tree Based Models*, 2010. URL http://CRAN.R-project.org/package=BayesTree. R package version 0.3-1.1.

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.

R.H. Dehejia and S. Wahba. Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84:151–61, 2002.

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*, 32:392–409, 2008.

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.

Jerome 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.

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

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: Characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–94, New York, 1997. Springer-Verlag.

S. Gruber. *Collaborative Targeted Maximum Likelihood Estimation*, 2010a. URL http://www.stat.berkeley.edu/~laan/Software/. R software version 0.5.

S. Gruber. *tmle: Targeted Maximum Likelihood Estimation*, 2010b. URL http://CRAN.R-project.org/package=tmle. R package version 1.1.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 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), 2010b.

F.R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–93, 1974.

F.E. Harrell Jr. *Hmisc: Harrell Miscellaneous*, 2010. URL `http://CRAN.R-project.org/package=Hmisc`. R package version 3.8-3.

T.J. Hastie. *gam: Generalized Additive Models*, 2009. URL `http://CRAN.R-project.org/package=gam`. R package version 1.01.

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.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, December 1991.

M. A. Hernan, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000a.

M. A. Hernan, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000b.

J.H. Holland and J.S. Reitman. Cognitive systems based on adaptive algorithms. *SIGART Bull.*, 63:49–49, 1977. ISSN 0163-5719. doi: http://doi.acm.org/10.1145/1045343.1045373.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *The Annals of Statistics*, 23:774–86, 1995.

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.

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

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.

M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, and A. Engelhardt. *caret: Classification and Regression Training*, 2010. URL `http://CRAN.R-project.org/package=caret`. R package version 4.72.

P. McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11:59–67, 1983.

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.

J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. *Proceedings of UAI*, 2010. forthcoming.

J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), 2010a.

J. Pearl. The causal foundations of structural equation modeling. Technical Report R-370, University of California, Los Angeles, Department of Computer Science, August 2010b.

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. van der Laan. Diagnosing and responding to violations in the positivity assumption. Technical report, Division of Biostatistics, Universtiy of California, Berkeley, 2010. URL `www.bepress.com/ucbbiostat/paper269/`.

E.C. Polley. *SuperLearner: Super Learner Prediction*, 2010. URL `http://www.stat.berkeley.edu/~ecpolley/SL/`. R package version 1.1-17.

G. Ridgeway and D.F. McCaffrey. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:4:540–443, 2007.

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

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica*, 11(4):920–936, 2001.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

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. 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. 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, 2000a.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials (Minneapolis, MN, 1997)*, pages 95–133. Springer-Verlag, New York, 2000b.

J.M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87:113–124, 2000.

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.

P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

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.

B. Rosner. Fev dataset, 1999a. URL `http://www.amstat.org/publications/jse/datasets`. Submitted by M.J. Kahn, Wheaton College, Norton, MA.

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

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 64:688–701, 1974.

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(448):1096–1120 (1121–1146), 1999.

J.S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Sotware, Forthcoming*, 2008.

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

O.M. Stitelman and M.J. van der Laan. Collaborative targeted maximum likelihood for time to event data. Technical Report 260, Division of Biostatistics, University of California, Berkeley, http://www.bepress.com/ucbbiostat/paper260, 2010.

T.A. Stukel, E.S. Fisher, D.E. Wennberg, D.A.Alter, D.J.Gottlieb, and M.J. Vermeulen. Analysis of observational studies in the presence of treatment selection bias: Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*, 297:278–85, 2007.

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.

Z. Tan. Bounded, efficient, and doubly robust estimation with inverse weighting. *Biometrika*, 94: 1–22, 2008.

Z. 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. URL `http://www.R-project.org`.

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 S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1), January 2010.

M.J. van der Laan and M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3, Issue 1, Article 3, 2007.

M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York, 2003.

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

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

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. Technical report 213, Division of Biostatistics, University of California, Berkeley, 2006b.

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 in targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 254*, 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-Verlag, New York, 4th edition, 2002a.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4th edition, 2002b. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

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.

J. Wooldridge. Should instrumental variables be used as matching variables? Tech. Rep. Michigan State University, MI., 2009.

Appendix A

Influence Curve Equations

A.1 Using the Delta Method to Derive Influence Curve Equations for log(OR) and log(RR) Parameters

Let $D^*(\mu)$ be the efficient influence curve for some parameter, μ . The delta method allows us to derive the efficient influence curve equation for a function $f(\mu)$, provided the first derivative of the function with respect to μ, $f'(\mu)$, exists and is non-zero.

$$D^*(f(\mu)) = f'(\mu)D^*(\mu)$$

The delta method can be used to construct the efficient influence curves for the log relative risk $(logRR)$ and log odds ratio$(logOR)$ parameters from the efficient influence curves for $\mu_0 = E(Y_0)$ and $\mu_1 = E(Y_1)$. When treatment assignment is binary, these are given by,

$$\begin{aligned} D^*(\mu_1)(O) &= \frac{A}{g_0(1 \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(1,W) - \mu_1 \\ D^*(\mu_0)(O) &= \frac{1-A}{1 - g_0(1 \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(0,W) - \mu_0 \end{aligned}$$

where O is the unit of observation, A is a binary treatment indicator, W is a vector of baseline covariates, $g(1 \mid W) = P(A = 1 \mid W)$, Y is an outcome measure, and $\bar{Q}(A,W)$ is the true conditional expectation of Y given A and W.

Log relative risk parameter, $\psi^{\mathbf{logRR}}$

$$\begin{aligned} \psi^{logRR} &= log\left(\frac{\mu_1}{\mu_0}\right) \\ &= log(\mu_1) - log(\mu_0). \end{aligned}$$

The efficient influence curve for this parameter can be written as

$$D^*(\psi^{logRR})(O) = \frac{\partial}{\partial \mu_1}(\psi^{logRR})D^*(\mu_1)(O) + \frac{\partial}{\partial \mu_0}(\psi^{logRR})D^*(\mu_0)(O)$$

where

$$\begin{aligned}\frac{\partial}{\partial \mu_1}(\psi^{logRR}) &= \frac{1}{\mu_1} \\ \frac{\partial}{\partial \mu_0}(\psi^{logRR}) &= -\frac{1}{\mu_0}\end{aligned}$$

By substitution,

$$\begin{aligned}D^*(\psi^{logRR})(O) &= \frac{1}{\mu_1}\left(\frac{A}{g(1 \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(1,W) - \mu_1\right) \\ &\quad -\frac{1}{\mu_0}\left(\frac{1-A}{1-g(1 \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(0,W) - \mu_0)\right).\end{aligned}$$

Log odds ratio parameter, $\psi^{\mathbf{logOR}}$

$$\begin{aligned}\psi^{log(OR)} &= log\left(\frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}\right) \\ &= log(\mu_1) - log(1-\mu_1) - log(\mu_0) + log(1-\mu_0)\end{aligned}$$

The efficient influence curve for this parameter can be written as

$$D^*(\psi^{logOR})(O) = \frac{\partial}{\partial \mu_1}(\psi^{logOR})D^*(\mu_1)(O) + \frac{\partial}{\partial \mu_0}(\psi^{logOR})D^*(\mu_0)(O)$$

where

$$\begin{aligned}\frac{\partial}{\partial \mu_1}(\psi^{logOR}) &= \frac{1}{\mu_1(1-\mu_1)} \\ \frac{\partial}{\partial \mu_0}(\psi^{logOR}) &= -\frac{1}{\mu_0(1-\mu_0)}\end{aligned}$$

By substitution,

$$\begin{aligned}D^*(\psi^{logOR})(O) &= \frac{1}{\mu_1(1-\mu_1)}\left(\frac{A}{g(1 \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(1,W)\right) \\ &\quad -\frac{1}{\mu_0(1-\mu_0)}\left(\frac{1-A}{1-g(1 \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(0,W)\right).\end{aligned}$$

A.2 Influence Curve Contribution Due to Estimating G for Point Treatment Effects

An estimator α_n of a parameter α that is asymptotically linear has the property that $\alpha_n - \alpha = P_n IC_\alpha + r(n)$, where $r(n)$ is a second order term. Consider g_α, the conditional distribution of binary treatment assignment A given baseline covariates, W, parameterized by α. The asymptotic linearity of α_n allows us to approximate $Q(g_{\alpha_n}) - Q(g_\alpha)$ as $P_0 D^*_{g_{\alpha_n}}(O) - D^*_{g_\alpha}(O)$. We can conclude from this that $\Phi(g_n)$ as an estimator of Φ_g is given by

$$[P_0 f(P_0)] \cdot IC_\alpha$$

where $[P_0 f(P_0)]$ is a parameter-specific matrix, and the vector influence curve IC_α is given by

$$IC_\alpha(O) = P_0 \left[W \frac{d}{d\alpha} g_\alpha(1 \mid W) \right]^{-1} \overrightarrow{W}(A - g_\alpha(1 \mid W)).$$

When $g(A \mid W)$ follows a logistic distribution,

$$\begin{aligned}
g_\alpha(1, W) &= (1 + e^{-\alpha_n W})^{-1} \\
g_\alpha(0, W) &= e^{-\alpha_n W}(1 + e^{-\alpha_n W})^{-1} \\
\frac{d}{d\alpha} g_\alpha(1, W) &= (1 + e^{-\alpha_n W})^{-2} e^{-\alpha_n W} W \\
&= g_\alpha(1, W) g_\alpha(0, W) W \\
\frac{d}{d\alpha} g_\alpha(0, W) &= \frac{d}{d\alpha}(1 - g_\alpha(1, W)) \\
&= -\frac{d}{d\alpha} g_\alpha(1, W)
\end{aligned}$$

By substitution,

$$IC_\alpha(O) = P_0 \left[W W^T g_\alpha(1 \mid W) g_\alpha(0 \mid W) \right]^{-1} (A - g_\alpha(1 \mid W)) \overrightarrow{W}.$$

This influence curve is unknown, and must be estimated from data.

$$\widehat{IC}_{\alpha_n}(O_i) = \left[\frac{1}{n} \sum_{i=1}^{n} \overrightarrow{W_i} \overrightarrow{W_i}^T g_{\alpha_n}(1 \mid W_i) g_{\alpha_n}(0 \mid W_i) \right]^{-1} (A_i - g_{\alpha_n}(1 \mid W_i)) \overrightarrow{W_i} \qquad \text{(A.1)}$$

For point treatment parameters the general form of the contribution to the influence curve from the estimation of g is given by

$$IC_g(O) = -\mathbf{a} \cdot (\mathbf{B}^{-1}(A - g_{\alpha_n}(1 \mid W) \overrightarrow{W}))$$

and estimated as

$$\widehat{IC}_g(O_i) = -\hat{\mathbf{a}} \cdot (\hat{\mathbf{B}}^{-1}(A_i - g_{\alpha_n}(1 \mid W_i)\overrightarrow{W_i}))$$

where $\hat{\mathbf{B}}$ is given by

$$\hat{\mathbf{B}} = \left[\frac{1}{n}\sum_{i=1}^{n} \overrightarrow{W_i}\overrightarrow{W_i}^T g_{\alpha_n}(1 \mid W_i) g_{\alpha_n}(0 \mid W_i)\right],$$

and $\hat{\mathbf{a}}$ is derived below for several target parameters: the additive treatment effect, population mean under missingness, relative risk, log relative risk, odds ratio, and log odds ratio.

