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

Year 2015 *Paper* 337

Second Order Inference for the Mean of a Variable Missing at Random

Ivan Diaz* Marco Carone†
Mark J. van der Laan‡

*Johns Hopkins University, ildiazm84@gmail.com
†University of Washington, mcarone@uw.edu
‡University of California, Berkeley, laan@berkeley.edu

Second Order Inference for the Mean of a Variable Missing at Random

Ivan Diaz, Marco Carone, and Mark J. van der Laan

Abstract

We present a second order estimator of the mean of a variable subject to missingness, under the missing at random assumption. The estimator improves upon existing methods by using an approximate second order expansion of the parameter functional, in addition to the first order expansion employed by standard doubly robust methods. This results in weaker assumptions about the convergence rates necessary to establish consistency, local efficiency, and asymptotic linearity. The general estimation strategy is developed under the targeted minimum loss based estimation (TMLE) framework. We present a simulation comparing the sensitivity of the first and second order estimators to the convergence rate of the initial estimators of the outcome regression and missingness score. In our simulation, the second order TMLE improved the coverage probability of a confidence interval by up to 53% for slow convergence rates of the initial estimators. In addition, we present a first order estimator inspired by a second order expansion of the parameter functional. This estimator only requires one-dimensional smoothing, whereas implementation of the second order TMLE generally requires kernel smoothing on the covariate space. The first order estimator proposed is expected to have improved finite sample performance, compared to existing first order estimators. In our simulations, the proposed first order estimator improved the coverage probability by up to 68% for slow convergence rates of the initial estimators of the outcome regression and the missingness score. We provide an illustration of our methods using a publicly available dataset targeting the effect of an anticoagulant on health outcomes of patients undergoing percutaneous coronary intervention. In our example, the use of an estimator with expected improved properties changes dramatically the substantive conclusions of the study. We provide R code to implement the proposed estimator.

1 Introduction

Estimation of the mean of an outcome subject to missingness is a problem extensively studied in the statistics literature. Under the assumption that missingness is independent of the outcome conditional on observed covariates, the marginal expectation is identified as a parameter depending on the conditional expectation given covariates among observed individuals (outcome regression henceforth) and the marginal distribution of the covariates. If the covariate vector consists of a few categorical variables, a nonparametric maximum likelihood estimator yields an optimal (consistent, asymptotically linear, and efficient) estimator of the mean outcome. However, if the covariate vector contains continuous variables or its dimension is large, estimation of the outcome regression requires smoothing on the covariate space. In the statistics literature, this has been often achieved by means of a parametric model. Unfortunately, the correct specification of a parametric model is a chimerical tasks in high-dimensional settings or in the presence of continuous variables [11], and data-adaptive prediction estimation methods such as those developed in the statistical learning literature (e.g., super learning, model stacking, bagging, etc.) must be used.

Our methods are developed in the context of targeted learning [13, 14], a field of statistics that deals with the use of data-adaptive methods coupled with optimal estimation theory for semi-parametric models and empirical process theory. In particular, the targeted minimum loss based estimation (TMLE) framework allows consistent and locally efficient estimation of arbitrary low-dimensional parameters in high dimensional models, under regularity and smoothness conditions. In our context, targeted learning allows the incorporation of flexible, data adaptive estimators of the outcome regression into the estimation procedure.

Several doubly robust and locally efficient estimators have been proposed for the missing data problem. These estimators are based on an expansion of the parameter functional in terms of a first order component and a remainder, and are asymptotically linear, under certain conditions. Arguably, the most important condition is that the outcome regression and the probability of missingness conditional on covariates (missingness score henceforth) are estimated consistently at an appropriate rate. A sufficient assumption for establishing $\sqrt{n}$-consistency of doubly robust estimators is that the outcome regression and the missingness score converge to their true values at rates faster than $n^{-1/4}$. We present a second order TMLE that incorporates a second order expansion of the parameter functional in order to relax this assumption, which may be implausible for high dimensions and certain data-adaptive estimators. The method we present is an application of the general higher order estimation theory we present in [3]. We refer to the second order estimator as 2-TMLE in contrast to the first order TMLE discussed by [14], referred to as 1-TMLE.

A complete literature review of higher order estimation theory is presented in [3]. The most relevant references for the problem studied here are [8] and [9]. In particular, [8] presents a particular second order expansion of the target parameter, as well as a second order estimator based on that expansion. This estimator directly uses inverse weighting by a kernel estimate of the covariate density. As a result of the course of dimensionality, the estimator is expected to have poor performance in finite samples as the dimension of the covariate space increases. Particularly, it may fall outside of the parameter space. In contrast, the 2-TMLE presented here is a substitution estimator

that falls in the parameter space with probability one.

As with the estimator presented in [8], implementation of the 2-TMLE requires approximating the second order influence function by means of kernel smoothing, which is subject to the curse of dimensionality. This issue may be circumvented by utilizing an alternative second order expansion that uses kernel smoothing on the missingness score, which is a one-dimensional function of the covariate vector. Since the true missingness score is generally unknown, implementation of this estimator must be carried out using an estimated missingness score. Unfortunately, as discussed in [3], introduction of the estimated missingness score in place of its true value yields a second order remainder term in the analysis of the estimator. As a consequence, the estimator obtained is not a second order estimator. We refer to this estimator as a $1^\star$-TMLE in accordance to this observation. Notably, the second order remainder term obtained with the $1^\star$-TMLE is different from that of the 1-TMLE, which implies they have different finite sample properties. We conjecture that the $1^\star$-TMLE improves finite sample performance over the 1-TMLE, and present a case study in which there are considerable finite sample gains.

Compared to the standard 1-TMLE, implementation of the $1^\star$-TMLE requires the inclusion of one additional covariate in the outcome regression. As a result, its implementation is straightforward and comes at no computational cost. In fact, the potential finite sample gains in performance can be overwhelming, as we illustrate in a simulation studying the coverage probability and mean squared error of the two estimators.

The paper is organized as follows. In Section 2 we review first order efficient estimation theory for the mean outcome in a missing data model. In Section 3 we present the second order expansion of the parameter functional and use it in Section 3.1 to construct a 2-TMLE, for a fixed bandwidth of a kernel smoother. In Section 3.2 we introduce the $1^\star$-TMLE discussed above. In Section 4 we describe a cross-validation method that may be used to select the bandwidth for the 2-TMLE as well as the $1^\star$-TMLE. Section 5 presents a simulation showing that the $1^\star$-TMLE and the 2-TMLE have improved coverage probabilities and mean squared error for slow convergence rates of the estimated outcome regression and missingness score. We conclude with Section 6 illustrating the use of the $1^\star$-TMLE.

