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

Year 2008 *Paper* 230

Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators

Oliver Bembom* Mark J. van der Laan†

*Division of Biostatistics, University of California, Berkeley, bembom@gmail.com

†Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper230

Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators

Oliver Bembom and Mark J. van der Laan

Abstract

Inverse-Probability-of-Treatment-Weighted (IPTW) estimators are becoming a popular analysis tool in causal inference. It is well known that these estimators suffer from high variability if some treatment probabilities are estimated to be close to zero. While it is a common recommendation for such situations to truncate the weights in order to reduce the mean squared error of the estimator, the current literature gives little guidance on how to select an appropriate truncation level. In this article, we develop a closed-form estimate for the mean squared error of a truncated IPTW estimator that can be used to select this truncation level data-adaptively. While the resulting estimator requires an estimate of an additional nuisance parameter, we show that its consistency does not rely on a consistent estimate of that nuisance parameter. For the case of a binary treatment variable, we present an approach for obtaining an estimate of this nuisance parameter that does not require the user to specify an appropriate parametric model.

We illustrate the practical performance of the proposed estimator in a number of simulation studies that show consistent gains in efficiency relative to more ad-hoc truncation approaches currently in use, with typical gains lying in the range from 1 to 15%. In fact, the estimator is seen to perform on par with an infeasible benchmark estimator that relies on knowledge of the true data-generating distribution. In an applied data analysis, the proposed methodology is estimated to achieve a 7% gain in efficiency relative to the non-truncated IPTW estimator, with truncation resulting in a non-significant finding becoming statistically significant. The methodology presented here has been implemented in an R package called tIPTW that can be downloaded at http://www.stat.berkeley.edu/∼laan/Software/.

1 Introduction

Many applications in epidemiology and clinical research center on estimating the causal effect of a treatment variable on an outcome of interest from observational data. Marginal structural models (MSMs) offer a powerful approach to this problem and are rapidly becoming a standard tool in causal inference (Robins, 1999; Robins et al., 2000; Robins, 2000). Among several estimators that have been proposed for such models, the Inverse-Probability-of-Treatment-Weighted (IPTW) estimator has become a particularly popular choice, due in large part to its straightforward implementation and intuitive interpretation (Robins et al., 2000). By weighting subjects by the inverse of the conditional probability of having selected their observed treatment, given available confounders, this estimator essentially works by creating a new sample in which treatment assignment is independent of the measured confounders.

This approach depends critically on the treatment probabilities used to weight observations. The assumption of experimental treatment assignment (ETA) requires that there exist no values of the confounding factors for which some treatment options have zero probability of being selected. If this assumption is violated, the new sample created by weighting observations fails to be representative of a target population in which treatment has been randomized, causing the IPTW estimator to become inconsistent. Recent work by Neugebauer and van der Laan (2005) has shown that the performance of the IPTW estimator can also be severely compromised if the ETA assumption is satisfied, but some treatment probabilities are very close to zero. First, such a practical violation of the ETA assumption can lead to bias in finite samples. Second, observations with very small treatment probabilities and corresponding large weights can dominate the remainder of the sample so that the estimator can also become highly variable.

Since the latter problem is generally the more pronounced one, one might hope to reduce the mean squared error of the estimator by using truncated weights that, at the price of a slight increase in bias, could lead to a dramatic reduction in variability. While this is a common recommendation for dealing with a practical violation of the ETA assumption, the literature currently gives little guidance on how to select an appropriate truncation level. The most common approach appears to consist of always truncating weights at a fixed level such as 10 or 20, regardless of the particular data set at hand. This seems highly unsatisfactory since the optimal bias-variance trade-off is strongly affected by sample size, with larger data sets in general requiring less truncation than smaller ones. This observation is taken into account by an alternative approach according to which weights are truncated at a level corresponding to, say, 10 or 20% of the sample size. While this is a step in the right direction, the optimal bias-variance trade-off can be expected to depend on a number of additional factors beyond sample size, such as the strength of the ETA violation or the amount of noise encountered in the data.