A.2.1 Additive treatment effect

$$\psi_0^{ATE} = E(Y_1) - E(Y_0)$$

The efficient influence curve for this parameter is given by

$$D^*(O) = \left(\frac{A}{g(1)} - \frac{(1-A)}{g(0)}\right)(Y - Q) + Q(1, W) - Q(0, W) - \psi(Q).$$

$$\begin{aligned}
Q(g_n) - Q(g) &\approx P_0 D^*_{g_n} - D^*_g \\
&= P_0\left[\left(\frac{A}{g_{\alpha_n}(1)} - \frac{(1-A)}{g_{\alpha_n}(0)}\right)(Y - Q) - \left(\left(\frac{A}{g_\alpha(1)} - \frac{(1-A)}{g_\alpha(0)}\right)(Y - Q)\right)\right] \\
&= P_0(Y - Q)\left[A\left(\frac{1}{g_{\alpha_n}(1)} - \frac{1}{g_\alpha(1)}\right) - (1-A)\left(\frac{1}{g_{\alpha_n}(0)} - \frac{1}{g_\alpha(0)}\right)\right] \\
&\leq P_0(Y - Q)\left[A\frac{g_\alpha(1) - g_{\alpha_n}(1)}{g^2_\alpha(1)} - (1-A)\frac{g_\alpha(0) - g_{\alpha_n}(0)}{g^2_\alpha(0)}\right] \\
&= P_0(Y - Q)\left[A\frac{-\frac{d}{d\alpha}g_\alpha(1)(\alpha_n - \alpha)}{g^2_\alpha(1)} + (1-A)\frac{\frac{d}{d\alpha}g_\alpha(0)(\alpha_n - \alpha)}{g^2_\alpha(0)}\right] \\
&= P_0\left[(Y - Q)A\frac{-\frac{d}{d\alpha}g_\alpha(1)}{g^2_\alpha(1)} - (1-A)\frac{\frac{d}{d\alpha}g_\alpha(1)}{g^2_\alpha(0)}\right](\alpha_n - \alpha) \\
&= -P_0\left[(Y - Q)\frac{d}{d\alpha}g_\alpha(1)\left(\frac{A}{g^2_\alpha(1)} + \frac{1-A}{g^2_\alpha(0)}\right)\right](\alpha_n - \alpha) \\
&= -P_0\left[(Y - Q)\frac{d}{d\alpha}g_\alpha(1)\left(\frac{A}{g^2_\alpha(1)} + \frac{1-A}{g^2_\alpha(0)}\right)\right]\frac{1}{n}\sum_{i=1}^{n} IC_\alpha(O_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} -P_0(Y - Q)\overrightarrow{W}\left(A\frac{g_\alpha(0)}{g_\alpha(1)} + (1-A)\frac{g_\alpha(1)}{g_\alpha(0)}\right) \cdot IC_\alpha
\end{aligned}$$

This unknown influence curve can be estimated from data,

$$\begin{aligned}\widehat{IC}_g(O_i) &= \hat{\mathbf{a}}_{\mathbf{ATE}} \cdot \widehat{IC}_{\alpha_n}(O_i), \\ \hat{\mathbf{a}}_{\mathbf{ATE}} &= \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Q}(W_i, A_i))\overrightarrow{W_i}\left(A_i\frac{g_{\alpha_n}(0 \mid W_i)}{g_{\alpha_n}(1 \mid W_i)} + (1-A_i)\frac{g_{\alpha_n}(1 \mid W_i)}{g_{\alpha_n}(0 \mid W_i)}\right)\right].\end{aligned}$$

$\widehat{IC}_{\alpha_n}(O_i)$ is defined in (A.1).

Population Mean under missingness

$$\psi_0^{EY1} = E(Y_1)$$

The estimate of $\hat{\mathbf{a}}_{\mathbf{EY1}}$ for for the population mean follows trivially from this result,

$$\hat{\mathbf{a}}_{\mathbf{EY1}} = \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Q}(W_i, A_i))\overrightarrow{W_i}A_i\frac{g_{\alpha_n}(0 \mid W_i)}{g_{\alpha_n}(1 \mid W_i)}\right].$$

A.2.2 Relative risk

$$\psi_0^{RR} = E(Y_1)/E(Y_0)$$

Let $\mu_1 = E(Y_1)$, $\mu_0 = E(Y_0)$. The efficient influence curve for the relative risk is given by

$$\begin{aligned}D^*(O) &= \frac{1}{\mu_0}\left(\frac{A}{g(1,W)}(Y - Q(A,W)) + Q(1,W)\right) \\ &\quad -\frac{\mu_1}{\mu_0^2}\left(\frac{1-A}{1-g(1,W)}(Y - Q(A,W)) + Q(0,W)\right)\end{aligned}$$

$$\begin{aligned}Q(g_n) - Q(g) &\approx P_0\left[\frac{A}{\mu_0}(Y-Q)\left(\frac{1}{g_{\alpha_n}(1)} - \frac{1}{g_\alpha(1)}\right) - \frac{\mu_1}{\mu_0^2}(1-A)(Y-Q)\left(\frac{1}{g_{\alpha_n}(0)} - \frac{1}{g_\alpha(0)}\right)\right] \\ &\leq P_0\left[\frac{A}{\mu_0}(Y-Q)\left(\frac{g_\alpha(1) - g_{\alpha_n}(1)}{g_\alpha^2(1)}\right) - \frac{\mu_1}{\mu_0^2}(1-A)(Y-Q)\left(\frac{g_\alpha(0) - g_{\alpha_n}(0)}{g_\alpha^2(0)}\right)\right] \\ &= P_0(Y-Q)\left[\frac{A}{\mu_0}\left(\frac{g_\alpha(1) - g_{\alpha_n}(1)}{g_\alpha^2(1)}\right) - \frac{\mu_1}{\mu_0^2}(1-A)\left(\frac{g_\alpha(0) - g_{\alpha_n}(0)}{g_\alpha^2(0)}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= P_0(Y-Q)\left[\frac{A}{\mu_0}\left(\frac{\frac{-d}{d\alpha}g_\alpha(1)(\alpha_n-\alpha)}{g_\alpha^2(1)}\right)-\frac{\mu_1}{\mu_0^2}(1-A)\left(\frac{\frac{-d}{d\alpha}g_\alpha(0)(\alpha_n-\alpha)}{g_\alpha^2(0)}\right)\right]\\
&= P_0(Y-Q)\left[\frac{A}{\mu_0}\left(\frac{\frac{-d}{d\alpha}g_\alpha(1)}{g_\alpha^2(1)}\right)-\frac{\mu_1}{\mu_0^2}(1-A)\left(\frac{\frac{-d}{d\alpha}g_\alpha(0)}{g_\alpha^2(0)}\right)\right](\alpha_n-\alpha)\\
&= -P_0(Y-Q)\left[\frac{A}{\mu_0}\left(\frac{\frac{d}{d\alpha}g_\alpha(1)}{g_\alpha^2(1)}\right)+\frac{\mu_1}{\mu_0^2}(1-A)\left(\frac{\frac{d}{d\alpha}g_\alpha(1)}{g_\alpha^2(0)}\right)\right](\alpha_n-\alpha)\\
&= -P_0(Y-Q)\frac{d}{d\alpha}g_\alpha(1)\left[\frac{1}{\mu_0}\left(\frac{A}{g_\alpha^2(1)}\right)+\frac{\mu_1}{\mu_0^2}\left(\frac{1-A}{g_\alpha^2(0)}\right)\right](\alpha_n-\alpha)
\end{aligned}
$$

By substitution

$$
\begin{aligned}
Q(g_n)-Q(g) &\approx \frac{1}{n}\sum_{i=1}^{n}-P_0(Y-Q)\overrightarrow{W}\left[\frac{Ag_\alpha(0)}{\mu_0 g_\alpha(1)}+\frac{(1-A)\mu_1}{\mu_0^2}\frac{g_\alpha(1)}{g_\alpha(0)}\right]\cdot IC_\alpha\\
&= \frac{1}{n}\sum_{i=1}^{n}-P_0(Y-Q)\overrightarrow{W}\left[\frac{Ag_\alpha(0)}{\mu_0 g_\alpha(1)}+\frac{(1-A)\mu_1}{\mu_0^2}\frac{g_\alpha(1)}{g_\alpha(0)}\right]\cdot IC_\alpha
\end{aligned}
$$

Estimated as

$$
\begin{aligned}
\widehat{IC}_g(O_i) &= \hat{\mathbf{a}}_{\mathbf{RR}}\cdot\widehat{IC}_{\alpha_n}(O_i),\\
\hat{\mathbf{a}}_{\mathbf{RR}} &= -\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Q}(W_i,A_i))\overrightarrow{W_i}\left(A_i\frac{g_{\alpha_n}(0\mid W_i)}{\hat{\mu}_0 g_{\alpha_n}(1\mid W_i)}+(1-A_i)\frac{\hat{\mu}_1}{\hat{\mu}_0^2}\frac{g_{\alpha_n}(1\mid W_i)}{g_{\alpha_n}(0\mid W_i)}\right).
\end{aligned}
$$

$\widehat{IC}_{\alpha_n}(O_i)$ is defined in (A.1).

A.2.3 Log relative risk

$$\psi_0^{logRR} = log(E(Y_1)/E(Y_0)))$$

Let $\mu_1 = E(Y_1)$, $\mu_0 = E(Y_0)$. The efficient influence curve for the log relative risk is given by

$$
\begin{aligned}
D^*(O) &= \frac{1}{\mu_1}\left(\frac{A}{g(1,W)}(Y-Q(A,W))+Q(1,W)-\mu_1\right)\\
&\quad -\frac{1}{\mu_0}\left(\frac{1-A}{1-g(1,W)}(Y-Q(A,W))+Q(0,W)-\mu_0\right)
\end{aligned}
$$

$$\begin{aligned}
Q(g_n) - Q(g) &\approx P_0\left[\frac{A}{\mu_1}(Y-Q)\left(\frac{1}{g_{\alpha_n}(1)} - \frac{1}{g_\alpha(1)}\right) - \frac{1-A}{\mu_0}(Y-Q)\left(\frac{1}{g_{\alpha_n}(0)} - \frac{1}{g_\alpha(0)}\right)\right] \\
&\leq P_0\left[\frac{A}{\mu_1}(Y-Q)\left(\frac{g_\alpha(1) - g_{\alpha_n}(1)}{g_\alpha^2(1)}\right) - \frac{1-A}{\mu_0}(Y-Q)\left(\frac{g_\alpha(0) - g_{\alpha_n}(0)}{g_\alpha^2(0)}\right)\right] \\
&= P_0(Y-Q)\left[\frac{A}{\mu_1}\left(\frac{\frac{-d}{d\alpha}g_\alpha(1)(\alpha_n - \alpha)}{g_\alpha^2(1)}\right) - \frac{1-A}{\mu_0}\left(\frac{\frac{-d}{d\alpha}g_\alpha(0)(\alpha_n - \alpha)}{g_\alpha^2(0)}\right)\right] \\
&= -P_0(Y-Q)\left[\frac{A}{\mu_1}\left(\frac{\frac{d}{d\alpha}g_\alpha(1)}{g_\alpha^2(1)}\right) + \frac{1-A}{\mu_0}\left(\frac{\frac{d}{d\alpha}g_\alpha(1)}{g_\alpha^2(0)}\right)\right](\alpha_n - \alpha) \\
&= -P_0(Y-Q)\frac{d}{d\alpha}g_\alpha(1)\left[\frac{1}{\mu_1}\left(\frac{A}{g_\alpha^2(1)}\right) + \frac{1}{\mu_0}\left(\frac{1-A}{g_\alpha^2(0)}\right)\right](\alpha_n - \alpha) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} -P_0(Y-Q)\overrightarrow{W}\left[\frac{Ag_\alpha(0)}{\mu_1 g_\alpha(1)} + \frac{(1-A)\mu_1}{\mu_0}\frac{g_\alpha(1)}{g_\alpha(0)}\right]\cdot IC_\alpha
\end{aligned}$$

This is estimated as

$$\begin{aligned}
\widehat{IC}_g(O_i) &= \hat{\mathbf{a}}_{\mathbf{logRR}} \cdot \widehat{IC}_{\alpha_n}(O_i), \\
\hat{\mathbf{a}}_{\mathbf{logRR}} &= -\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Q}(W_i, A_i))\overrightarrow{W_i}\left(A_i\frac{g_{\alpha_n}(0 \mid W_i)}{\hat{\mu}_1 g_{\alpha_n}(1 \mid W_i)} + (1-A_i)\frac{\hat{\mu}_1}{\hat{\mu}_0}\frac{g_{\alpha_n}(1 \mid W_i)}{g_{\alpha_n}(0 \mid W_i)}\right)
\end{aligned}$$

$\widehat{IC}_{\alpha_n}(O_i)$ is defined in (A.1).