2 Review of First Order Estimation Theory

Let W denote a d-dimensional vector of covariates, and let Y denote an outcome of interest measured only when a missingness indicator A is equal to one. We assume that Y is binary or continuous taking values in the interval $(0, 1)$. The observed data $O = (W, A, AY)$ is assumed to have a distribution P_0 in the nonparametric model $\mathcal{M}$. Assume we observe an i.i.d. sample $O_1, \ldots, O_n$, and denote its empirical distribution P_n. For every element $P \in \mathcal{M}$, we define

$$\begin{aligned} Q_W(P)(w) &:= P(W \leq w) \\ g(P)(w) &:= P(A = 1|W = w) \\ \bar{Q}(P)(w) &:= E_P(Y|A = 1, W = w), \end{aligned}$$

where E_P denotes expectation under P. We denote $Q_{W,0} := Q_W(P_0)$, $g_0 := g(P_0)$, and $\bar{Q}_0 := \bar{Q}(P_0)$. We refer to $\bar{Q}$ as the *outcome regression*, and to g as the *missingness score*. We suppress the argument P from the notation $Q_W(P)$, $g(P)$, and $\bar{Q}(P)$ whenever it does not cause confusion. For a function f of o, we use the notation $Pf := \int f(o)\mathrm{d}P(o)$. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a parameter mapping defined as $\Psi(P) := E_P(\bar{Q}(W))$, and let $\psi_0 := \Psi(P_0)$. Under the assumptions that

i) $A \perp\!\!\!\perp Y|W$ under P_0 (missing at random) and

ii) $P_0(g_0(W) > 0) = 1$ (positivity),

it can be shown that $\psi_0 = E_{F_0}(Y)$, where F_0 is the true distribution of the full data (W, Y). Because Ψ depends on P only through $Q := (Q_W, \bar{Q})$, we also use the alternative notation $\Psi(Q)$ to refer to $\Psi(P)$.

First order inference for ψ_0 is based on the following expansion of the parameter functional $\Psi(P)$ around the true P_0:

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)}(P) + R_2(P, P_0), \tag{1}$$

where $D^{(1)}(P)$ is a function of an observation $o = (w, a, y)$ that depends on P, and $R_2(P, P_0)$ is a second order remainder term. The super index (1) is used to denote a first order approximation. This expansion may be seen as analogous to a Taylor expansion when P is finite dimensional, and the expression *second order* may be interpreted in the same way.

We use the expression *first order estimator* to refer to estimators based on first order approximations as in equation (1). Analogously, the expression *second order estimator* is used to refer to estimators based on second order approximations, e.g., as presented in Section 3 below.

Doubly robust locally efficient inference is based on approximation (1) with

$$D^{(1)}(P)(o) = \frac{\mathbb{1}\{a = 1\}}{g(w)}\{y - \bar{Q}(w)\} + \bar{Q}(w) - \Psi(P), \tag{2}$$

$$R_2(P, P_0) = \int \left\{1 - \frac{g_0(w)}{g(w)}\right\}\{\bar{Q}(w) - \bar{Q}_0(w)\}\mathrm{d}Q_{W,0}(w). \tag{3}$$

It is a straightforward algebra exercise to check that equation (1) holds with the definitions given above. $D^{(1)}$ above is referred to as the canonical gradient, or the efficient influence function referred to in the literature [1, 13].

For the sake of limiting the discussion we focus on the analysis of a first order TMLE, other doubly robust locally efficient estimators may be analyzed with similar arguments. First order targeted minimum loss based estimation of ψ_0 proceeds by constructing an estimator $\hat{P}$ of P_0 satisfying $P_n D^{(1)}(\hat{P}) = 0$, and using equation (1) to obtain

$$\hat{\psi} - \psi_0 = (P_n - P_0)D(\hat{P}) + R_2(\hat{P}, P_0),$$

where $\hat{\psi} := \Psi(\hat{P})$. Assuming

i) $D^{(1)}(\hat{P})$ converges to $D^{(1)}(P_0)$ in $L_2(P_0)$ norm,

ii) the size of the class of functions considered for estimation of $\hat{P}$ is bounded (technically, there exists a Donsker class $\mathcal{H}$ so that $D^{(1)}(\hat{P}) \in \mathcal{H}$ with probability tending to one),

empirical process theory (e.g., theorem 19.24 of [16]) shows that

$$\hat{\psi} - \psi_0 = (P_n - P_0)D^{(1)}(P_0) + R_2(\hat{P}, P_0).$$

In addition, if

$$R_2(\hat{P}, P_0) = o_P(1/\sqrt{n}), \tag{4}$$

we obtain

$$\hat{\psi} - \psi_0 = (P_n - P_0)D^{(1)}(P_0) + o_P(1/\sqrt{n}).$$

This implies, in particular, that $\hat{\psi}$ is a $\sqrt{n}$-consistent estimator of ψ_0, it is asymptotically normal, and it is locally efficient.

In this paper we discuss ways of constructing an estimator that requires a consistency assumption weaker than (4). Note that (4) is an assumption about the convergence rate of a second order term involving the product of the differences $\hat{\bar{Q}} - Q_0$ and $\hat{g} - g_0$. Using the Cauchy-Schwarz inequality repeatedly, $|R_2(\hat{P}, P_0)|$ may be bounded as

$$|R_2(\hat{P}, P_0)| \leq ||1/\hat{g}||_\infty \, ||\hat{g} - g_0||_{P_0} \, ||\hat{\bar{Q}} - \bar{Q}_0||_{P_0},$$

where $||f||_P^2 := \int f^2(o) \mathrm{d}P(o)$, and $||f||_\infty := \sup\{f(o) : o \in \mathcal{O}\}$. A set of sufficient conditions for assumption (4) to hold is, for example,

i) $\hat{g}$ is bounded away from zero with probability tending to one

ii) $\hat{g}$ is the MLE of $g_0 \in \mathcal{G} = \{g(w;\beta) : \beta \in \mathbb{R}^d\}$ (i.e., g_0 is estimated in a correctly specified parametric model.) This implies $||\hat{g} - g_0||_{P_0} = O_P(1/\sqrt{n})$.

iii) $||\hat{\bar{Q}} - \bar{Q}_0||_{P_0} = o_P(1)$.

Similarly, a set of sufficient conditions is that $\hat{g}$ is bounded away from zero with probability tending to one, $\hat{\bar{Q}}$ is an MLE in a correctly specified parametric model, and the $L_2(P_0)$ norm of $\hat{g} - g_0$ converges to zero in probability. As discussed in [11], however, correct specification of a parametric models is hardly achievable in high dimensional settings.

Data adaptive estimators must then be used for the outcome regression and missingness score, but they may potentially yield a remainder term R_2 with a convergence rate slower than $n^{-1/2}$. In the next section we present a second order expansion of the parameter functional that allows the construction of estimators that require consistency assumptions weaker than (4).

3 Second Order Estimation

Let us first introduce some notation. For a function $f^{(2)}$ of a pair of observations (o_1, o_2), let $P_0^2 f^{(2)} := \int\int f^{(2)}(o_1, o_2) \mathrm{d}P_0(o_1) \mathrm{d}P_0(o_2)$ denote the expectation of $f^{(2)}$ with respect to the product measure P_0^2.

Second order estimators are based on second order expansions of the parameter functional of the form

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)}(P) - \frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0), \tag{5}$$

where $D^{(2)}$ is a function of a pair of observations (o_1, o_2), and R_3 is a third order term. This representation exists only if W has finite support [3]. If the support of W is infinite, it is necessary to use an approximate second order influence function relying on kernel smoothing, which may introduce challenges due to the curse of dimensionality. In this section we discuss two possible estimation strategies: (i) an estimator that requires kernel smoothing on the covariate vector, and (ii) an estimator that requires kernel smoothing on the missingness score. Strategy (i) is only practical in the presence of a few, possibly data-adaptively selected covariates, whose quantity may increase with sample size. Strategy (ii) requires a-priori knowledge of the true missingness score, and is therefore not applicable in most practical situations. As a solution, we propose to use strategy (ii) with the estimated missingness score, to obtain an estimator denoted as $1^\star$-TMLE. As discussed in [3], the $1^\star$-TMLE is not a second order estimator, since introduction of an estimated missingness score yields a second order term in the remainder term. Nevertheless, the potential finite sample gains obtained with the $1^\star$-TMLE compared to the standard 1-TMLE are worth further investigation. In Section 5.2 we present a simulation study in which the $1^\star$-TMLE showed considerable finite sample improvement of the mean squared error and the coverage probability.