Wang et al. (2006) recently suggested a more systematic approach to examining the behavior of IPTW estimators that relies on parametric bootstrap samples. First, an estimate of the data-generating distribution is obtained that allows one to simulate realizations of the observed data structure. For this estimated data-generating distribution, the true value of the parameter of interest can be computed by G-computation (Robins, 1986, 1987). The sampling distribution of IPTW estimates obtained by applying the IPTW estimator to a large number of parametric bootstrap samples can therefore be used to compute estimates of the variance, bias, and mean squared error of this estimator. While Wang et al. propose this approach primarily as a tool for diagnosing bias due to a violation of the ETA assumption, they also illustrate its use for quantifying the effects of different choices of truncation levels on the mean squared error of the estimator. The authors stop short, however, of recommending this approach as a formal method for data-adaptive selection of the truncation level, based primarily on the argument that the mean squared error estimates obtained in this manner require a number of additional assumptions beyond those needed for consistency of the IPTW estimator. Specifically, the IPTW estimator only relies on a correct model for the treatment mechanism, i.e. the conditional probability of selecting a treatment option given measured confounders; the parametric bootstrap, however, requires a consistent estimate of the entire data-generating distribution. An additional limitation in selecting the truncation level based on this parametric bootstrap approach lies in its reliance on a large number of simulated data sets. First, such simulations can be computationally intensive so that the approach would not scale well to larger data sets or to applications in which a number of marginal structural models are investigated simultaneously. Second, unless an enormous number of data sets are simulated, the

mean squared error estimates can be expected to be quite sensitive to the exact number of simulated data sets used.

In this article, we develop a methodology for data-adaptively selecting the truncation level that addresses these limitations. The approach is based on mean squared-error estimates that are obtained in closed form, thus avoiding the implementational problems associated with relying on the parametric bootstrap. In addition, the proposed data-adaptive estimator is shown to be consistent under the same conditions required for consistency of the conventional IPTW estimator. Specifically, while the methodology requires an estimate of an additional nuisance parameter beyond the treatment mechanism needed by the IPTW estimator, we show that the data-adaptive estimator converges to the conventional non-truncated estimator even if this nuisance parameter is mis-specified. In addition, we present an approach for avoiding reliance on a parametric model for this nuisance parameter that can be employed if the treatment variable is binary. The remainder of the article is organized as follows: After introducing the methodology in section 2, we illustrate its finite-sample performance in a set of simulation studies. In section 4, we apply the proposed methodology in a data analysis aimed at estimating the causal effect of recent leisure-time physical activity on mortality in the elderly. We then close with a brief discussion of possible extensions to the methodology described here.

2 Methods

2.1 IPTW estimators in causal inference

We consider the common point-treatment data structure consisting of n i.i.d. copies of $O = (W, A, Y)$, where W denotes the collection of measured confounders, A gives the treatment variable, and Y is the outcome of interest. Within the counterfactual framework for causal inference, as first introduced by Neyman (1923) and further developed by Rubin (1978) and Robins (1986, 1987), this observed data structure O is viewed as a coarsened version of a hypothetical full data structure $X = (W, (Y_a : a \in \mathcal{A}))$ that contains the counterfactual outcome Y_a we would have observed on a given subject had she been assigned to treatment a for all a in the collection $\mathcal{A}$ of candidate treatments.

We are frequently interested in parameters of the distribution F_X generating the full data structure X. Examples of such parameters include the mean of a counterfactual outcome such as $E[Y_a]$, representing the mean outcome we would observe if every member of our target population were assigned to treatment $A = a$; the marginal variable importance $E[Y_a - Y_0]$ of treatment A on Y, representing the additive effect on the mean outcome corresponding to changing the treatment assignment of each subject from a reference level $A = 0$ to $A = a$ (van der Laan, 2006); or the parameters of a marginal structural model $E[Y_a \mid V] = m(a, V \mid \beta)$ that models the mean counterfactual outcome $E[Y_a]$ as a function of a within strata defined by a subset V of the baseline covariates W (Robins, 1999; Robins et al., 2000; Robins, 2000).

The observed data O are derived from the full data X through the treatment mechanism $g = g(a \mid X) \equiv P(A = a \mid X)$ that selects for each subject a single counterfactual outcome Y_A corresponding to the observed treatment A: $O = (W, A, Y_A)$. The observed data-generating distribution P_0 is thus defined by F_X and g. Parameters of the full data structure X are identifiable from the observed data if g satisfies the randomization assumption

$$g(a \mid X) = g(a \mid W) \tag{1}$$

and the assumption of experimental treatment assignment (ETA)

$$g(a \mid W) > 0 \text{ for all } a \; F_W\text{-a.e..} \tag{2}$$

The first assumption requires that there are no unmeasured confounders of the relationship between A and Y. According to the second assumption, the confounders W cannot take on values for which certain treatments have zero probability of being selected. If this is not true, the ETA assumption is said to be theoretically violated. If there exist values a of A and w of W such that $g(a \mid w)$ is very close to zero, the ETA assumption is said to be practically violated (Neugebauer and van der Laan, 2005).