A.2.4 Odds ratio

$$\psi_0^{OR} = \frac{EY_1/(1-EY_1)}{EY_0/(1-EY_0)}$$

Let $\mu_1 = E(Y_1)$, $\mu_0 = E(Y_0)$. The efficient influence curve for the odds ratio is given by

$$\begin{aligned}
D^*(O) &= \frac{1-\mu_0}{\mu_0(1-\mu_1)^2}\left(\frac{A}{g(1)}(Y - Q(1,W) + Q(1,W) - \mu_1\right) \\
&\quad -\frac{\mu_1}{\mu_0^2(1-\mu_1)}\left(\frac{1-A}{g(0)}(Y - Q(0,W) + Q(0,W) - \mu_0\right)
\end{aligned}$$

$$\begin{aligned}
\text{Let } c_1 &= \frac{1-\mu_0}{\mu_0(1-\mu_1)^2}, \\
c_2 &= \frac{\mu_1}{\mu_0^2(1-\mu_1)}
\end{aligned}$$

$$\begin{aligned} Q(g_n) - Q(g) &\approx P_0(Y-Q)\left[Ac_1\left(\frac{1}{g_{\alpha_n}(1)} - \frac{1}{g_\alpha(1)}\right) - (1-A)c_2\left(\frac{1}{g_{\alpha_n}(0)} - \frac{1}{g_\alpha(0)}\right)\right] \\ &\leq P_0(Y-Q)\left[Ac_1\left(\frac{g_\alpha(1) - g_{\alpha_n}(1)}{g_\alpha(1)^2}\right) - (1-A)c_2\frac{g_\alpha(0) - g_{\alpha_n}(0)}{g_\alpha(0)^2}\right] \\ &= P_0(Y-Q)\left[Ac_1\frac{\frac{-d}{d\alpha}g_\alpha(1)}{g_\alpha(1)^2} - (1-A)c_2\frac{\frac{d}{d\alpha}g_\alpha(1)}{g_\alpha(0)^2}\right](\alpha_n - \alpha) \\ &\approx -P_0(Y-Q)\overrightarrow{W}\left[Ac_1\frac{g_\alpha(0)}{g_\alpha(1)} + (1-A)c_2\frac{g_\alpha(1)}{g\alpha(0)}\right]\cdot IC_\alpha \end{aligned}$$

Estimated as,

$$\begin{aligned} \widehat{IC}_g(O_i) &= \hat{\mathbf{a}}_{\mathbf{OR}} \cdot \widehat{IC}_{\alpha_n}, \\ \hat{\mathbf{a}}_{\mathbf{OR}} &= \frac{1}{n}\sum_{i=1} n(Y_i - \hat{Q}(W_i, A_i))\overrightarrow{W_i}\left[A_i\hat{c}_1\frac{g_{\alpha_n}(0,W_i)}{g_{\alpha_n}(1,W_i)} + (1-A_i)\hat{c}_2\frac{g_{\alpha_n}(1,W_i)}{g_{\alpha_n}(0,W_i)}\right]. \end{aligned}$$

$\widehat{IC}_{\alpha_n}(O_i)$ is defined in (A.1).

A.2.5 Log odds ratio

$$\psi_0^{logOR} = log\left(\frac{EY_1(1-EY_1)}{EY_0(1-EY_0)}\right)$$

Let $\mu_1 = E(Y_1)$, $\mu_0 = E(Y_0)$. The efficient influence curve for the log odds ratio is given by

$$\begin{aligned} D^*(O) &= \frac{1}{\mu_1(1-\mu_1)}\left(\frac{A}{g_0(1 \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W)\right) \\ &\quad -\frac{1}{\mu_0(1-\mu_0)}\left(\frac{1-A}{g_0(0 \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(0,W)\right) \end{aligned}$$

$$\begin{aligned} \text{Let } c_3 &= \frac{1}{\mu_1(1-\mu_1)}, \\ c_4 &= \frac{1}{\mu_0(1-\mu_0)} \end{aligned}$$

$$
\begin{aligned}
Q(g_n) - Q(g) &\approx P_0(Y-Q)\left[Ac_3\left(\frac{1}{g_{\alpha_n}(1)} - \frac{1}{g_\alpha(1)}\right) - (1-A)c_4\left(\frac{1}{g_{\alpha_n}(0)} - \frac{1}{g_\alpha(0)}\right)\right] \\
&\leq P_0(Y-Q)\left[Ac_3\left(\frac{g_\alpha(1) - g_{\alpha_n}(1)}{g_\alpha(1)^2}\right) - (1-A)c_4\frac{g_\alpha(0) - g_{\alpha_n}(0)}{g_\alpha(0)^2}\right] \\
&= P_0(Y-Q)\left[Ac_3\frac{\frac{-d}{d\alpha}g_\alpha(1)}{g_\alpha(1)^2} - (1-A)c_4\frac{\frac{d}{d\alpha}g_\alpha(1)}{g_\alpha(0)^2}\right](\alpha_n - \alpha) \\
&\approx -P_0(Y-Q)\overrightarrow{W}\left[Ac_3\frac{g_\alpha(0)}{g_\alpha(1)} + (1-A)c_4\frac{g_\alpha(1)}{g\alpha(0)}\right]\cdot IC_\alpha
\end{aligned}
$$

Estimated as,

$$
\begin{aligned}
\widehat{IC}_g(O_i) &= \hat{\mathbf{a}}_{\mathbf{logOR}} \cdot \widehat{IC}_{\alpha_n}, \\
\hat{\mathbf{a}}_{\mathbf{logOR}} &= \frac{1}{n}\sum_{i=1} n(Y_i - \hat{Q}(W_i, A_i))\overrightarrow{W_i}\left[A_i\hat{c}_3\frac{g_{\alpha_n}(0, W_i)}{g_{\alpha_n}(1, W_i)} + (1-A_i)\hat{c}_4\frac{g_{\alpha_n}(1, W_i)}{g_{\alpha_n}(0, W_i)}\right].
\end{aligned}
$$

$\widehat{IC}_{\alpha_n}(O_i)$ is defined in (A.1).

Appendix B

tmle: An R Package for Targeted Maximum Likelihood Estimation of Binary Point Treatment Effects

Targeted maximum likelihood estimation (TMLE) represents an approach for construction of an efficient double-robust semi-parametric substitution estimator of a target feature of the data generating distribution, such as a variable importance or causal effect parameter. **tmle** is a newly developed R package that implements TMLE for estimation of the effect of a binary treatment at a single point in time on an outcome of interest, controlling for a user supplied covariates: the additive treatment effect, the relative risk, the odds ratio. The package allows that the outcome is subject to missingness, and that one experimental unit contributes repeated records of the point-treatment data structure, thereby allowing this package to analyze longitudinal data structures. The TMLE of the direct effect of the binary treatment, controlling for a binary intermediate variable on the pathway from treatment to the outcome, is also implemented. Relevant factors of the likelihood may be modeled or fit by user-specified commands, or fit data-adaptively internally. Effect estimates, variances, p values, and 95% confidence intervals are provided by the software.

B.1 Introduction

Research in fields such as econometrics, biomedical research, and epidemiology can involve collecting data on a sample from a population in order to assess the population or group level effect of a treatment, exposure, or intervention on a measurable outcome of interest. Obtaining an unbiased and efficient estimate of the statistical parameter of interest necessitates accounting for potential bias introduced through model misspecification or informative treatment assignment or missingness in the outcome data. Due to the curse of dimensionality, parametric estimation approaches are not feasible for high dimensional data without restrictive simplifying modeling assumptions. However, high dimensional data is increasingly common, for example in datasets

used for longitudinal studies, comparative effectiveness research (administrative databases), and genomics. Targeted maximum likelihood estimation (TMLE) is an efficient, double robust, semi-parametric methodology that has been successfully applied in these settings (van der Laan and Rubin, 2006a; van der Laan et al., 2009). The development of the **tmle** package for the R statistical programming environment (Team, 2010) was motivated by the growing need for a user-friendly tool for effective semi-parametric estimation.

TMLE can be applied across a broad range of problems to estimate association and causal effect parameters. The methodology readily incorporates domain knowledge, user-specified parametric models, and optionally allows flexible data-adaptive estimation. The **tmle** package provides an implementation of TMLE for estimating a variety of binary point treatment effect parameters. These parameters include the additive treatment effect for continuous outcomes, risk difference, risk ratio and odds ratio parameters for binary outcomes. The package allows that the outcome is subject to missingness, where it is assumed that the missingness mechanism satisfies the missing at random assumption. Controlled direct effect estimation, estimating the effect of treatment on the outcome at different levels of a (binary) intermediate variable, is also available. (Pearl (2010a) provides a discussion of controlled direct effects.)

Causal effect estimation provides a useful context for describing TMLE methodology and the software implementation. The counterfactual framework discussed in Rubin (1974) frames the estimation of causal effects as a missing data problem. Suppose we are interested in assessing the marginal difference in an outcome, Y, if everyone received treatment ($A = 1$) vs. everyone not receiving treatment ($A = 0$). If we could actually measure the outcome under both scenarios for all individuals, the full data would be given as $X^{Full} = (Y_1, Y_0, W)$, where Y_1 is the counterfactual outcome corresponding to treatment $(A = 1)$, Y_0 is the counterfactual outcome under no treatment $(A = 0)$, and W is a vector of baseline covariates. A causal quantity of interest could be the additive causal effect $E_0Y_1 - E_0Y_0$. The effect estimate could easily be calculated as the average difference over all n subjects in X^{Full}, $1/n\sum_{i=1}^{n} Y_{1i} - Y_{0i}$. Parameters of the full data easily shed light on questions of scientific interest, however in reality the full data can never be known. For each subject we can only observe the outcome corresponding to the actual treatment received. The unobserved counterfactual outcome is missing. Assume the observed data consists of n i.i.d. copies of $O = (W, A, Y = Y_A) \sim P_0$, where P_0 is an unknown underlying probability distribution in a model space $\mathcal{M}$, that gives rise to the data, W is a vector of measured baseline covariates, A is a treatment variable, and Y is the outcome observed under treatment assignment A. By assuming the coarsening at random (CAR) assumption, and a positivity assumption, it follows that the distribution of Y_a can be identified from the observed data distribution P_0 (see e.g., van der Laan and Robins (2003), for general identifiability results for CAR-censored data structures). In this special example, CAR is equivalent with assuming the randomization assumption that A is independent of X, given W.