3.1 Second Order Estimator with Kernel Smoothing on the Covariate Vector

Assume W contains only discrete variables. Then the second order expansion (5) holds with

$$D^{(2)}(P)(o_1, o_2) = \frac{2a_1 \mathbb{1}\{w_1 = w_2\}}{g(w_1) q_W(w_1)} \left(1 - \frac{a_2}{g(w_1)}\right) (y_1 - \bar{Q}(w_1)),$$
$$R_3(P, P_0) = \int \left(1 - \frac{g_0(w) q_{W,0}(w)}{g(w) q_W(w)}\right) \left(1 - \frac{g_0(w)}{g(w)}\right) \left(\bar{Q}(w) - \bar{Q}_0(w)\right) \mathrm{d}Q_{W,0}(w),$$

where q_W denotes the density associated to Q_W, and $D^{(1)}$ is defined in (2). It is a straightforward algebra exercise to explicitly check that equation (5) holds.

In many practical situations, however, W is high-dimensional or it contains continuous variables so that the indicator $\mathbb{1}\{w_1 = w_2\}$ has no support. To circumvent this issue, we propose to use the above expansion replacing the indicator function with a kernel function $K_h(w_1 - w_2)$, for a given bandwidth h. We denote such approximation by $D_h^{(2)}$. The conditions under which $D_h^{(2)}$ is an appropriate approximation of $D^{(2)}$ are discussed in [3].

Analogous to the 1-TMLE discussed in the previous section, we construct an estimator $\hat{P}$ satisfying $P_n D^{(1)}(\hat{P}) = P_n^2 D_h^{(2)}(\hat{P}) = 0$. This allows us to exploit expansion (5) and construct a $\sqrt{n}$-consistent estimator in which assumption $R_2(\hat{P}, P_0) = o_P(1/\sqrt{n})$ is replaced by the weaker assumption $R_3(h, \hat{P}, P_0) = o_P(1/\sqrt{n})$, where $R_3(h, \hat{P}, P_0)$ is a third order term similar to $R_3(\hat{P}, P_0)$, defined explicitly in Theorem 3 of [3]. For a complete treatment of other necessary empirical process conditions we refer the interested reader to [3].

For a fixed bandwidth h, the proposed 2-TMLE is given by the following algorithm:

Step 1. *Initial estimators.* Obtain initial estimators $\hat{g}$ and $\hat{\bar{Q}}$ of g_0 and $\bar{Q}_0$. In general, the functional form of g_0 and $\bar{Q}_0$ will be unknown to the researcher. Since consistent estimation of these quantities is key to achieve consistency of $\hat{\psi}$, we advocate for the use of data-adaptive predictive methods that allow flexibility in the specification of these functional forms.

Step 2. *Compute auxiliary covariates.* For each subject i, compute auxiliary covariates $\hat{H}^{(1)}(A_i, W_i)$ and $\hat{H}_h^{(2)}(A_i, W_i)$:

$$\hat{H}^{(1)}(W_i) := \frac{1}{\hat{g}(W_i)}$$
$$\hat{H}_h^{(2)}(W_i) := \frac{1}{\hat{g}(W_i)} \left(1 - \frac{\hat{g}_h(W_i)}{\hat{g}(W_i)}\right),$$

where

$$\hat{g}_h(w) = \frac{\sum_{i=1}^n K_h(w - W_i) A_i}{\sum_{i=1}^n K_h(w - W_i)}$$

is a kernel regression estimator of g_0.

Step 3. *Solve estimating equations.* Estimate the parameter $\epsilon = (\epsilon_1, \epsilon_2)$ in the logistic regression model

$$\text{logit}\, \hat{\bar{Q}}_{\epsilon,h}(w) = \text{logit}\, \hat{\bar{Q}}(w) + \epsilon_1 \hat{H}^{(1)}(w) + \epsilon_2 \hat{H}_h^{(2)}(w), \tag{6}$$

by fitting a standard logistic regression model of Y_i on $\hat{H}^{(1)}(W_i)$ and $\hat{H}_h^{(2)}(W_i)$, with no intercept and with offset $\text{logit}\, \hat{\bar{Q}}(W_i)$, among observations with $A = 1$.

Step 4. *Update initial estimator and compute 2-TMLE.* Update the initial estimator as $\hat{\bar{Q}}_h^\star(w) = \hat{\bar{Q}}_{\hat{\epsilon},h}(w)$, and define the h-specific 2-TMLE as $\hat{\psi}_h = \Psi(\hat{\bar{Q}}_h^\star)$

This algorithm is implemented in the R code provided in Appendix A.

Rationale Behind the Estimation Algorithm: Solving the Relevant Estimating Equations. The main arguments to prove that an estimator $\hat{P}$ solving the estimating equations $P_n D^{(1)}(\hat{P}) = P_n^2 D_h^{(2)}(\hat{P}) = 0$ is asymptotically linear under the assumption that $R_3(\hat{P}, P_0) = o_P(1/\sqrt{n})$ are discussed in [3]. We provide a heuristic argument that the previous algorithm solves these estimating equations. The score equations of the logistic regression model (6) are equal to

$$\sum_{i=1}^{n} \hat{H}^{(1)}(Y_i - \hat{\bar{Q}}_{\epsilon,h}(W_i)) = 0$$
$$\sum_{i=1}^{n} \hat{H}_h^{(2)}(Y_i - \hat{\bar{Q}}_{\epsilon,h}(W_i)) = 0.$$

Because the maximum likelihood estimator solves the score equations, it can be readily seen that

$$\sum_{i=1}^{n} \hat{H}^{(1)}(Y_i - \hat{\bar{Q}}_h^{\star}(W_i)) = 0$$
$$\sum_{i=1}^{n} \hat{H}_h^{(2)}(Y_i - \hat{\bar{Q}}_h^{\star}(W_i)) = 0,$$

which, from the definitions of $\hat{H}^{(1)}$ and $\hat{H}_h^{(2)}$, correspond to $P_n D^{(1)}(\hat{P}) = 0$ and $P_n^2 D_h^{(2)}(\hat{P}) = 0$, respectively.

Comparison with Alternative Second Order Estimators. To the best of our knowledge, the only second order estimator preceding our proposal is discussed in [8]. For a fixed bandwidth h, their estimator is defined as

$$\hat{\psi}_h = \Psi(\hat{P}) + P_n D^{(1)}(\hat{P}) + \frac{1}{2} P_n^2 D_h^{(2)}(\hat{P}). \tag{7}$$

Unlike our proposal, this estimator involved direct computation of D_h^2, which in turn involves inverse weighting by an estimated multivariate density $\hat{q}_W(w)$. As a consequence of the curse of dimensionality these weights may be very unstable, which may lead to a highly variable estimator. In addition, the above estimator does not always satisfy global constraints on the parameter space. In contrast, our proposed 2-TMLE is always in the parameter space, since it is defined as a substitution estimator.