Within the general estimating function methodology described in van der Laan and Robins (2003), estimators of a parameter β of F_X can be obtained in two steps. First, an unbiased estimating function $D^{Full}(X \mid F_X)$ of the full data is found, i.e. a function satisfying

$$E_0 D^{Full}(X \mid F_{X,0}) = 0. \tag{3}$$

This function is then mapped into an observed-data estimating function $D(O \mid P)$ that is unbiased under the true data-generating distribution P_0:

$$E_0 D(O \mid P_0) = 0. \tag{4}$$

A popular mapping is given by the IPTW mapping that produces estimating functions $D^{IPTW}(O \mid g, \beta)$ of the observed data that are indexed by the parameter of interest β and the treatment mechanism g. If the parameter of interest is defined as the mean counterfactual outcome $E[Y_a]$, for example, the full-data and IPTW estimating functions are given by

$$\begin{aligned} D^{Full}(X \mid F_X) &= Y_a - \beta & (5)\\ D^{IPTW}(O \mid g, \beta) &= \frac{I(A=a)}{g(A \mid W)}\Big(Y - \beta\Big). & (6) \end{aligned}$$

If the parameter of interest is defined through a marginal structural model $E[Y_a \mid V] = m(a, V \mid \beta)$, we have

$$\begin{aligned} D^{Full}(X \mid F_X) &= \sum_{a \in \mathcal{A}} h(a,V) \frac{\partial}{\partial \beta} m(a, V \mid \beta)\Big(Y_a - m(a, V \mid \beta)\Big) & (7)\\ D^{IPTW}(O \mid g, \beta) &= \frac{h(A,V)}{g(A \mid W)} \frac{\partial}{\partial \beta} m(A, V \mid \beta)\Big(Y - m(A, V \mid \beta)\Big), & (8) \end{aligned}$$

where h is a user-supplied weight function.

If the true treatment mechanism g_0 is known to the investigator, the IPTW estimator is given by the solution of the estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^{n} D^{IPTW}(O_i \mid g_0, \beta). \tag{9}$$

In the more typical scenario of an unknown treatment mechanism, we obtain the estimator as the solution of

$$0 = \frac{1}{n} \sum_{i=1}^{n} D^{IPTW}(O_i \mid g_n, \beta), \tag{10}$$

where g_n is an estimate of g_0. This estimator is consistent if g_n is a consistent estimate of g_0.

2.2 Truncated IPTW estimators

IPTW estimators work by assigning each subject a weight $wt(A, W)$ that is inversely proportional to the conditional probability of having selected their observed treatment, given available confounders. In the case of estimating a mean counterfactual outcome, for example, the estimator relies on weights

$$wt(A, W) = \frac{I(A=a)}{g(A \mid W)}. \tag{11}$$

For marginal structural models, the weights are given by

$$wt(A, W) = \frac{h(A,V)}{g(A \mid W)}, \tag{12}$$

with $h(A,V)$ commonly chosen as $h(A,V) = g(A \mid V) = P(A \mid V)$ to obtain so-called stabilized weights (Robins et al., 2000). Down-weighting observations that were likely to have received their observed treatment and up-weighting those that were instead unlikely to have been observed with the treatment we recorded for them, the IPTW approach essentially creates a new sample in which treatment assignment is independent of the baseline covariates, making it straightforward to estimate the causal parameter of interest.

If some treatment probabilities $g(A \mid W)$ are close to zero, a few observations with correspondingly large weights $wt(A,W)$ may dominate the remainder of the sample and cause the estimator to become highly variable. We are therefore interested in studying IPTW estimators that rely on weights $wt_M(A,W) \equiv min(wt(A,W), M)$ that are truncated at a given constant M. We denote the resulting IPTW estimating functions by $D_M^{IPTW}(O \mid g, \beta)$; depending on whether or not g_0 is known to the investigator, the corresponding estimators are denoted by $\beta_{M,n}(g_0)$ or $\beta_{M,n}(g_n)$. Our goal is to obtain an estimate of the mean squared error of such estimators as a function of M, which would then allow us to define a data-adaptive IPTW estimator based on the truncation constant corresponding to the minimal estimated mean squared error. Since the mean squared error of an estimator can be decomposed into its variance and the square of its bias, it suffices to obtain estimates of these latter two quantities.