Non-parametric structural equation modeling (NPSEM) provides an alternative paradigm for defining causal effect parameters (Pearl, 2010b). The following system of equations expresses the

knowledge about the data generating mechanism:

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned}$$

where U_W, U_A, and U_Y are exogenous error terms. This NPSEM allows the definition of counterfactual outcomes $Y_a = f_Y(W, a, U_Y)$, corresponding with the intervention that sets the treatment node A equal to a, and thereby identifies the causal quantity of interest, such as $E_0Y_1 - E_0Y_0$. The functions f_W, f_A, f_Y may be unspecified, one might assume exclusion restriction assumptions, or one might even assume parametric forms. The randomization assumption corresponds with assuming that U_A is independent of U_Y.

The NPSEM approach and the counterfactual framework offer distinct formulations for discussing causality, yet each provides an equivalent foundation for defining causal effects as parameters of statistical distributions. With these definitions in place, we turn our focus to obtaining an efficient, unbiased estimate of the statistical target parameter. Section B.2 provides background on causal effect estimation and defines parameters commonly reported in the literature, the additive effect (risk difference), risk ratio (relative risk), and odds ratio. This section introduces TMLE methodology and describes influence curve-based inference. Section B.3 discusses the implementation in the **tmle** package, including a brief discussion of data-adaptive estimation using the **Super Learner** package, (Polley, 2010), and extensions to missing outcome data and controlled direct effect estimation. An application of the **tmle** program to the analysis of a publicly available dataset is provided in Section B.4. Extensions to the methodology and the software are described in the Discussion section. An FAQ provides answers to commonly asked questions regarding the practical application of TMLE using the software provided in the R package. **tmle** is available for download from the Comprehensive R Archive Network at `http://cran.r-project.org/web/packages/tmle/`.

B.2 Targeted maximum likelihood estimation

B.2.1 Causal inference

Consider the additive effect of a binary treatment on a binary outcome with no missingness. This parameter is defined non-parametrically on full data X^{Full} as $\psi_0^F = E_0Y_1 - E_0Y_0$, and identified from the observed data $O = (W, A, Y = Y_A)$ as $\Psi(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$ under the randomization assumption and positivity assumption. Here ψ_0^F denotes the causal quantity of interest, and ψ_0 is the statistical counterpart that can be interpreted as the causal effect ψ_0^F under these assumptions. We note that Ψ represents a mapping from a probability distribution of O into a real number, and Ψ is called the target parameter mapping. The statistical association measure ψ_0 can be interpreted as the additive causal effect of A on Y providing two

assumptions are met: 1) Coarsening at random (CAR) is an assumption of conditional independence between treatment assignment and the full data given measured covariates, $A \perp X \mid W$ (Heitjan and Rubin (1991), Jacobsen and Keiding (1995), Gill et al. (1997)). This assumption indicates there are no unmeasured confounders of the effect of treatment on the outcome, i.e., U_A is independent of U_Y in the NPSEM. 2) The positivity assumption, also known as the experimental treatment assignment assumption (ETA) is that $\forall a \in \mathcal{A}, P(A = a \mid W) > 0$. In other words, if no observations within some stratum defined by W receive treatment at level $A = a$, then the data do not provide sufficient information to compare the effect of treatment at level a with no treatment, or with treatment at some other level. The parameter is borderline identifiable when there is a practical ETA violation, $\exists a \in \mathcal{A} : P(A = a \mid W) < \epsilon$, for some small ϵ relative to sample size.

A number of estimation procedures have been applied to causal effect estimation, including the maximum likelihood-based G-computation estimator (Robins, 1986), the inverse-probability-of-treatment-weighted (IPTW) estimator (Hernan et al., 2000b; Robins, 2000b), the augmented IPTW estimator (Robins and Rotnitzky, 2001; Robins et al., 2000; Robins, 2000a). Scharfstein et al. (1999) presented a doubly robust regression-based estimator for the treatment specific mean, later extended to time-dependent censoring (Bang and Robins, 2005). See Rosenblum and van der Laan (2010) for a discussion of TMLE in relation to these other estimators. TMLE is a maximum likelihood based G-computation estimator that targets the fit of the data generating distribution towards reducing bias in the parameter of interest, generally one particular low-dimensional feature of the true underlying distribution. TMLE is more generally referred to as Targeted Minimum Loss-based Estimation. At its core, in the above application, TMLE methodology involves fluctuating an initial estimate of the conditional mean outcome, and minimizing a loss function to select the magnitude of the fluctuation. The targeting fluctuation is parameter-specific. The loss function is not unique, but must be chosen with care to ensure that the fluctuated estimate is a parametric sub-model $M \in \mathcal{M}$, and that the risk of the loss function is indeed minimized at the truth. Targeted *maximum likelihood* estimation corresponds with choosing the negative log-likelihood loss function.

An orthogonal factorization of the likelihood of the data is given by

$$\mathcal{L}(O) \;=\; P(Y \mid A, W)P(A \mid W)P(W).$$

We refer to $P(W)$ and $P(Y \mid A, W)$ as the Q portion of the likelihood, $Q = (Q_W, Q_Y)$, and $P(A \mid W)$ as the g portion of the likelihood. Further define

$$\begin{aligned} \bar{Q}_0(A, W) &\equiv E_0(Y \mid A, W), \\ g_0(1 \mid W) &\equiv P_0(A = 1 \mid W), \end{aligned}$$

where the subscript '0' denotes the truth, and a subscript 'n' will denote the corresponding quantity estimated from data. $P_0(W)$ is estimated by the empirical distribution on W, the non-parametric MLE. $\bar{Q}_n(A, W)$ can be obtained by regressing Y on A and W. For some applications g_0 may be known, (e.g., treatment assignment in randomized controlled trials), so that consistent estimation

will be guaranteed. It has been shown that estimation of g_0 leads to increased efficiency even when the true g_0 is known (van der Laan and Robins, 2003).

The additive treatment effect, also referred to as the risk difference when the outcome is binary, is defined non-parametrically as $E(Y_1)-E(Y_0)$. If we let $\mu_1 = E(Y_1)$ and $\mu_0 = E(Y_0)$, the additive treatment effect (ATE) risk ratio (RR) and odds ratio (OR) parameters for binary outcomes are defined as:

$$\begin{aligned} \psi_0^{ATE} &= \mu_{1_0} - \mu_{0_0}, \\ \psi_0^{RR} &= \frac{\mu_{1_0}}{\mu_{0_0}} \\ \psi_0^{OR} &= \frac{\mu_{1_0}/(1-\mu_{1_0})}{\mu_{0_0}/(1-\mu_{0_0})}. \end{aligned} \tag{B.1}$$

Because each of these parameters is a function of (μ_0, μ_1), understanding TMLE of the parameters μ_1 and μ_0 provides a sound basis for understanding the estimation of each point treatment parameter available in the package. Notice that these parameters are functions of the Q portion of the likelihood. TMLE of a target parameter $\Psi(Q_0)$ for a specified target parameter mapping $\Psi()$ is a substitution estimator of the form $\Psi(Q_n^*)$ obtained by plugging in an estimator Q_n^* of Q_0 into the parameter mapping. The g portion of the likelihood is an ancillary nuisance parameter. If $O = (W, A, \Delta, \Delta Y_A)$, then the g-factor factorizes into a treatment assignment mechanism, $g(A \mid W)$ and a missingness mechanism, $\pi(\Delta = 1 \mid A, W)$, where $\Delta = 1$ indicates the outcome is observed, $\Delta = 0$ indicates the outcome is missing. We will first discuss TMLE estimation when there is no missingness, then show how missingness is incorporated into the estimation procedure, and describe estimation of the population mean outcome when a subset of outcomes are unmeasured.

B.2.2 TMLE methodology

TMLE is a regular, asymptotically linear (RAL) estimator. Theory tells us that an efficient RAL estimator solves the efficient influence curve equation for the target parameter up to a second order term (Bickel et al., 1997). Hampel (1974) introduces influence curves and discusses their role in robust estimation. Briefly, an influence curve is a function that describes the behavior of an estimator under slight perturbations of the empirical distribution. For asymptotically linear estimators, the empirical mean of the influence curve of the estimator provides the linear approximation of the estimator. As a consequence, the variance of the influence curve provides the asymptotic variance of the estimator. Among all influence curves for RAL estimators, the one having the smallest variance is known as the efficient influence curve. Because TMLEs solve the efficient influence curve equation, and the efficient influence curves satisfies a so called double robustness property, TMLEs are guaranteed to be asymptotically unbiased if either Q_0 or g_0 is consistently estimated. When both are consistently estimated, TMLEs achieve the semi-parametric efficiency bound, under appropriate conditions.

TMLE is a two-stage procedure. In stage one an initial estimate of the conditional mean outcome, $\bar{Q}_n^0(A,W)$ is obtained. The second stage targets the initial estimate to reduce any residual bias in the estimate of the parameter of interest. This is accomplished by fluctuating the initial estimate in a manner that exploits information in the g portion of the likelihood, designed to ensure that the TMLE solves the efficient influence curve estimating equation.

Given $\bar{Q}_n^0$ and g_n, fluctuating the initial density estimate is straightforward. The direction of the fluctuation determined by the efficient influence curve equations for the target parameters $E(Y_1), E(Y_0)$ is given by

$$H_0^*(A,W) = \frac{I(A=0)}{g(0 \mid W)}, \tag{B.2}$$

$$H_1^*(A,W) = \frac{I(A=1)}{g(1 \mid W)}. \tag{B.3}$$

The TMLE targeting step for updating $\bar{Q}_n^0$ with respect to (EY_1, EY_0), EY_0, and EY_1 is as follows:

$$\begin{aligned} logit(\bar{Q}_n^1(A,W)) &= logit(\bar{Q}_n^0(A,W)) + \epsilon_0 H_0^*(A,W) + \epsilon_1 H_1^*(A,W), \\ logit(\bar{Q}_n^1(0,W)) &= logit(\bar{Q}_n^0(0,W)) + \epsilon_0 H_0^*(0,W), \\ logit(\bar{Q}_n^1(1,W)) &= logit(\bar{Q}_n^0(1,W)) + \epsilon_1 H_1^*(1,W). \end{aligned}$$

Maximum likelihood (*glm*) is used to fit the fluctuation parameter $\epsilon = (\epsilon_0, \epsilon_1)$ that controls the magnitude of the fluctuation. The MLE for ϵ is obtained by a logistic regression of Y on $H_0^*(A,W), H_1^*(A,W)$, with offset $logit(Q_n^0(A,W))$. This fluctuation procedure is generally iterated until convergence. For some parameters, including $E(Y_1)$ and $E(Y_0)$, one-step convergence is guaranteed, hence $\bar{Q}_n^*(A,W) = \bar{Q}_n^1(A,W)$.

The magnitude of ϵ determines the degree of perturbation of the initial estimate, and is a direct function of the degree of residual confounding. For example, when $\bar{Q}_n^0$ is correct, ϵ is essentially 0. It is important to avoid overfitting $\bar{Q}_n^0$, as this minimizes the signal in the residuals needed for bias reduction. Section B.2.4 describes how carrying out the fluctuation on the logit scale even when Y is continuous ensures that the parametric sub-model stays within the defined model space, $\mathcal{M}$.