3.2 Second Order Estimator with Kernel Smoothing on the Propensity Score

Smoothing on the support of W to compute $H_h^{(2)}$ might lead to sparsity issues when d is large. As a solution to this problem, consider the second order expansion (5) of parameter $\Psi(P)$ for discrete

W with:

$$D^{(2)}(P)(o_1,o_2) = \frac{2a_1\mathbb{1}\{g_0(w_1)=g_0(w_2)\}}{g(w_1)q_W(w_1)}\left(1-\frac{a_2}{g(w_1)}\right)(y_1-\bar{Q}(w_1)),$$
$$R_3(P,P_0) = \int\left(1-\frac{g_0(w)q_{W,0}(w)}{g(w)q_W(w)}\right)\left(1-\frac{g_0(w)}{g(w)}\right)\left(\bar{Q}(w)-\bar{Q}_0(w)\right)\mathrm{d}Q_{W,0}(w).$$

In contrast to the previous section, here $q_{W,0}(w)$ represents the density function $\frac{d}{dx}P_0(g_0(W)\leq x)|_{x=g_0(w)}$, and $q_W(w)$ represents $\frac{d}{dx}P(g_0(W)\leq x)|_{x=g_0(w)}$. Analogous to the multivariate case, it is often necessary to consider a kernel function $K_h(g_0(w_1)-g_0(w_2))$ instead of the indicator $\mathbb{1}\{g_0(w_1)=g_0(w_2)\}$, which may not be well supported in the data. We denote the approximate second order influence function obtained with such an approximation by $D_h^{(2)}$, to emphasize the dependence on the choice of bandwidth. Using this approximation the estimation procedure described in the previous section may be carried out in exactly the same fashion, but with $\hat{g}_h$ replaced by

$$\hat{g}_h(w) = \frac{\sum_{i=1}^n K_h(g_0(w)-g_0(W_i))A_i}{\sum_{i=1}^n K_h(g_0(w)-g_0(W_i))}.$$

This algorithm yields an asymptotically linear estimator of ψ_0 under the assumption that $R_3(\hat{P},P) = o_P(1/\sqrt{n})$, among other regularity assumptions.

Since g_0 is often unknown, it is necessary to replace it by its estimate $\hat{g}$:

$$\hat{g}_h(w) = \frac{\sum_{i=1}^n K_h(\hat{g}(w)-\hat{g}(W_i))A_i}{\sum_{i=1}^n K_h(\hat{g}(w)-\hat{g}(W_i))}.$$

Unfortunately, a careful analysis of the remainder term associated to this estimator reveals that the introduction of an estimate $\hat{g}$ in place of g_0 yields a second order remainder term. This implies that asymptotic linearity of this estimator (denoted $1^\star$-TMLE) requires the convergence of a second order term in order to be $\sqrt{n}$-consistent. The second order term associated to the $1^\star$-TMLE, however, is different from R_2 defined in (3), required for asymptotic linearity of the 1-TMLE. As a consequence, these estimators are expected to have different finite sample properties. We conjecture that the $1^\star$-TMLE of this section has improved finite sample properties over the 1-TMLE, and present a case study in Section 5 supporting our conjecture.

4 Choosing the Optimal Bandwidth

The estimators presented in the previous section required a user-given bandwidth h. Several options are available for choosing this bandwidth. In principle, it is possible to select the optimal bandwidth using the log-likelihood loss function for estimation of the density q_0. However, because this choice is targeted to estimation of q_0, it may result sub-optimal for estimation of ψ_0. In this section we propose an alternative loss function that targets directly the first order expansion of the parameter of interest. The bandwidth estimator proposed is equivalent to the first step of the collaborative

TMLE (C-TMLE) presented in [15]. This approximation of the C-TMLE is computationally more tractable and works well in our simulations and data analysis, presented below.

Following [4], let $s \in \{1, \ldots, S\}$ index a random sample split into a validation sample $V(s)$ and a training sample $T(s)$. The cross-validation bandwidth selector is defined as

$$\hat{h} := \arg\min_h \{cvRSS(h) + cvVar(h) + n \times [cvBias(h)]^2\},$$

where

$$\begin{aligned}
cvRSS(h) &:= \sum_{s=1}^{S} \sum_{i \in V(s)} (Y_i - \hat{\bar{Q}}^{\star}_{h,s}(W_i))^2, \\
cvVar(h) &:= \sum_{s=1}^{S} \sum_{i \in V(s)} \left[\frac{A_i}{\hat{g}_s(W_i)} (Y_i - \hat{\bar{Q}}^{\star}_{h,s}(W_i)) + \hat{\bar{Q}}^{\star}_{h,s}(W_i) - \hat{\psi}_{h,s} \right]^2, \\
cvBias(h) &:= \frac{1}{S} \sum_{s=1}^{S} \left(\hat{\psi}_{h,s} - \hat{\psi}_h \right),
\end{aligned}$$

are the cross-validated residual sum of squares (RSS), cross-validated variance estimate, and cross-validated bias estimate, respectively. The key idea is to select the bandwidth h that makes $\hat{H}^{(2)}_h$ most predictive of Y, while adding an asymptotically negligible penalty term for increases in bias and variance in estimation of ψ_0. Here $\hat{\bar{Q}}^{\star}_{h,s}$, $\hat{\psi}_{h,s}$, and $\hat{g}_s$ are the result of applying the estimation algorithms described in Section 3, using only data in the training sample $T(s)$.

This loss function is the result of adding a mean squared error (MSE) term $cvVar(h) + n \times [cvBias(h)]^2$ to the usual RSS loss function used in regression problems. Since he MSE contribution to the loss function is asymptotically negligible compared to the RSS, this loss function yields a valid loss function for the parameter $\bar{Q}_0$.

This bandwidth selection algorithm is implemented in the R code provided in Appendix A.

5 Simulation

In this section we present the results of two simulation studies. In Section 5.1, we illustrate the improvements obtained by a 2-TMLE compared to the 1-TMLE. In Section 5.2, we present a simulation study exploring the potential finite sample improvements obtained with the $1^{\star}$-TMLE, compared to the 1-TMLE.

5.1 2-TMLE Smoothing on the Covariate Space

Simulation Setup For each sample size $n \in \{500, 1000, 2000, 10000\}$, we simulated 1000 datasets from the joint distribution implied by the conditional distributions

$$\begin{aligned} W &\sim 6 \times Beta(1/2, 1/2) - 3 \\ A|W &\sim Ber(\text{expit}(1 + W)) \\ Y|A = 1, W &\sim Ber(\text{expit}(-2 + \exp(W) + W)), \end{aligned}$$

where $Ber(\cdot)$ denotes the Bernoulli distribution, expit denotes the inverse of the logit function, and $Beta(a, b)$ denotes the Beta distribution.

For each dataset, we fitted correctly specified parametric models for $\bar{Q}$ and g. For a perturbation parameter p, we then varied the convergence rate of $\hat{\bar{Q}}$ by adding a random variable with Gaussian distribution and mean $3 \times n^{-p}$ and standard deviation n^{-p} to the linear predictor in $\hat{\bar{Q}}$. Analogously, the convergence rate of $\hat{g}$ was varied using a perturbation parameter q by adding Gaussian noise with mean $3 \times n^{-q}$ and standard deviation n^{-q} to the linear predictor. We varied the values of p and q in a grid on the square $(0, 0.5) \times (0, 0.25)$.

We computed a 1-TMLE as well as a 2-TMLE for each initial estimator $(\hat{\bar{Q}}, \hat{g})$ obtained through this perturbation. We compare the performance of the two estimators through their relative MSE compared to the nonparametric efficiency bound, and the coverage probability of confidence interval assuming a known variance. We assume the variance is known in order to isolate any randomness due to its estimation. The MSE and coverage probabilities are approximated through empirical means across the 1000 simulated datasets.