The variance of an asymptotically linear estimator can be estimated in a straightforward manner from its influence curve. We will use this approach to estimate the variance of the truncated IPTW estimator $\beta_{M,n}(g_0)$, for which the influence curve is easy to derive. The influence curve of $\beta_{M,n}(g_n)$ is given by the projection of the influence curve for $\beta_{M,n}(g_0)$ onto the orthogonal complement of the tangent space corresponding to the model used for estimating the treatment mechanism (van der Laan and Robins, 2003). Since it therefore depends on the particular model used for estimating the treatment mechanism, this influence curve cannot be derived in general. We will instead estimate the variance of $\beta_{M,n}(g_n)$ based on the influence curve for $\beta_{M,n}(g_0)$. In finite samples, the variance of $\beta_{M,n}(g_n)$ tends to be somewhat smaller than that of $\beta_{M,n}(g_0)$ (van der Laan and Robins, 2003); intuitively, the former estimator can be thought of as adjusting for additional empirical confounding that is not captured by the latter estimator. By using the influence curve of the latter estimator, we will therefore tend to overestimate the variance of the former estimator to some extent.

An analytic estimate for the bias of a truncated IPTW estimator is difficult to obtain in finite samples. We will therefore restrict ourselves to the asymptotic bias of the estimator, ignoring any additional finite-sample bias that may be present as well. Since the two estimators $\beta_{M,n}(g_0)$ and $\beta_{M,n}(g_n)$ are asymptotically equivalent, we only need to derive a single bias estimate. By focusing on asymptotic bias, we will tend to underestimate the bias of the estimator somewhat. Since we will also tend to overestimate the variance of $\beta_{M,n}(g_n)$ slightly, one might hope that these two opposing biases might neutralize each other to some extent in estimating the mean squared error of this estimator. Since g_0 is rarely known to the investigator, this estimator is more frequently encountered in practice than $\beta_{M,n}(g_0)$.

2.3 Estimating the bias of a truncated IPTW estimator

In order to obtain an estimate of the asymptotic bias of a truncated IPTW estimator, we will first derive the limit $\beta_M(P)$ of this estimator under a given data-generating distribution P. The asymptotic bias of the estimator under P is then given by the difference between $\beta_M(P)$ and the corresponding value $\beta(P)$ of the parameter of interest under P. Finally, we estimate the asymptotic bias under P_0 by the plug-in estimate $B_n(M) \equiv \beta_M(P_n) - \beta(P_n)$, where P_n is an estimate of P_0.

The limit $\beta_M(P)$ of the estimator $\beta_{M,n}(g_0)$ under P is given by the solution of

$$0 = E_P D_M^{IPTW}(O \mid g_0, \beta). \tag{13}$$

Assuming that g_n is a consistent estimate of g_0, this also defines the limit of the estimator $\beta_{M,n}(g_n)$ under P. We propose to compute $\beta_M(P)$ by applying the Newton-Raphson algorithm to the function

$$\phi_M : \beta \rightarrow \phi_M(\beta) \equiv E_P D_M^{IPTW}(O \mid g_0, \beta), \tag{14}$$

using as starting value $\beta^0 = \beta(P)$. Given a current estimate β_M^k, this iterative algorithm defines the update

$$\beta^{k+1} = \beta^k - \Big[\frac{\partial}{\partial\beta}\phi_M(\beta)\Big|_{\beta=\beta^k}\Big]^{-1}\phi_M(\beta^k). \tag{15}$$

This update step is carried out until the algorithm converges. If the mapping ϕ is linear, this occurs in a single step so that the asymptotic bias under P can be written as

$$\beta_M(P) - \beta(P) = -\Big[\frac{\partial}{\partial\beta}\phi_M(\beta)\Big|_{\beta=\beta(P)}\Big]^{-1}\phi_M(\beta(P)). \tag{16}$$

Parameters of interest for which this is the case include the mean counterfactual outcome, the marginal variable importance in a non-parametric model, and the parameters of a linear marginal structural model.