As discussed above, estimating two parameters $E(Y_1)$ and $E(Y_0)$ allows us to calculate any of the causal effect parameters available for estimation in the **tmle** package. The TMLE estimate of $E(Y_1)$ is given by the G-computation formula $E_{W,n}(\bar{Q}_n^*(1,W)) = \frac{1}{n}\sum_{i=1}^n \bar{Q}_n^*(1,W_i)$, where the marginal distribution of W is estimated with the empirical distribution of $W_1, \ldots, W_n$. The estimate of $E(Y_0)$ has an analogous definition, $E_{W,n}(\bar{Q}_n^*(0,W)) = \frac{1}{n}\sum_{i=1}^n \bar{Q}_n^*(0,W_i)$. The implementation in the **tmle** package targets these two parameters simultaneously. It is also possible to target them separately, or to directly target any specific parameter. For example, the covariate used to target the ATE parameter is given by $H_{ATE}^*(A,W) = I(A=1)/(g(1 \mid W) - I(A=0)/g(0 \mid W)$, where $g(0 \mid W) = 1 - g(1 \mid W)$. However, simultaneous targeting eliminates duplicate calculations, so is sensible from a computational perspective.

B.2.3 Missing outcomes

One problem that frequently arises when analyzing study data is that the outcome may not have been recorded for some observations. A naive estimation approach that considers only complete cases is inefficient, and will be biased when missingness is informative.

Causal inference parameters Consider a randomized clinical trial measuring the effect of treatment on subsequent mortality in which a subset of people in the treatment group become ill, drop out of the study, and die shortly after being lost to follow-up. Because they are no longer in the study, outcome data is missing for these subjects. Further assume that members of the treatment group who remain healthy tend to stay in the study. If observations with missing outcomes are discarded before analyzing the data the estimate of the effect of treatment on mortality will be overly optimistic. An unbiased estimator of the treatment effect must be able to account for this informative missingness.

TMLE does this by exploiting covariate information to reduce both bias and variance. The data are represented in a more general data structure given by $O = (W, A, \Delta, \Delta Y)$, where $\Delta = 1$ indicates the outcome is observed, $\Delta = 0$ indicates the outcome is missing, and $\Delta Y = Y$ when $\Delta = 1$, 0 otherwise. The g-factor of the likelihood now further factorizes into g_A, the treatment mechanism described above, and g_Δ, the missingness mechanism: $g_0 = P(A \mid W)P(\Delta \mid A, W)$. The identifiability result for $E_0 Y_a$ is now given by $E_0 \bar{Q}_0(a, W)$, where $\bar{Q}_0(a, W) = E_0(Y \mid A = a, W, \Delta = 1)$. The clever covariate for targeting the initial estimator of $\bar{Q}_0(A, W) = E_0(Y \mid A, W, \Delta = 1)$ with respect to EY_a is now given by $I(A = a, \Delta = 1)/g(A, \Delta \mid W)$. Thus the above clever covariates are now multiplied by $\Delta / P(\Delta = 1 \mid A, W)$. The regression $\bar{Q}_0$ is now estimated based on the complete observations only.

Population mean outcome Another common research question is determining the marginal mean outcome when some observations are missing the outcome, in the absence of any treatment assignment. The mean outcome conditional on observing the outcome is a biased estimate of the marginal mean outcome ($E(Y_1)$ parameter) when missingness is informative. TMLE can reduce this bias when missingness is a function of measured baseline covariates.

B.2.4 Logistic loss function for continuous outcomes

One obvious approach to applying TMLE with continuous outcomes is to carry out the procedures described above on the linear scale instead of the logit scale, and indeed this has been done successfully in the past. However, particularly when there are ETA violations, this approach can lead to violations of the requirement that the fluctuations of the initial density estimate is a parametric *sub-model* of the observed data model $\mathcal{M}$. A linear fluctuation provides no assurance that the targeted estimate of the conditional mean remains within the parameter space. Gruber and van der Laan (2010a) demonstrates that the negative log-likelihood for binary outcomes is also a valid loss function for continuous outcomes bounded between 0 and 1, and provides a procedure

for mapping Y, a continuous outcome bounded by (a, b), into Y^*, a continuous outcome bounded by (0,1): $Y^* = (Y - a)/(b - a)$. Estimates on the Y^* scale are easily mapped to their counterparts on the original scale:

$$\begin{aligned} EY_0 &= EY_0^*(b-a) + a \\ EY_1 &= EY_1^*(b-a) + a. \end{aligned}$$

Parameter estimates $\psi_n^{ATE}, \psi_n^{RR}, \psi_n^{OR}$ are then calculated as in (B.1).

B.2.5 Controlled direct effect estimation

The **tmle** package also offers controlled direct effect (CDE) estimation. Suppose that in addition to affecting outcome Y directly, treatment A gives rise to an intermediate random variable, Z, that itself has an effect on Y. For example, consider the effect of exercise, A, on weight, Y. Exercise burns calories, directly causing weight loss. Exercise may also affect caloric intake (Z), which has its own effect on weight. One research question might be, *How does weight change with daily exercise?* A second researcher might ask, *What is the effect of daily exercise on weight if caloric intake remains unchanged?* The former requires estimation of the full treatment effect of A on Y, as described above. The latter is an example of a causal effect mediated by an intermediate variable, and requires a modified estimation approach.

The data consists of n i.i.d. copies of $O = (W, A, Z, \Delta, \Delta Y) \sim P_0$, and the likelihood now factorizes as $\mathcal{L}(O) = P(Y \mid \Delta = 1, Z, A, W)P(\Delta = 1 \mid Z, A, W)P(Z \mid A, W)P(A \mid W)P(W)$. Each factor can again be estimated from the data. The **tmle** package restricts controlled direct effect estimation to mediation by a binary variable, Z. Continuing the weight loss example, $Z = 0$ could indicate caloric intake is unaffected by the exercise program, while $Z = 1$ indicates increased caloric intake. CDE estimates calculated at each level of Z provide answers to the second research question posed above.

The first stage of the modified TMLE procedure estimates $\bar{Q}_0(Z, A, W)$. In the second stage $Q_n^0(Z, A, W)$ is fluctuated separately at each level of Z, using modified covariates:

$$\begin{aligned} H_0^*(\Delta, Z, A, W) &= \frac{I(Z = z)}{g_Z(z \mid A, W)} \frac{I(A = 0)}{g_A(1 \mid W)} \frac{1}{g_\Delta(1 \mid Z, A, W)}, \\ H_1^*(\Delta, Z, A, W) &= \frac{I(Z = z)}{g_Z(z \mid A, W)} \frac{I(A = 0)}{g_A(0 \mid W)} \frac{1}{g_\Delta(1 \mid Z, A, W)}. \end{aligned}$$

Here g_Z refers to the conditional distribution of Z given A and W, and ϵ is fit using observations where $\Delta = 1$ and $Z = z$, by default using a logistic fluctuation model.

B.2.6 Inference

Theory tells us that the difference between a parameter estimate obtained from an RAL estimator and the true parameter value converges at a root-n rate to a Normal limit distribution,

$\sqrt{n}(\psi_n - \psi_0) \xrightarrow{D} N(0, \Sigma)$, where Σ is the covariance matrix of the (possibly multi-dimensional) parameter. In practice, this provides a means for estimating the variance of the estimator as the variance of the empirical influence curve divided by the number of i.i.d. units of observation, n. The parameter-specific influence curves are given below. Asymmetric confidence intervals for the RR and OR parameters are constructed on the log scale, based on the influence curves for the log(RR) and log(OR), respectively.

$$\begin{aligned} IC^{EY_1}(O) &= \frac{\Delta}{g_{0_\Delta}(1 \mid A, W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(1, W) - \psi_0^{EY_1} \\ IC^{ATE}(O) &= \left(\frac{A}{g_{0_A}(1 \mid W)} - \frac{1 - A}{g_{0_A}(0 \mid W)}\right) \frac{\Delta}{g_{0_\Delta}(1 \mid A, W)}(Y - \bar{Q}_0(A, W)) \\ &\quad + \bar{Q}_0(1, W) - \bar{Q}_0(A, W) - \psi_0^{ATE} \end{aligned}$$

$$\begin{aligned} IC^{logRR}(O) &= \frac{1}{\mu_{1_0}}\left(\frac{A}{g_{0_A}(1 \mid W)}\frac{\Delta}{g_{0_\Delta}(1 \mid A, W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(1, W) - \mu_{1_0}\right) \\ &\quad - \frac{1}{\mu_{0_0}}\left(\frac{1 - A}{1 - g_{0_A}(1 \mid W)}\frac{\Delta}{g_{0_\Delta}(1 \mid A, W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(0, W) - \mu_{0_0})\right) \end{aligned}$$

$$\begin{aligned} IC^{logOR}(O) &= \frac{1}{\mu_{1_0}(1 - \mu_{1_0})}\left(\frac{A}{g_{0_A}(1 \mid W)}\frac{\Delta}{g_{0_\Delta}(1 \mid A, W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(1, W)\right) \\ &\quad - \frac{1}{\mu_{0_0}(1 - \mu_{0_0})}\left(\frac{1 - A}{1 - g_{0_A}(1 \mid W)}\frac{\Delta}{g_{0_\Delta}(1 \mid A, W)}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(0, W)\right) \end{aligned}$$

Each IC is evaluated by substituting estimates of the true unknown quantities in the above formulas, $\hat{\mu}_0, \hat{\mu}_1, g_{n_A}, g_{n_\Delta}$, and in particular, the targeted estimate $\bar{Q}_n^*(A, W)$ in place of $\bar{Q}_0(A, W)$. A conservative estimate of the variance of the parameter estimate is given by $\hat{\sigma}^2 = var(\widehat{IC}(O))/n$, where n is the number of i.i.d. units of observation. If the dataset contains repeated measures on independent subjects, the subject is considered the unit of observation, and the unit's contribution to the influence curve is equal to the mean contribution for that subject. Ninety-five percent confidence intervals are calculated as $\psi_n(Q_n^*) \pm 1.96\hat{\sigma}/\sqrt{n}$ for the ATE and EY_1 parameters, and $exp(log(\psi_n(Q_n^*)) \pm 1.96\hat{\sigma}/\sqrt{n})$ for the RR and OR parameters, with $\hat{\sigma}$ equal to the estimated standard error of the $log(RR)$ or $log(OR)$ estimates, respectively.

For CDE parameters a term reflecting the contribution of estimating Z is incorporated into each

influence curve:

$$IC^{EY_1}(O) = \frac{I(Z=z)}{g_{0_Z}(Z \mid A,W)} \frac{\Delta}{g_{0_\Delta}(1 \mid A,W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) - \psi_0^{EY_1}$$

$$IC^{ATE}(O) = \frac{I(Z=z)}{g_{0_Z}(Z \mid A,W)} \left(\frac{A}{g_{0_A}(1 \mid W)} - \frac{1-A}{g_{0_A}(0 \mid W)} \right) \frac{\Delta}{g_{0_\Delta}(1 \mid A,W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) - \bar{Q}_0(A,W) - \psi_0^{ATE}$$

$$IC^{logRR}(O) = \frac{1}{\mu_{1_0}} \left(\frac{I(Z=z)}{g_{0_Z}(Z \mid A,W)} \frac{A}{g_{0_A}(1 \mid W)} \frac{\Delta}{g_{0_\Delta}(1 \mid A,W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) - \mu_{1_0} \right) - \frac{1}{\mu_{0_0}} \left(\frac{I(Z=z)}{g_{0_Z}(Z \mid A,W)} \frac{1-A}{g_{0_A}(0 \mid W)} \frac{\Delta}{g_{0_\Delta}(1 \mid A,W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(0,W) - \mu_{0_0}) \right)$$

$$IC^{logOR}(O) = \frac{1}{\mu_{1_0}(1-\mu_{1_0})} \left(\frac{I(Z=z)}{g_{0_Z}(Z \mid A,W)} \frac{A}{g_{0_A}(1 \mid W)} \frac{\Delta}{g_{0_\Delta}(1 \mid A,W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) \right) - \frac{1}{\mu_{0_0}(1-\mu_{0_0})} \left(\frac{I(Z=z)}{g_{0_Z}(Z \mid A,W)} \frac{1-A}{g_{0_A}(0 \mid W)} \frac{\Delta}{g_{0_\Delta}(1 \mid A,W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(0,W) \right)$$

B.3 Implementation in the tmle package

The TMLE algorithm is given by:

1. Obtain $\bar{Q}_n^0(A,W)$, an initial estimate of $P(Y \mid A,W)$
2. Estimate g factors needed to fluctuate $\bar{Q}_n^0(A,W)$ to obtain targeted estimate, $\bar{Q}_n^*(A,W)$
3. Apply target parameter mapping Ψ to targeted estimate Q_n^* using the empirical distribution as estimator of the distribution of W.