Simulation Results Table 1 shows the relative MSE (rMSE, defined as n times the MSE divided by the efficiency bound) of the 1-TMLE and the 2-TMLE for several sample sizes and selected values of the perturbation parameter (p, q) (those for which we found large differences.) Figure 1 shows the relative MSE for all values of (p, q) used in the simulation.

For all sample sizes, the 2-TMLE improves the rMSE, particularly for slow convergence rates of the outcome regression and the missingness score (small values of the perturbation parameters p and q.) In addition, the rMSE of the 1-TMLE seems to diverge very fast for small values of p and q, as compared to the 2-TMLE which diverges more slowly.

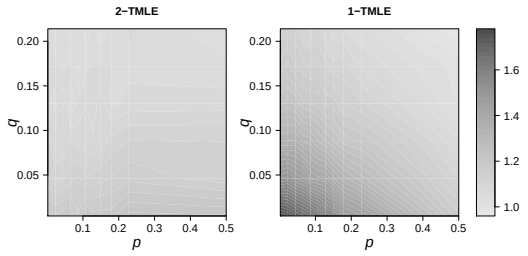
We computed the coverage probability of a Gaussian-based 95% confidence interval based on the 1-TMLE and the 2-TMLE, using the true variance of the estimators. Table 2 shows the results for selected values of the perturbation parameter (p, q). Figure 2 shows the results for all the values of (p, q) used in the simulation.

For all sample sizes, the 2-TMLE has a coverage probability close to the nominal level 0.95. In addition, the coverage probability of the 1-TMLE converges to zero much faster than that of the 2-TMLE for small values of p and q, in agreement with the results for the rMSE shown in Table 1. The largest improvement obtained with the 2-TMLE occurs at a sample size of $n = 10000$ and $(p, q) = (0.002, 0.004)$, where the 1-TMLE has a coverage probability of 0.06 compared to

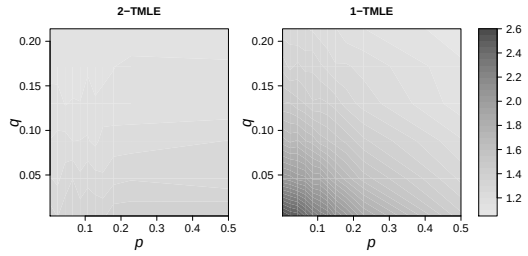
Table 1: Relative MSE of the second order TMLE (2-TMLE) and the first order TMLE (1-TMLE) for different sample sizes and convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.

		1-TMLE				2-TMLE			
		n				n			
q	p	500	1000	2000	10000	500	1000	2000	10000
	0.002	1.76	2.57	3.73	13.91	1.19	1.32	1.39	1.88
0.004	0.044	1.71	2.49	3.48	12.48	1.19	1.33	1.39	1.97
	0.107	1.63	2.24	3.09	10.24	1.20	1.32	1.39	2.10
	0.002	1.50	2.04	2.69	8.05	1.13	1.25	1.29	1.55
0.046	0.044	1.48	1.98	2.58	7.34	1.12	1.27	1.29	1.61
	0.107	1.41	1.89	2.32	6.09	1.12	1.29	1.31	1.74
	0.002	1.24	1.51	1.72	3.31	1.06	1.16	1.18	1.24
0.130	0.044	1.24	1.47	1.65	3.05	1.10	1.15	1.17	1.25
	0.107	1.20	1.42	1.58	2.58	1.08	1.18	1.20	1.29

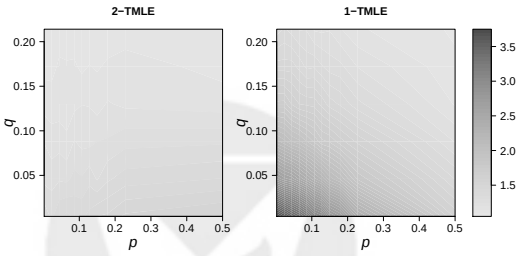
Figure 1: Relative MSE of the second order TMLE (2-TMLE) and the first order TMLE (1-TMLE) for different sample sizes and convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.



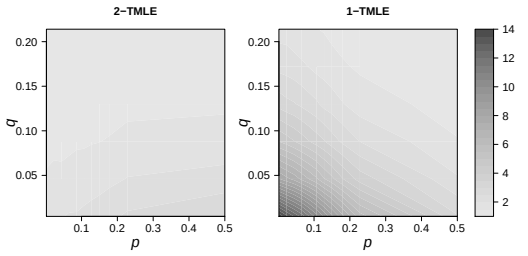
(a) $n = 500$



(b) $n = 1000$



(c) $n = 2000$



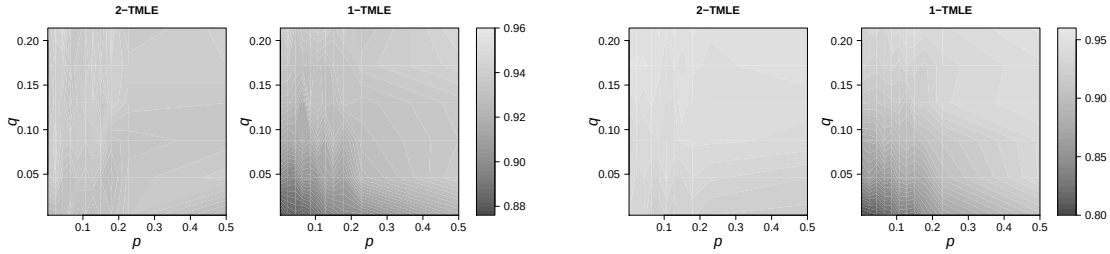
(d) $n = 10000$

0.84 obtained with the 2-TMLE. This simulation illustrates the potential for dramatic improvement obtained by using the 2-TMLE in finite samples.

Table 2: Coverage probabilities of the second order TMLE (2-TMLE) and the first order TMLE (1-TMLE) for different sample sizes and varying convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.

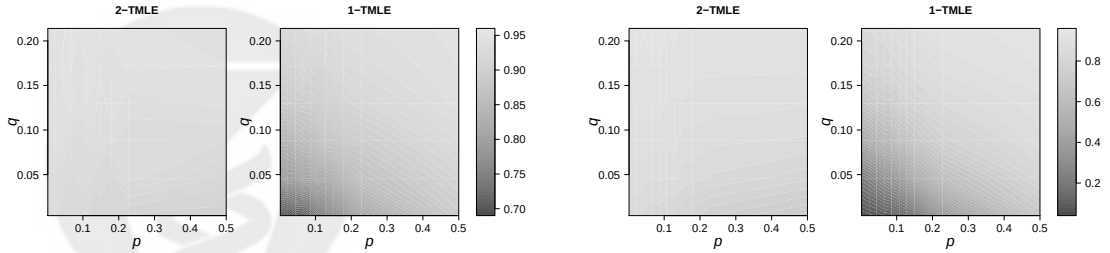
		1-TMLE				2-TMLE			
		n				n			
q	p	500	1000	2000	10000	500	1000	2000	10000
	0.002	0.88	0.80	0.71	0.06	0.93	0.94	0.92	0.84
0.004	0.044	0.89	0.81	0.71	0.08	0.93	0.93	0.92	0.84
	0.107	0.89	0.83	0.74	0.13	0.93	0.93	0.93	0.82
	0.002	0.90	0.85	0.79	0.23	0.94	0.94	0.94	0.88
0.046	0.044	0.90	0.85	0.80	0.29	0.93	0.94	0.93	0.86
	0.107	0.91	0.88	0.82	0.38	0.93	0.94	0.93	0.86
	0.002	0.93	0.91	0.88	0.67	0.94	0.94	0.94	0.91
0.130	0.044	0.92	0.91	0.89	0.71	0.94	0.94	0.94	0.91
	0.107	0.93	0.92	0.90	0.76	0.94	0.94	0.95	0.90

Figure 2: Coverage probabilities of the second order TMLE (2-TMLE) and the first order TMLE (1-TMLE) for different sample sizes and varying convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.