Three ingredients are required to compute the asymptotic bias in this manner. First, given a data-generating distribution P, we need to identify the value $\beta(P)$ of the parameter of interest. Second, we need to be able to compute the expectation of the IPTW estimating function at a given β and a given truncation level M. Third, we need to find the inverse of the derivative of ϕ with respect to β. We will now illustrate these steps for parameters of interest defined on the basis of a marginal structural model $E[Y_a \mid V] = m(a, V \mid \beta)$. In this case, the value of the parameter of interest $\beta(P)$ at a particular data-generating distribution P can be obtained as the solution to

$$\begin{aligned} 0 &= ED^{Full}(X \mid \beta) \\ &= E\sum_{a\in\mathcal{A}} h(a,V)\frac{\partial}{\partial\beta}m(a,V\mid\beta)\Big(Y_a - m(a,V\mid\beta)\Big) \\ &= E_W\sum_{a\in\mathcal{A}} h(a,V)\frac{\partial}{\partial\beta}m(a,V\mid\beta)\Big(Q(a,W) - m(a,V\mid\beta)\Big), \end{aligned} \tag{17}$$

where $Q(A,W) \equiv E[Y \mid A,W]$ and all expectations are under P. Equivalently, we have that

$$\begin{aligned} \beta(P) &= \arg\min_\beta E\sum_{a\in\mathcal{A}} h(a,V)\Big(Y_a - m(a,V\mid\beta)\Big)^2 \\ &= \arg\min_\beta E_W\sum_{a\in\mathcal{A}} h(a,V)\Big(Q(a,W) - m(a,V\mid\beta)\Big)^2. \end{aligned} \tag{18}$$

The expectation of $D_M^{IPTW}(O \mid g_0, \beta)$ at β^k is given by

$$\begin{aligned} \phi_M(\beta^k) &= E\left[wt_M(A,W)\frac{\partial}{\partial\beta}m(A,V\mid\beta)\Big|_{\beta=\beta^k}\Big(Y - m(A,V\mid\beta^k)\Big)\right] \\ &= EE_{Y|A,W}\left[wt_M(A,W)\frac{\partial}{\partial\beta}m(A,V\mid\beta)\Big|_{\beta=\beta^k}\Big(Y - m(A,V\mid\beta^k)\Big)\Big|A,W\right] \\ &= E_WE_{A|W}\left[wt_M(A,W)\frac{\partial}{\partial\beta}m(A,V\mid\beta)\Big|_{\beta=\beta^k}\Big(Q(A,W) - m(A,V\mid\beta^k)\Big)\Big|W\right] \\ &= E_W\left[\sum_{a\in\mathcal{A}(W)} wt_M(a,W)\frac{\partial}{\partial\beta}m(a,V\mid\beta)\Big|_{\beta=\beta^k}\Big(Q(a,W) - m(a,V\mid\beta^k)\Big)g(a\mid W)\right], \end{aligned} \tag{19}$$

where $\mathcal{A}(W) \equiv \{a \in \mathcal{A} : g(a \mid W) > 0\}$. The derivative of ϕ_M with respect to β at β^k is given by

$$\begin{aligned}
\frac{\partial}{\partial\beta}\phi_M(\beta)\Big|_{\beta=\beta^k} &= E\frac{\partial}{\partial\beta}\left[wt_M(A,W)\frac{\partial}{\partial\beta}m(A,V \mid \beta)\Big(Y - m(A,V \mid \beta)\Big)\right]\Big|_{\beta=\beta^k} \\
&= E\left[wt_M(A,W)\frac{\partial^2}{\partial\beta^T\partial\beta}m(A,V \mid \beta)\Big|_{\beta=\beta^k}\Big(Y - m(A,V \mid \beta^k)\Big)\right] - \\
&\quad E\left[wt_M(A,W)\frac{\partial}{\partial\beta}m(A,V \mid \beta)\Big|_{\beta=\beta^k}\frac{\partial}{\partial\beta^T}m(A,V \mid \beta)\Big|_{\beta=\beta^k}\right] \\
&= -E\left[wt_M(A,W)\frac{\partial}{\partial\beta}m(A,V \mid \beta)\Big|_{\beta=\beta^k}\frac{\partial}{\partial\beta^T}m(A,V \mid \beta)\Big|_{\beta=\beta^k}\right] \\
&= -E_W\left[\sum_{a\in\mathcal{A}(W)} wt_M(a,W)\frac{\partial}{\partial\beta}m(a,V \mid \beta)\Big|_{\beta=\beta^k}\frac{\partial}{\partial\beta^T}m(a,V \mid \beta)\Big|_{\beta=\beta^k}g(a \mid W)\right] \quad (20)
\end{aligned}$$