The *tmle* function determines which causal effect parameter(s) to estimate based on the values of arguments specified by the user. The data arguments *Y, A, W, Z, Delta*, are the outcome, binary treatment, baseline covariates, mediating binary variable, and missingness indicator, respectively. Only *(Y, W)* must be specified. If A is *NULL* or has no variation (all A are set to 1, or all A are set to 0), the EY_1 parameter estimate is returned. When A is specified, the additive treatment effect is evaluated. If Y is binary, the RR and OR estimates are returned as well. If Z is not *NULL*, the parameter estimates are calculated at each level of $Z \in (0,1)$. Each of these estimation procedures refers to *Delta* to take missingness into account, but missingness does not dictate which parameters are estimated.

When the logistic fluctuation is specified for continuous outcomes, an internal pre-processing step maps $Y \in [a, b]$ to $Y^* \in [0, 1]$ prior to calling the *estimateQ* function to carry out Step 1. *estimateQ* returns an estimate of $\bar{Q}_n^0(A, W)$ on the scale of the linear predictors needed for Step 2: the logit scale for a logistic fluctuation, linear scale for a linear fluctuation. In Step 2, the *estimateG* function is called to estimate each factor of the nuisance parameter required for calculating $H_0^*(A, W)$ and $H_1^*(A, W)$, ϵ is fit using maximum likelihood, and $\bar{Q}_n^*(A, W)$ is calculated. The *calcParameters* function estimates each parameter value, variance, p value, and constructs a 95% confidence interval. The function returns these estimates, along with values for $\bar{Q}_n^0(A, W)$, $\bar{Q}_n^*(A, W)$, and each factor of g. The package provides flexible options for estimating each relevant factor of the likelihood, allowing the procedure to be tailored to the needs of the analysis. These options and their effects are described next.

B.3.1 Stage 1: Estimating $\bar{Q}$

The goal of the first stage of the TMLE procedure is to fit $\bar{Q}_0$ well. A target parameter estimate based on an initial fit that explains a large portion of the variance in Y generally has smaller variance than a target parameter based on a poor initial fit, and a good initial fit also minimizes the reliance on the targeted bias reduction step. Several arguments to the *tmle* function provide flexibility in how the initial fitted values are obtained:

- *Q* optional fitted values for $\bar{Q}_n^0(A, W)$
  - *Q* an $n \times 2$ matrix, $(E(Y \mid A = 0, W), E(Y = 1 \mid A = 1, W))$ For CDE estimation, these values should be fitted with $Z = 0$
  - *Q.Z1* an $n \times 2$ matrix of values fitted with $Z = 1$, only used for CDE estimation
- *Qform* optional regression formula of the form *Y A + W*, suitable for call to *glm*
- *Qbounds* optional truncation levels for Y and $\bar{Q}_n^0(A, W)$ for continuous outcomes
- *Q.SL.library* optional vector of prediction algorithms for data-adaptive estimation

If values are provided for all of the first three arguments to the function, user-specified values, (*Q, Q.Z1*), take precedence. Data-adaptive estimation only occurs if both *Q* and *Qform* are *NULL*. The *Q* argument allows the user to incorporate any estimation procedure into *tmle* by running that procedure externally, obtaining fitted (predicted) values for each counterfactual outcome, $\bar{Q}_n^0(0, W)$ and $\bar{Q}_n^0(1, W)$ and supplying these to the *tmle* procedure. In essence, this option provides unlimited flexibility in obtaining the required stage one estimate of the conditional mean of Y given A and W.

The code snippet below shows a simple application of the *tmle* function using user-specified parametric models to estimate $\bar{Q}$ and g. The models are passed as arguments to the function, along with data arguments (Y, A, W). Default settings imply there is no missing outcome data and that

observations are i.i.d. A sample of size $n = 250$ is drawn from a data generating distribution with true parameter values $\psi_0^{ATE} = 0.216, \psi_0^{RR} = 1.395, \psi_0^{OR} = 2.659$.

```
R> n <- 250
R> W <- matrix(rnorm(n * 3), ncol = 3)
R> colnames(W) <- paste("W", 1:3, sep = "")
R> A <- rbinom(n,1, plogis(0.6 * W[,1] + 0.4 * W[,2]
+                          + 0.5 * W[,3]))
R> Y <- rbinom(n,1, plogis(A + 0.2 * W[,1] + 0.1 * W[,2]
+                          + 0.2 * W[,3]^2))
```

Next parameters are estimated based on correctly specified models for the Q and g factors of the likelihood.

```
R> result.Qcgc <- tmle(Y, A, W, family = "binomial",
+         Qform = Y~ A + W1 + W2 + W3, gform = A ~ W1 + W2 + W3)
R> result.Qcgc
 Additive Effect
   Parameter Estimate:  0.21157
   Estimated Variance:  0.0044941
              p-value:  0.0015995
    95% Conf Interval: (0.080178, 0.34297)

 Relative Risk
   Parameter Estimate:  1.3966
              p-value:  0.0025233
    95% Conf Interval: (1.1244, 1.7347)

              log(RR):  0.33406
    variance(log(RR)):  0.012232

 Odds Ratio
   Parameter Estimate:  2.5554
              p-value:  0.0025418
    95% Conf Interval: (1.3895, 4.6995)

              log(OR):  0.93822
    variance(log(OR)):  0.096621
```

tmle can provide data-adaptive estimation when the **Super Learner** package is installed Polley (2010). Super learning is an ensemble method that relies on proven oracle properties of V-fold cross validation to ascertain an optimal convex combination of estimates obtained from application

of each algorithm in a user-specified library of prediction algorithms (van der Laan et al., 2007). Because one cannot know in advance which class of procedures will be most successful for a given problem, an important aspect of super learning is ensuring that the library of prediction algorithms includes a variety of approaches that search over a large space of possible models. For example, one might include a collection of pre-specified regression models (main terms, main terms plus key interaction terms) along with other flexible modeling approaches, such as non-linear models, cubic splines, and classifiers.

The following example applies super learning to the data generated above in order to estimate $\bar{Q}_0$. The user-specified library contains three prediction algorithms: 1) *SL.glm* is a main terms regression of Y on A and W, 2) *SL.step* calls the *step* function distributed with the base R installation, (Team, 2010), with forward and backward moves incorporating quadratic terms, and 3) SL.DSA.2 calls the *DSA* function in the suggested **DSA** package that uses deletion and addition moves to search over a space of polynomial models that is in this case constrained to order two (Neugebauer and Bullard, 2010). In contrast to the AIC criterion used by the *step* procedure, *DSA* model selection is based on cross-validation (Sinisi and van der Laan, 2004).

```
R> result.QSLgc <- tmle(Y, A, W, family="binomial",
+  Q.SL.library = c("SL.glm", "SL.step", "SL.DSA.2"),
+                 gform = A ~ W1 + W2 + W3,)
R> summary(result.QSLgc)
 Initial estimation of Q
 Procedure: SuperLearner
 Model:
 Y ~  SL.glm_All + SL.step_All + SL.DSA.2_All

 Coefficients:
      SL.glm_All    0
     SL.step_All    0
    SL.DSA.2_All    1


 Estimation of g (treatment mechanism)
 Procedure: user-supplied regression formula
 Model:
 A ~  (Intercept) + W1 + W2 + W3

 Coefficients:
     (Intercept)   -0.01499195
              W1    0.7587852
              W2    0.2719946
              W3    0.3438723
```

```
Estimation of g.Z (intermediate variable assignment mechanism)
Procedure: No intermediate variable

Estimation of g.Delta (missingness mechanism)
Procedure: No missingness

Bounds on g: ( 0.025 0.975 )

Additive Effect
  Parameter Estimate:  0.20889
  Estimated Variance:  0.0045076
             p-value:  0.0018622
   95% Conf Interval: (0.077302, 0.34049)

Relative Risk
  Parameter Estimate:  1.3884
             p-value:  0.0027473
   95% Conf Interval: (1.1201, 1.721)

              log(RR):  0.32814
   variance(log(RR)):  0.012006


Odds Ratio
  Parameter Estimate:  2.5336
             p-value:  0.0030238
   95% Conf Interval: (1.3705, 4.6839)

              log(OR):  0.92965
   variance(log(OR)):  0.098287
```

These parameter estimates and variances using super learning are very similar to those obtained using the correctly specified regression model for $\bar{Q}$, signaling that data-adaptive estimation was successful at discovering the true regression of Y on A and W. *tmle's* default library for estimating $\bar{Q}_0$ contains the three algorithms explicitly included in the example. However, a larger library that incorporates additional estimation procedures is recommended. If the **SuperLearner** package is not available, in the absence of a user-specified regression formula $\bar{Q}_0$ will be estimated using a main terms regression of Y on A and W.

The summary method for *tmle* objects lists the procedures used to estimate the relevant Q and g factors of the likelihood. The super learner is a convex combination of predicted values. When

super learning is used, coefficients reported in the summary reflect each prediction algorithm's contribution. A coefficient of 0 signifies that incorporating predictions from that algorithm does not substantially improve the overall fit given the predictions from algorithms with non-zero coefficients, however, this should not be interpreted as a goodness-of-fit measure. For example, if two model selection algorithms arrive at the same model, at most one will have a non-zero coefficient.

It is important to avoid overfitting $\bar{Q}_n^0$, as this minimizes the signal in the residuals needed for bias reduction. The *tmle* function provides an option for guarding against overfits by cross-validating the initial super learner estimate of $\bar{Q}_0$. The independent units of observation are evenly divided among V folds. Observational units are identified by the *id* variable, an optional argument to the function that if not specified signifies observations are i.i.d. A super learner fit is obtained for each leave-one-fold-out subset of the data, and used to obtain predicted values for observations in the omitted fold. This procedure is invoked by setting *cvQinit = TRUE*.

B.3.2 Stage 2: Targeting the initial estimate

The estimate of the parameter of interest can be biased when $\bar{Q}_n^0$ does not consistently estimate $\bar{Q}_0$. van der Laan and Rubin (2006a) provides a theoretical foundation for constructing a parametric sub-model with fluctuation parameter ϵ that reduces residual bias that is a function of measured covariates. As mentioned above, this fluctuation involves estimating nuisance parameter g_0. Several arguments to the *tmle* function give the user control over the estimation procedure. For estimating the treatment mechanism, g_A:

- *g1W*: the conditional probability of receiving treatment given baseline covariates W
- *gform*: a logistic regression model specification
- *g.SL.library*: a super learner library of prediction algorithms
- *gbound*: a value indicating symmetrical upper and lower bounds on predicted conditional treatment assignment probabilities $(gbound, 1 - gbound)$

The first three of these are similar to the options available for estimating $\bar{Q}_0$. The *gbound* argument is a tuning parameter, conforming with the theoretical guideline that $g_n(A, W)$ must be bounded away from 0 and 1 (van der Laan and Robins, 2003). Bounding will have no effect when no treatment assignments are rare within strata defined by W, e.g., $gbound < g_n < (1 - gbound)$. However, when there is sparsity in the data causing a practical ETA violation, some treatment assignment probabilities will be quite small. As a consequence, some values of $H^*(A, W)$ will be very large for a subset of observations. This lack of identifiability leads to estimates with high variability. Bounding g_n away from (0,1) tends to have a beneficial effect on the variance of the resulting estimate. However, truncation can introduce bias, necessitating a trade-off. These effects are most pronounced when the linear fluctuation is used for continuous outcomes, and largely mitigated by fluctuating on the logit scale (the default).