(a) $n = 500$

(b) $n = 1000$



(c) $n = 2000$

(d) $n = 10000$

5.2 Estimator Smoothing on the Estimated Propensity Score

In order to demonstrate the finite sample benefits obtained with the $1^\star$-TMLE, we performed a simulation comparing it to the 1-TMLE in terms of coverage probability and rMSE.

Simulation Setup For each sample size $n \in \{500, 1000, 2000, 10000\}$, we simulated 1000 datasets from the joint distribution implied by the conditional distributions

$$\begin{aligned} W = (W_1, W_2, W_3) &\sim Clayton(2) \\ A|W &\sim Ber(\text{expit}(2 - 2W_1 - 2W_3^2)) \\ Y|A = 1, W &\sim Ber(\text{expit}(-2 + \exp(W_3) + W_1 W_2)), \end{aligned}$$

where $Ber(\cdot)$ denotes the Bernoulli distribution, expit denotes the inverse of the logit function, and $Clayton(\theta)$ denotes the Clayton copula of dimension 3 with distribution function

$$P(W_1 \leq w_1, W_2 \leq w_2, W_3 \leq w_3) = \left(-2 + w_1^{-\theta} + w_2^{-\theta} + w_3^{-\theta}\right)^{1/\theta}.$$

For each dataset, we fitted correctly specified parametric models for $\bar{Q}$ and g. We then varied the convergence rate of $\hat{\bar{Q}}$ and $\hat{g}$ by adding random variables with Gaussian distribution in the same fashion of the previous section.

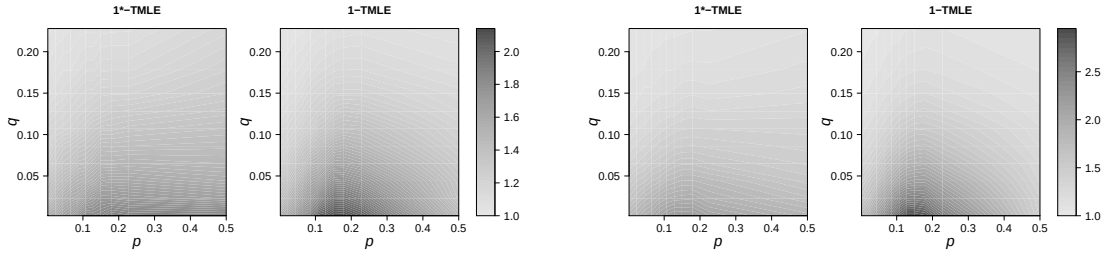
Simulation Results Table 3 shows the relative MSE of each estimator for different sample sizes and selected values of the convergence perturbation (p, q). Figure 3 shows the same results for all values of (p, q) used in the simulation.

The most important result in terms of rMSE is that it converges to infinity much slower for the $1^\star$-TMLE, as compared to the 1-TMLE, for small vales of p and q (Table 3). Figure 3 shows that, when the estimator of g_0 converges very slowly, the 1-TMLE has very bad performance if the estimator of $\bar{Q}_0$ has a moderately slow convergence rate (e.g., $0.05 < p < 3.5$ for $n = 1000$). This is quite interesting since it seems to suggest that for slow convergence rates of $\hat{g}$ it is preferable to inconsistently estimate $\bar{Q}_0$ rather than estimating it at a slow rate, in finite samples. This fact is in accordance to the findings of [6], who state that "... in at least some settings, two wrong models are not better than one." In our simulation, it seems that two models converging at slow rates are not better than one model converging at a slow rate. This fact is attenuated with the use of the $1^\star$-TMLE.

Table 3: Relative mean squared error of the 1*-TMLE and the 1-TMLE for different sample sizes and varying convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.

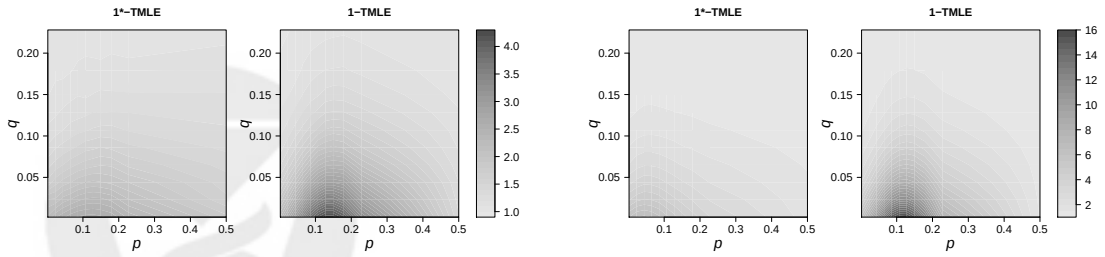
		1-TMLE				1*-TMLE			
		n				n			
q	p	500	1000	2000	10000	500	1000	2000	10000
	0.002	1.55	1.79	2.21	6.16	1.56	1.80	2.22	6.00
0.002	0.065	1.82	2.35	3.41	13.00	1.70	2.01	2.56	6.47
	0.149	2.12	2.91	4.26	14.64	1.84	2.20	2.62	4.73
	0.002	1.32	1.43	1.58	3.02	1.34	1.48	1.63	3.35
0.044	0.065	1.53	1.81	2.30	6.54	1.47	1.65	1.89	3.83
	0.149	1.77	2.17	2.89	8.03	1.60	1.80	2.01	3.05
	0.002	1.18	1.24	1.26	1.77	1.20	1.27	1.32	2.02
0.086	0.065	1.33	1.44	1.63	3.35	1.31	1.38	1.47	2.33
	0.149	1.49	1.70	2.03	4.27	1.42	1.50	1.60	2.04

Figure 3: Relative mean squared error of the 1*-TMLE and the 1-TMLE for different sample sizes and varying convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.



(a) $n = 500$

(b) $n = 1000$



(c) $n = 2000$

(d) $n = 10000$

The coverage probabilities of a Gaussian-based 95% confidence interval assuming known variances are presented in Table 4, for selected values of (p, q), and in Figure 4, for all the values of

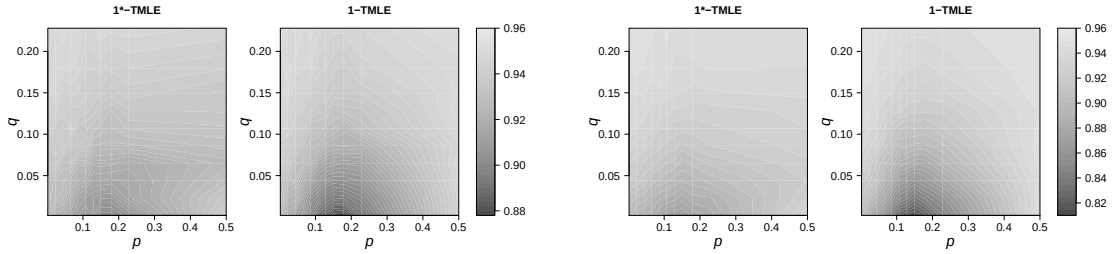
(p, q) considered. Figure 4 shows that the coverage probability of the 1-TMLE degrades as sample size increases for small values of p and q, reaching values as low as 0.12 when $n = 10000$. In comparison, the coverage probability of the $1^\star$-TMLE is never below 0.50 even in the most extreme cases $(p < 0.1, q < 0.05, n = 10000)$.