To obtain an estimate of the asymptotic bias of $\beta_{M,n}(g_0)$, consider now an estimate P_n of the true data-generating distribution P_0 that estimates the marginal distribution of W by its empirical distribution in the sample, the treatment mechanism by g_n, and the regression $Q(A,W)$ by $Q_n(A,W)$. According to (18), the parameter of interest $\beta(P_n)$ under P_n can then be obtained by creating a new data set that for each subject contains one line for each treatment $a \in \mathcal{A}$ and then regressing the estimated expected outcomes $Q_n(a,W_i)$ under those treatments on the model $m(a,V \mid \beta)$ using weights $h(a,V_i)$. The quantities (19) and (20) can likewise be obtained by plugging in g_n and Q_n for g and Q, respectively, and replacing the expectation over W by the empirical mean over W in the sample.

As an example, consider a linear marginal structural model of the form $m(a,V \mid \beta) = Z_a\beta$, where β is a d-dimensional column vector and Z_a is a d-dimensional row vector, e.g. $Z_a = (1,a,V)$. Then

$$\frac{\partial}{\partial\beta}m(a,V \mid \beta)\Big|_{\beta=\beta^k} = Z_a^T \quad (21)$$

so that we have

$$\phi_M(\beta^k) = E_W\left[\sum_{a\in\mathcal{A}(W)} wt_M(a,W)Z_a^T\Big(Q(a,W) - Z_a\beta^k\Big)g(a \mid W)\right] \quad (22)$$

and

$$\frac{\partial}{\partial\beta}\phi_M(\beta)\Big|_{\beta=\beta^k} = -E_W\sum_{a\in\mathcal{A}(W)} wt_M(a,W)Z_a^TZ_ag(a \mid W). \quad (23)$$

2.4 Estimating the variance of a truncated IPTW estimator

If the treatment mechanism is known, the IPTW estimating function lies in the orthogonal complement of the nuisance tangent space, so that the corresponding estimator is asymptotically linear,

$$\sqrt{n}(\beta_{M,n}(g_0) - \beta_M(P_0)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC_M^{IPTW}(O_i \mid g_0, \beta_M(P_0)) + o_p(1), \quad (24)$$

with influence curve IC_M^{IPTW} equal to an appropriately standardized version of the estimating function itself (Bickel et al., 1993):

$$IC_M^{IPTW}(O \mid g_0, \beta_M(P_0)) = c^{-1}D_M^{IPTW}(O \mid g_0, \beta_M(P_0)), \quad (25)$$

where

$$\begin{aligned}
c &= \frac{\partial}{\partial\beta}ED_M^{IPTW}(O \mid g_0,\beta)\Big|_{\beta=\beta_M(P_0)} \\
&= \frac{\partial}{\partial\beta}\phi_M(\beta)\Big|_{\beta=\beta_M(P_0)}. \quad (26)
\end{aligned}$$

Let c_n be an estimate of c that can be obtained as above by plugging in an estimate P_n of the true data-generating distribution P_0 in (26). The variance of the truncated IPTW estimator $\beta_{M,n}(g_0)$ can then be estimated by $1/n$ times the empirical variance of the estimated influence curve

$$IC_{M,n}^{IPTW}(O \mid g_0, \beta_{M,n}(g_0)) = c_n^{-1} D_M^{IPTW}(O \mid g_0, \beta_{M,n}(g_0)). \tag{27}$$

As described above, we will estimate the variance of the estimator $\beta_{M,n}(g_n)$ based on the corresponding estimated influence curve

$$IC_{M,n}^{IPTW}(O \mid g_n, \beta_{M,n}(g_n)) = c_n^{-1} D_M^{IPTW}(O \mid g_n, \beta_{M,n}(g_n)). \tag{28}$$

Denote the resulting variance estimate by $V_n(M)$.