Though the logistic fluctuation is strongly recommended, the package also provides a linear fluctuation option for continuous outcomes by setting the argument *fluctuation = 'linear'*. Bounding g_n very close to (0,1) typically has little effect on TMLEs obtained using the logistic fluctuation. In contrast, estimates obtained using the linear fluctuation are particularly sensitive to the level of bounding of g_n.

The next coding example illustrates typical effects of different choices of bounds on $g_n(A \mid W)$ on estimation when there is sparsity in the data. The true value for the additive treatment effect for the simulated data is $\psi_0 = 1$. Conditional treatment assignment probabilities $g_A(1 \mid W)$ range from 0.02 to 0.99. The user-supplied regression model for estimating $\bar{Q}_0$ is deliberately misspecified so that estimation is forced to rely on g. The regression formula for $g(1 \mid W)$ is correctly specified, but practical ETA violations lead to estimates with increased bias and variance when the linear fluctuation is employed, as compared to the logistic fluctuation when bounds on g_n are smaller than (0.05, 0.95). Parameter estimates are obtained for 250 samples of size 250.

```
R> n <- 250
R> niterations <- 250
R> gbd <- c(0, 0.01, 0.025, 0.05, 0.1)
R> ngbd <- length(gbd)
R> result.Qmgc <- matrix(NA, nrow = niterations, ncol = 2 * ngbd)

R> for(i in 1:niterations){
+     W <- matrix(rnorm(n * 3), ncol = 3)
+     colnames(W) <- paste("W", 1:3, sep = "")
+     logitA <- 0.5 + 0.9 * W[,1] + 0.5 * W[,2] + 0.7 * W[,3]
+     A <- rbinom(n,1, plogis(logitA))
+     Y <- A + 4 * W[,1] + 4 * W[,2] + 3 * W[,3] + rnorm(n)
+     result.Qmgc[i,] <- c(unlist(sapply(gbd, function(x){
+       tmle(Y, A, W, Qform = Y ~ A, gform = A ~ W1 + W2 + W3,
+       fluctuation = "linear", gbound = x)$estimates$ATE[1]})),
+   unlist(sapply(gbd, function(x){
+      tmle(Y, A, W, Qform = Y ~ A, gform = A ~ W1+ W2 + W3,
+      fluctuation = "logistic", gbound = x)$estimates$ATE[1]})))
+ }
```

Results in table B.1 indicate that the bias of estimates arising from the logistic fluctuation is robust with respect to the choice of bound on g_n, until the bias introduced by bounding at (0.1, 0.9) begins to make a sizable contribution to the MSE. The default bound is $(0.025, 0.975)$.

Recall that the logistic fluctuation for continuous Y requires that Y be bounded by (a, b). When these upper and lower bounds on Y are not provided by the user via the *Qbounds* argument, the default is to use the range of the observed outcomes. This may be problematic when there is missingness in the outcome if the distribution of observed outcomes is truncated with respect to

	Linear			Logistic		
g_n bounds	Bias	Var	MSE	Bias	Var	MSE
(0, 1)	−0.52	0.96	1.24	−0.03	0.11	0.11
(0.01, 0.99)	−0.40	0.56	0.72	−0.03	0.11	0.11
(0.025, 0.975)	−0.21	0.23	0.28	−0.03	0.09	0.09
(0.05, 0.95)	0.03	0.07	0.07	0.07	0.05	0.06
(0.1, 0.9)	0.41	0.07	0.24	0.41	0.07	0.24

Table B.1: A comparison of the effect of bounding g_n using a logistic or linear fluctuation in a sparse data setting.

the true distribution of the outcome, thus using domain knowledge to specify bounds on $\bar{Q}_n$ is encouraged.

B.3.3 Examples with missing outcomes

The *Delta* argument to the *tmle* function is used to indicate which observations have missing outcomes, with *Delta = 1* indicating that the outcome is observed. The *tmle* function ignores the Y value for observations having $\Delta = 0$, so in practice, no special value is reserved to signify missing, and $Y_i = 0$ for observation i is understood to be a valid value when $\Delta_i = 1$. When not explicitly specified, *Delta = 1* is assigned to all observations, signifying that no observations have missing outcomes.

When *Delta = 0* for one or more observations, the missingness mechanism is estimated from the data, or can be user-supplied. When there is missingness, bounds on g_n apply to the product $g_n(\Delta, A, W) = g_A(A \mid W) * g_\Delta(\Delta \mid A, W)$.

The same options are available for estimating g_Δ as for estimating g_A. The relevant arguments to the *tmle* function are:

- *pDelta1* the conditional probability of being observed given treatment assignment A and baseline covariates.
- *g.Deltaform* can be used to specify a regression formula for the regression of Δ on A and W.
- *g.SL.library* specifies a super learner library of prediction algorithms. This same library is used for all factors of g.

When there is no mediating variable, Z, optional argument *pDelta1*, if specified, should be an $n \times 2$ matrix, $P(\Delta = 1 \mid A = 0, W), P(\Delta = 1 \mid A = 1, W)$. When there is a mediating variable, an $n \times 4$ matrix is required: $P(\Delta = 1 | Z = z, A = a, W)$, with (z, a) set to $(0, 0), (0, 1), (1, 0), (1, 1)$, respectively.

Covariates $H_0^*(A,W)$ and $H_1^*(A,W)$ for this more general data structure are given by:

$$\begin{aligned} H_0^*(\Delta,A,W) &= \frac{I(A=0)}{g_A(0 \mid W)} \frac{1}{g_\Delta(1 \mid A,W)}, \\ H_1^*(\Delta,A,W) &= \frac{I(A=1)}{g_A(1 \mid W)} \frac{1}{g_\Delta(1 \mid A,W)}, \end{aligned}$$

and reduce to (B.2) and (B.3), respectively, when there is no missingness. The fluctuation parameter ϵ is fit on observations where $\Delta = 1$. Counterfactual outcomes are obtained for all observations. Accounting for missingness increases efficiency, thus this is beneficial even when missingness is non-informative.

Population mean outcome example The population mean outcome parameter, $E(Y_1)$, is estimated when there is no variation in the treatment assignment for all observations, or when *A = NULL*, and $\Delta = 0$ for some observations. In the next example the true parameter value is $\psi_0^{EY_1} = 0$. $\bar{Q}_n^0$ is based on a deliberately misspecified regression model that is fit on observations where the outcome is observed, i.e., those for which $\Delta = 1$. Because a correctly specified regression model is used to estimate $P(\Delta = 1 \mid W)$, bias is expected to be on the order of $1/\sqrt{n}$. At sample size $n = 250$ used in the example, this is approximately 0.06.

```
R> set.seed(1960)
R> n <- 250
R> W <- matrix(rnorm(n * 3), ncol = 3)
R> colnames(W) <- paste("W",1:3, sep = "")
R> Delta <- rbinom(n, 1, plogis(0.8 + 0.3*W[,1]))
R> Y <- 2 * W[,1] + 4 * W[,2] + 3 * W[,3]+ rnorm(n)
R> Y[Delta == 0] <- NA
R> result.EY1 <- tmle(Y,A = rep(1, n), W, Qform = Y ~ W3,
+                     g.Deltaform = Delta ~ W1, Delta = Delta)
R> result.EY1
 Population Mean
   Parameter Estimate:  -0.043213
   Estimated Variance:  0.15326
              p-value:  0.9121
    95% Conf Interval: (-0.81052, 0.72409)
```

B.3.4 Controlled direct effect estimation example

The first stage of the modified TMLE procedure for CDE estimates $\bar{Q}_0(Z,A,W)$. All estimation options remain available to the user: user-specified values, user-specified parametric model, super learning, cross-validated super learning. Optional user supplied values must be specified

at each level of Z for each subject: the *Q* argument is used to pass in an $n \times 2$ matrix of user-determined values for $\bar{Q}_n^0(Z = 0, A, W)$. The *Q.Z1* argument is used to pass in an $n \times 2$ matrix of user-determined values for $\bar{Q}_n^0(Z = 1, A, W)$.

In the second stage $Q_n^0(Z, A, W)$ is fluctuated separately for $Z = 0$ and $Z = 1$. This requires estimation of an additional nuisance parameter, $g_\Delta = P(\Delta = 1 \mid Z = z, A = a, W))$. The *pZ1* argument allows the user to pass in an $n \times 2$ matrix of conditional probabilities $P(Z = 1 \mid A = 0, W), P(Z = 1 \mid A = 1, W)$. Alternatively, a valid regression formula can be supplied via the *g.Zform* argument. User-supplied values for the conditional mean of Z given A and W may be specified as an $n \times 2$ matrix, *pZ1* = $(P(Z = 1 \mid A = 1, W), P(Z = 1 \mid A = 0, W))$

The following example illustrates CDE estimation in conjunction with missingness in the outcome. A sample of size 1000 is generated, and approximately 25% of outcomes are set to missing.

```
R> n <- 1000
R> W <- matrix(rnorm(n*3), ncol = 3)
R> colnames(W) <- paste("W", 1:3, sep = "")
R> A <- rbinom(n,1, plogis(0.6*W[,1] + 0.4*W[,2] + 0.5*W[,3]))
R> Z <- rbinom(n,1, plogis(0.5 + A))
R> Y <- A + A*Z+ 0.2*W[,1] + 0.1*W[,2] + 0.2*W[,3]^2 + rnorm(n)
R> Delta <- rbinom(n,1, plogis(Z + A))
R> pDelta1 <- cbind(rep(plogis(0), n), rep(plogis(1), n),
+                   rep(plogis(1), n), rep(plogis(2), n))
R> colnames(pDelta1) <- c("Z0A0", "Z0A1", "Z1A0", "Z1A1")
R> Y[Delta == 0] <- NA
```

The regression formula for estimation of $\bar{Q}_0$ is deliberately misspecified in the next call to *tmle*. Super learning is used to estimate the g_A factor of the likelihood, but the specified library contains only one algorithm, *SL.glm*, which performs a main terms regression of the outcome on all available covariates. Estimates of g_Z and g_Δ are passed in to the function. Parameter estimates are reported at each level of Z. The true parameter values are $\psi_{0_{Z0}}^{ATE} = 1$, $\psi_{0_{Z1}}^{ATE} = 2$.

```
R> result.Z.missing <- tmle(Y, A, W, Z, Delta = Delta,
+     pDelta1= pDelta1,  Qform = Y ~ 1, g.SL.library = "SL.glm")
R> result.Z.missing
Controlled Direct Effect
           ----- Z = 0 -----
 Additive Effect
   Parameter Estimate:  1.1094
   Estimated Variance:  0.034713
              p-value:  2.6122e-09
    95% Conf Interval: (0.74419, 1.4745)

           ----- Z = 1 -----
```

```
Additive Effect
  Parameter Estimate:  1.9056
  Estimated Variance:  0.011937
            p-value:  <2e-16
   95% Conf Interval: (1.6914, 2.1197)
```

B.4 FEV data analysis

TMLE was applied to assess the marginal effect of smoking on forced expiratory volume (FEV) using data originally introduced in Rosner (1999b) and discussed in Kahn (2005). The data consists of 654 observations with five variables recorded for each subject: *age* (years), *fev* (liters), *ht* (height in inches), *sex* (0=female, 1=male), *smoke* (0=non smoker, 1=smoker) (Rosner, 1999a). FEV is a measure of pulmonary function that is related to body size and lung capacity. Thus, the relationship between smoking and FEV is likely to be confounded by height, age and sex, all of which influence FEV, and are associated with smoking status. Though height does not have an obvious link to smoking behavior, it may serve as a proxy for health and social factors that influence the decision to smoke. The data are from an observational study of children 3 -19 years old. No children younger than nine years old smoked cigarettes. Therefore, any attempt to estimate a marginal effect of smoking on FEV adjusted for age incurs a theoretical ETA violation due to a complete lack of support in the data. For this reason we restrict the analysis to the subset of data containing $n = 439$ observations on subjects ages 9 - 19.