The largest gain in performance for the $1^\star$-TMLE is obtained when $n = 10000$, $p = 0.149$, and $q = 0.002$ (see Table 4). In that case, the coverage probability of the 1-TMLE is 0.12, whereas the $1^\star$-TMLE has a coverage of 0.65.

Table 4: Coverage probabilities of the $1^\star$-TMLE and the 1-TMLE for different sample sizes and varying convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.

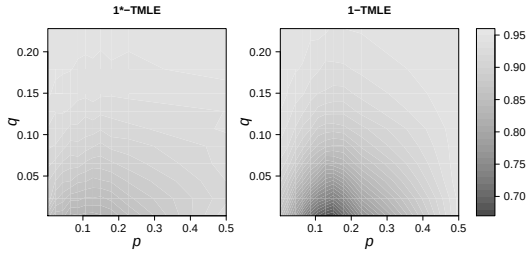
		1-TMLE				$1^\star$-TMLE			
		n				n			
q	p	500	1000	2000	10000	500	1000	2000	10000
0.002	0.002	0.93	0.90	0.86	0.49	0.93	0.90	0.87	0.51
	0.065	0.91	0.86	0.75	0.16	0.92	0.89	0.83	0.47
	0.149	0.88	0.81	0.68	0.12	0.90	0.87	0.83	0.65
0.044	0.002	0.93	0.92	0.90	0.75	0.93	0.92	0.90	0.73
	0.065	0.92	0.89	0.84	0.44	0.93	0.91	0.88	0.67
	0.149	0.90	0.86	0.78	0.35	0.92	0.90	0.87	0.76
0.086	0.002	0.94	0.94	0.93	0.87	0.94	0.93	0.92	0.85
	0.065	0.93	0.92	0.89	0.70	0.93	0.92	0.91	0.81
	0.149	0.92	0.90	0.86	0.61	0.93	0.91	0.90	0.85

Figure 4: Coverage probabilities of the 1*-TMLE and the 1-TMLE for different sample sizes and varying convergence rates of the initial estimators of $\bar{Q}_0$ and g_0.

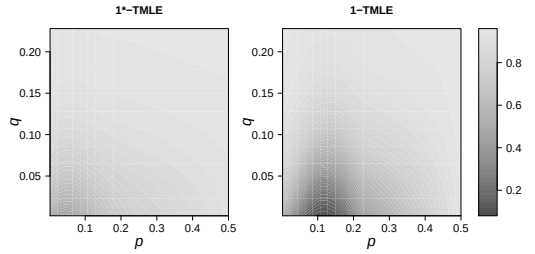


(a) $n = 500$

(b) $n = 1000$



(c) $n = 2000$



(d) $n = 10000$

6 Example

In order to illustrate the method presented, we make use of the dataset `lindner` available in the R package `PSAgraphics`. The dataset contains data on 996 patients treated at the Lindner Center, Christ Hospital, Cincinnati in 1997, and were originally analyzed in [2]. All patients received a Percutaneous Coronary Intervention (PCI). One of the primary goals of the original study was to assess whether administration of Abciximab (an anticoagulant) during PCI improves short and long term health outcomes of patients undergoing (PCI). We reanalyze the `lindner` dataset focusing on the cardiac related costs incurred within 6 months of patients initial PCI as an outcome. The covariates measured are: indicator of coronary stent deployment during the PCI, height, sex, diabetes status, prior acute myocardial infarction, left ejection fraction, and number of vessels involved in the PCI.

As noted by several authors [e.g. 10, 1, 7], causal inference problems may be tackled using methods for outcomes missing at random. Let T denote an indicator of having received Abciximab. Adopting the potential outcomes framework, consider the potential outcomes $Y_t : t \in \{0, 1\}$, given by the outcomes that would have been observed in a hypothetical world if, contrary to the fact, $P(T = t) = 1$. The consistency assumption states that $A = t \rightarrow Y_t = Y$, where Y is the observed outcome. Thus, $E(Y_t)$ may be estimated using methods for missing outcomes, where Y_t

is observed only when $T = t$. In particular, estimation of $E(Y_1)$ and $E(Y_0)$ is carried out using the methods described in the previous sections with $A = T$ and $A = 1 - T$, respectively. Our parameter of interest is the average treatment effect $E(Y_1) - E(Y_0)$.

Since the outcome is continuous, we first used the transformation $(y - \min(y))/(\max(y) - \min(y))$ to map it to the interval $[0, 1]$. We then used the approach outlined in [5] to construct the 1-TMLE and the $1^\star$-TMLE. The distribution of both estimators was estimated with the bootstrap.

The regression of the outcome conditional on covariates was estimated separately for the two treatment groups. Both the outcome regression and the treatment mechanism where estimated using a model stacking technique called super learning [12]. Super learning takes a collection of candidate estimators and combines them in a weighted average, where the weights are chosen to minimize the cross-validated prediction error of the final predictor. The collection of algorithms used is described in Table 5. Table 6 shows the cross-validated risks of the algorithms as well as their weights in the final predictor of $\bar{Q}_0$ and g_0. This illustration is an example of a situation in which the estimator of $\bar{Q}_0$ and g_0 may converge at rates slower than $n^{-1/4}$. Thus, we expect the $1^\star$-TMLE to match or improve the performance of the 1-TMLE.

The R code used to implement the estimators is provided in Appendix A.

Table 5: Prediction algorithms used to estimate $\bar{Q}_0$ and g_0

Algorithm	Description
GLM	Generalized linear model. The logit link was used for g_0 and the identity for Q_0.
BayesGLM	Bayesian GLM. Weakly informative priors were used as implemented by default in the function `bayesglm` of the `arm` package in R.
GAM	Generalized additive model as implemented in the R package `gam`.
PolyMARS	Multivariate adaptive polynomial spline regression implemented in the R package `polspline`.
Earth	Multivariate adaptive regression splines implemented in the R package `earth`.

The unadjusted dollar difference in the outcome between the two groups is equal to US$2216. The first and second order TMLE give an adjusted difference of US$2993.9 and US$2835.7, respectively. Figure 5 presents the bootstrap distributions of the estimators. In particular, the 95% confidence intervals for the 1-TMLE and the $1^\star$-TMLE are $(1115.5, 4858.7)$ and $(-233.9, 4845.8)$, respectively. Note that an analysis based on the 1-TMLE would have concluded an effect significantly different from zero with a 0.05 type I error probability, whereas an analysis based on the $1^\star$-TMLE would have concluded that the effect is not significant. In this illustration the use of an estimator with improved asymptotic properties changes dramatically the substantive conclusion of the study.

Table 6: Cross-validated risk and weight of each algorithm in the super learner for estimation of $\bar{Q}_0$ and g_0.