The observed data consists of n i.i.d. copies of $O = (W, A, Y) \sim P_0$, where $W =$ (*age, ht, sex*), A is an indicator of smoking status, and Y is a continuous measure of FEV. The outcome of interest is the marginal additive effect of smoking on FEV, defined as $E_W[E(Y \mid A = 1, W) - E(Y \mid A = 0, W)]$. If the true regression of Y on A and W were a main terms linear regression, this parameter would correspond to the coefficient in front of the treatment term. However, there is no reason to believe that is the case, and an estimate of the treatment effect based on this misspecified model for $\bar{Q}$ is likely to be biased. The double-robustness property of TMLE tells us that even given a misspecified $\bar{Q}_n^0$, the targeting step can reduce this bias, given a consistent estimate of the treatment mechanism. In the next example we deliberately supply a main terms model for $\bar{Q}$ that we assume is misspecified, and use super learning to estimate $g_A(1 \mid W)$. The algorithms included in the super learner library are:

- *SL.glm* main terms logistic regression of A on W (Team, 2010)
- *SL.step* stepwise forward and backward model selection using AIC criterion, restricted to second order polynomials (Team, 2010)
- *SL.DSA.2* DSA algorithm searching over second order polynomials, substitution and addition moves enabled (Neugebauer and Bullard, 2010)
- *SL.loess* local fitting of a polynomial response surface (*span = 0.75*) (Team, 2010)

- *SL.caret* random forest, with data-adaptively selected value for *mtry* parameter (Kuhn et al., 2010)
- *SL.bart* a classifier based on a Bayesian sum-of-trees model (*ntree = 300*) (Chipman and McCulloch, 2010)
- *SL.knn, SL.knn20, SL.knn40, SL.knn60* k-nearest neighbor algorithm, with neighborhood size, k, set to 10, 20 ,40, 60 (Venables and Ripley, 2002a).

```
R> fev <- read.table("fev.dat",
+            col.names = c("age", "fev", "ht", "sex", "smoke"))
R> fev <- fev[fev$age >= 9,]
R> g.SL.library <- c("SL.glm", "SL.step", "SL.DSA.2","SL.loess",
+                  "SL.caret",  "SL.bart", "SL.knn", "SL.knn20",
+                  "SL.knn40", "SL.knn60")
R> smoke.Qmis <- tmle(Y = fev$fev, A = fev$smoke,
+                     W = fev[ ,c(1, 3, 4)],
+                     Qform = Y ~ ., g.SL.library = g.SL.library)
R> smoke.Qmis
 Additive Effect
   Parameter Estimate:  -0.099653
   Estimated Variance:  0.0045071
              p-value:  0.13771
    95% Conf Interval: (-0.23124, 0.031932)
```

The parameter estimate after targeting is $1/n \sum_{i=1}^{n} \bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i) = -0.10$. Users are often curious about how targeting affects the parameter estimate. The function returns initial (untargeted) predicted values, $\bar{Q}_n^0(0, W), \bar{Q}_n^0(1, W)$. This allows the user to calculate a parameter estimate of -0.16 based on the initial estimate of $\bar{Q}_0$ as follows:

```
R> EY0 <- mean(smoke.Qmis$Qinit$Q[,"Q0W"])
R> EY1 <- mean(smoke.Qmis$Qinit$Q[,"Q1W"])
R> EY1 - EY0
[1] -0.1574331
```

Recall that TMLE is asymptotically efficient when both $\bar{Q}_0$ and g_0 are estimated consistently. In the next example, instead of starting with a deliberately misspecified model for $\bar{Q}_0$, super learning is applied to estimate $\bar{Q}_0$. The prediction algorithm library includes all the algorithms specified for the estimation of g that do not require a binary outcome (everything except the k-nearest neighbor algorithms), and also a linear regression of Y on A and W that includes main terms and all interactions of A and W. We begin by defining a new super learner wrapper function, *SL.glm.int*:

```
R> SL.glm.int <- function(Y.temp, X.temp, newX.temp, family,...){
+    Aint <- paste("A",  colnames(X.temp)[-c(1, 2)], sep = "*")
+    form <- paste("Y.temp ~ Z + ", paste(Aint, collapse = "+"))
+    fit.glm <- glm(form, data = data.frame(Y.temp, X.temp),
+                   family = family)
+    out <- predict(fit.glm, newdata = newX.temp,
+                   type = "response")
+    fit <- list(object = fit.glm)
+    foo <- list(out = out, fit = fit)
+    class(foo$fit) <- c("SL.glm.int")
+    return(foo)
+ }
R> Q.SL.library <- c("SL.glm", "SL.glm.int", "SL.DSA.2",
+                    "SL.loess", "SL.caret", "SL.bart")
```

The library for estimating $\bar{Q}_0$ is passed into the *tmle* function. Because the predicted values for $g_A(1 \mid W)$ are not affected by altering the method used to estimate $\bar{Q}_0$, this next example illustrates a way to reduce computation time by passing in the treatment assignment probabilities obtained from the previous invocation of the function.

```
R> smoke.QSL <- tmle(Y = fev$fev, A = fev$smoke,
+         W = fev[, c(1,3,4)], Q.SL.library = Q.SL.library,
+         g1W = smoke.Qmis$g$g1W)
R> smoke.QSL
Additive Effect
   Parameter Estimate:  -0.11117
   Estimated Variance:  0.0038465
              p-value:  0.073065
    95% Conf Interval: (-0.23272, 0.010393)
```

When a data-adaptive approach to estimating $\bar{Q}_0$ is used, the parameter estimate of -0.11 is quite close to -0.10, the estimate obtained when TMLE was forced to incorporate the misspecified model for $\bar{Q}_n^0$. Super learning also improves efficiency.

Stage one of the TMLE procedure is concerned with explaining the most variance in the outcome. Because ψ_0 is a function of the Q portion of the likelihood, improving the estimate of $\bar{Q}_0$ tends to improve the estimate of ψ_0. However, estimation procedures for $\bar{Q}_0$ have a different goal with respect to the bias/variance tradeoff than do estimators of ψ_0. TMLE's goal is to optimize the tradeoff with respect to ψ_0. Though each TMLE point estimate indicates that smoking decreases FEV, neither estimate is statistically significant. Both analyses are shown in order to highlight salient aspects of the procedure. In practice, the use of super learning and the algorithms included in the library should be a priori specified.

B.5 Discussion

The **tmle** package was designed to provide a flexible, easily customizable implementation of TMLE for binary point treatment effects. A novice user has only to supply the data, while advanced users can control the estimation procedure by overriding default specifications and/or supplying values for $\bar{Q}_n^0$ and g_n from any external estimation procedure. The function can internally estimate any factor of the likelihood with user-supplied linear or logistic regression models, or can use super learning to obtain data-adaptive fits. Covariate information is exploited to reduce bias and increase efficiency in estimates when outcome data is missing. Influence curve-based inference readily accounts for repeated measures. Additionally, the ability to incorporate data-adaptive machine learning techniques while still providing valid inference is a desirable feature of TMLE.

Future extensions to the package include incorporating external weights on observations, and providing TMLE for marginal structural models. Additional loss functions and fluctuation models that increase robustness with respect to outliers and sparsity are under development. TMLE applications to estimate causal effects of multiple time-point interventions while controlling for time-dependent covariates is under development.

Another open area of research is finding an optimal strategy for nuisance parameter estimation. van der Laan and Gruber (2010) presents a theorem on collaborative double robustness of the efficient influence curve that sheds light on this problem. The theorem indicates that depending on the difference $(Q_n - Q_0)$, in addition to g_0 there may exist one or more conditional nuisance parameter distributions that together with the initial estimate solve the estimating equation at the true parameter value, ψ_0. The paper describes a collaborative targeted forward selection algorithm for fitting g that is guided by the goodness-of-fit for the corresponding TMLE of Q_0, and thus on its utility for estimating ψ_0 (see also Gruber and van der Laan (2010b)). A beta version of R software for collaborative TMLE (C-TMLE) is available (Gruber, 2010a).

B.6 Answers to some frequently asked questions (FAQ)

Is there a way to see the parameter estimates based on the initial (untargeted) estimate $\bar{Q}_n^0$? The *tmle* function returns the initial estimates for $\bar{Q}(0, W), \bar{Q}(1, W)$, as a matrix, *result$Qinit$Q*. EY_0 can be estimated as *mean(Qinit$Q[,'Q0W'])*, EY_1 can be estimated as *mean(Qinit$Q[,'Q1W'])*, From there any desired parameter estimate can be calculated. For CDE estimation, *result[[1]]*$Qinit$*Q* corresponds to values obtained when $Z = 0$, and *result[[2]]*$Qinit$*Q* corresponds to values obtained by setting $Z = 1$.

Can I use the package for count data (poisson regression)? Data-adaptive estimation of $\bar{Q}_0$ is not available for count data, but the package can estimate the additive effect of point treatment on a poisson-distributed outcome variable by supplying a formula for poisson regression (log link only), and setting *family = 'poisson'*. The fluctuation will be carried out on the logit scale, unless *fluctuation='linear'* is specified. In this case, despite the name, pois-

son regression will be used to fit ϵ. If data-adaptive estimation of $\bar{Q}_0$ is desired, specify *family='gaussian'*, and externally enforce the constraint that predicted values cannot be less than 0 by specifying *Qbounds = c(0, ub)*, with an appropriate value filled in for the upper bound. Although this will ensure that the initial estimate of the conditional mean outcome is non-negative, unless the logistic fluctuation is used there is no guarantee that the targeted estimate will respect this constraint.

Can I call the *tmle* function a second time without having to re-do the initial estimation of $\bar{Q}_0$? Yes. Predicted values based on the initial estimate $\bar{Q}_0$ are returned as *result$Qinit$Q* (assuming the result of the first call to *tmle* was assigned to the variable named *result*). These values can be passed into a second call to *tmle* by specifying a value for the *Q* argument: *Q = result$Qinit$Q*. For CDE estimation, values for two arguments must be supplied, *Q = result[[1]]$Qinit$Q, Q.Z1 = result[[2]]$Qinit$Q*.

Values for the conditional probabilities for treatment assignment, intermediate variable, and missingness are also available to be examined or passed into a second invocation of the tmle function: *g1W = result$g$g1W, pZ1 = resultg.Zg1W, pDelta1 = result$g.Delta$g1W*. These are untruncated values, regardless of the value of the *gbound* argument.