	$\bar{Q}_0$ Treated		$\bar{Q}_0$ Untreated		g_0	
Algorithm	CV Risk	Weight	CV Risk	Weight	CV Risk	Weight
GLM	0.00275	0.00000	0.00684	0.00000	0.19506	0.00000
BayesGLM	0.00275	0.00000	0.00684	0.00000	0.19502	0.13993
GAM	0.00274	0.65699	0.00679	0.57261	0.19495	0.00000
PolyMARS	0.00280	0.15156	0.00709	0.21333	0.18905	0.62503
Earth	0.00281	0.19145	0.00688	0.21405	0.19332	0.23504

The $1^\star$-TMLE puts in more effort to remove bias caused by slow convergence rates of the initial estimators. As a result, its variance provides confidence intervals that contain the true parameter value with higher probability. However, to obtain this benefit of the $1^\star$-TMLE, it is necessary to use the bootstrap or a second order Taylor expansion, instead of the influence function $D^{(1)}$ which is often used to estimate the variance of the 1-TMLE.

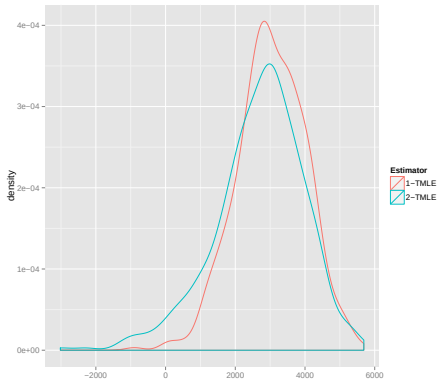


Figure 5: Bootstrap estimate of the density of the estimators.

References

[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] Micheal E Bertrand, Maarten L Simoons, Keith AA Fox, Lars C Wallentin, Christian W Hamm, PJ Feyter, G Specchia, Witold Ruzyllo, and EP McFadden. Management of acute coronary syndromes in patients presenting without persistent st-segment elevation. *European heart journal*, 2002.

[3] Marco Carone, Iván Díaz, and Mark J van der Laan. Higher-order targeted minimum loss-based estimation. 2014.

[4] Susan Gruber and Mark van der Laan. C-tmle of an additive point treatment effect. In Mark van der Laan and Sherri Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

[5] Susan Gruber and Mark 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), 2010.

[6] 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.

[7] Karthika Mohan, Judea Pearl, and Jin Tian. Missing data as a causal inference problem. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*. Citeseer, 2013.

[8] James Robins, Lingling Li, Eric Tchetgen, and Aad W van der Vaart. Quadratic semiparametric von mises calculus. *Metrika*, 69(2-3):227–247, 2009.

[9] James Robins, Eric Tchetgen Tchetgen, Lingling Li, Aad van der Vaart, et al. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009.

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

[11] Richard JCM Starmans. Models, inference, and truth: probabilistic reasoning in the information era. In Mark van der Laan and Sherri Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

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

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

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

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

[16] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Appendix A R code

```
require('SuperLearner')

## Function to truncate probabilities
trunc <- function(x){
  x[x >= 0.999] <- 0.999
  x[x <= 0.001] <- 0.001
  x
}

## Kernel regression
kernelreg <- function(A, X, h){

  reg <- ksmooth(X, A, bandwidth = h, n.points = 1000)
  x <- reg$x[!is.na(reg$y)]
  y <- reg$y[!is.na(reg$y)]
  return(splinefun(x, y))

}

## Function to cross-validate the estimators.
## datat is the training data
## datav is the validation data, may be set to NULL
## h is the bandwidth
## lib is the super learner library

cvfun <- function(datat, datav, h, lib){

  attach(datat)
  on.exit(try(detach(datat), silent = T))

  Y[A==0] <- 9999

  n <- length(A)

  gnfit <- try(SuperLearner(A, W, family = binomial(), SL.library = lib))
  Qnfit <- try(SuperLearner(Y[A==1], W[A==1,], family = gaussian(), SL.library = lib))

  predA <- function(x)trunc(predict(gnfit, newdata=x)$pred[,1])
  predY <- function(x)trunc(predict(Qnfit, newdata=x)$pred[,1])

  Qn  <- try(trunc(predY(W)))
  gn  <- try(trunc(predA(W)))

  predg <- kernelreg(A, gn, h=h)

  gnn <- try(trunc(predg(gn)))

  H1 <- 1/gn
  H2 <- 1/gn * (1 - gnn/gn)

  epsso <- try(coef(glm(Y ~ 0 + offset(qlogis(Qn)) + H1 + H2, family = binomial, subset = A == 1)))

  Qn1 <- trunc(plogis(qlogis(Qn) + epsso[1] / gn + epsso[2]/gn * (1 - gnn/gn)))

  psi <- mean(Qn1)
```

```
  detach(datat)

  if(!is.null(datav)){

    attach(datav)
    on.exit(try(detach(datav), silent = T))

    Y[A==0] <- 9999

    Qnv    <- trunc(predY(W))
    gnv    <- trunc(predA(W))
    gnnv   <- try(predg(gnv))

    Qn1v   <- trunc(plogis(qlogis(Qnv) + epsso[1] / gnv + epsso[2]/gnv * (1 - gnnv/gnv)))
    Dv     <- A/gn * (Y - Qn1v) + Qn1v - psi

    detach(datav)
    resv <- list(psi=psi, Dv=Dv, Qv=Qn1v)

    return(resv)

  } else {

    return(psi)

  }

}

## Function to compute the risk of a candidate h with V-fold cross-validation
cvrisk <- function(h, data, V, lib){

  n   <- length(data$A)
  psi <- cvfun(data, datav=NULL, h, lib=lib)

  valid <- split(sample(1:n), rep(1:V, length = n))

  res <- sapply(valid, function(i){

    datat <- lapply(data, function(x)subset(x, !(1:n %in% i)))
    datav <- lapply(data, function(x)subset(x, 1:n %in% i))
    out   <- cvfun(datat, datav, h, lib)

    cvRSS  <- sum(((datav$Y - out$Qv)[datav$A==1])^2)
    cvVar  <- sum((out$Dv)^2)
    cvBias <- 1/V * (out$psi - psi)

    return(c(cvRSS=cvRSS,cvVar=cvVar,cvBias=cvBias))

  })

  fac <- apply(res, 1, sum)
  risk <- fac[1] + fac[2] + n*fac[3]^2

  return(risk)

}
```

```
require('SuperLearner')
require('PSAgraphics')

data(lindner)
n <- dim(lindner)[1]
A <- lindner[,'abcix']
Y <- lindner[,'cardbill']
range <- range(Y)
Y <- (Y - range[1])/diff(range)
W <- lindner[, c('stent', 'height', 'female', 'diabetic', 'acutemi', 'ejecfrac', 'ves1proc')]

data1 <- list(W=W, A=A, Y=Y)
data0 <- list(W=W, A=1-A, Y=Y)
rm(A, W, Y)

## lib <- c('SL.glm', 'SL.bayesglm', 'SL.gam', 'SL.polymars', 'SL.earth')
## Runs faster:
lib <- c('SL.glm', 'SL.gam')

## Compute the optimal bandwidth
hopt1 <- optimize(cvrisk, interval=c(0.00000001, 5), data1, V=5, lib=lib)$minimum
hopt0 <- optimize(cvrisk, interval=c(0.00000001, 5), data0, V=5, lib=lib)$minimum
## Compute the estimates
psi1  <- cvfun(data1, datav=NULL, h=hopt1, lib=lib)
psi0  <- cvfun(data0, datav=NULL, h=hopt0, lib=lib)

## Back to original Y scale
psi <- (psi1-psi0)*diff(range) + range[1]
```