2.5 A data-adaptively truncated IPTW estimator

Let

$$MSE_n(M) = V_n(M) + B_n^2(M) \tag{29}$$

denote the estimated mean squared error of an IPTW estimator using a truncation level M. We now define a data-adaptively truncated IPTW estimator based on the truncation level $\underline{M}_n$ that is estimated to lead to a minimal mean squared error. To simplify arguments regarding the consistency of this estimator, the precise definition of $\underline{M}_n$ relies on the quantity

$$MSE_{\epsilon,n}(M) = \max(MSE_n(M), \epsilon), \tag{30}$$

for a small positive ϵ such as $\epsilon = 10^{-16}$, that can be viewed as a version of $MSE_n(M)$ as stored by a computer that is unable represent numbers smaller than ϵ. The selected truncation level $\underline{M}_n$ is then defined as

$$\underline{M}_n = \max\Big\{M > 1 : MSE_{\epsilon,n}(M) = \min_M MSE_{\epsilon,n}(M)\Big\}. \tag{31}$$

If $MSE_n(M) > \epsilon$ for $M > 1$, (31) simply selects the minimizer of $MSE_n(M)$. If $MSE_n(M) \leq \epsilon$ for some values of M, however, we will select the largest values of $M > 1$ for which this is the case, which may not necessarily be the true minimizer of $MSE_n(M)$.

While the variance estimate $V_n(M)$ only requires a consistent estimate of the treatment mechanism, as also needed by the IPTW estimator itself, the bias estimate $B_n(M)$ depends in addition on an estimate Q_n of the regression Q. One might therefore be concerned that the data-adaptively truncated estimator defined here could be inconsistent in situations in which the conventional non-truncated IPTW is consistent, namely if the model for g is correctly specified while Q is estimated inconsistently. To show that the consistency of the estimator does in fact not depend on a consistent estimate of the additional nuisance parameter Q, note first that $B_n(\infty) = 0$ as long as g_n satisfies the ETA assumption, regardless of the estimate Q_n of Q. This is true since the non-truncated IPTW estimator is consistent and therefore asymptotically unbiased if the randomization assumption and ETA assumption hold and g is estimated consistently. The asymptotic bias estimate $B_n(M)$ is obtained under an estimated data-generating distribution P_n for which the randomization assumption is trivially satisfied; in addition, the derivation of $B_n(M)$ incorporates explicitly the assumption that g is estimated consistently. It follows therefore, that $B_n(\infty) = 0$ as long as the estimated treatment mechanism satisfies the ETA assumption. Since $V_n(M) \to 0$ at $\sqrt{n}$-rate, it follows then that the same condition implies $MSE_n(\infty) \to 0$. This in turn guarantees that $\underline{M}_n \to \infty$ so that the data-adaptively truncated estimator introduced here converges to the conventional non-truncated IPTW estimator as $n \to \infty$. Since this convergence does not depend on the estimate Q_n for Q, the data-adaptively truncated estimator is consistent as long as the randomization assumption and ETA assumption hold and g is estimated consistently, i.e. under precisely the same conditions that ensure consistency of the conventional IPTW estimator.

In spite of this consistency result, the finite-sample performance of the data-adaptively truncated IPTW estimator cannot be expected to be impervious to grossly inconsistent estimates Q_n. For this reason, it would be appealing to obtain these estimates in a manner that does not require the user to specify a parametric

model for Q. A particularly straightforward approach to this problem exists for the important special case of a binary treatment variable A. In that case, we may obtain an estimate of Q by regressing Y not on A and W, but instead on A and the estimated propensity score $g_n(1 \mid W)$. The propensity score represents a univariate summary measure of the potentially high-dimensional collection of confounders W that captures all the information necessary to adjust for confounding by W (Rosenbaum and Rubin, 1983). Since $g_n(1 \mid W)$ is a one-dimensional covariate, it is considerably easier to obtain a flexible data-adaptive fit for a regression of Y on A and $g_n(1 \mid W)$ than for a regression of Y on A and W. Specifically, we propose to use a generalized additive model (Hastie and Tibshirani, 1990) for this purpose that simply includes an indicator variable for A and a smoothing spline with four degrees of freedom for the logit of the estimated propensity score. We denote the resulting data-adaptively truncated IPTW estimator by $\beta_n(g_0)$ or $\beta_n(g_n)$ depending on whether the treatment mechanism is known or not. Alternatively, the user may supply a parametric model for Q to obtain an estimator $\beta_n(g_0, Q_n)$ or $\beta_n(g_n, Q_n)$.

3 Simulation study

In this section, we present simulation studies aimed at examining the finite-sample performance of the data-adaptively truncated estimators $\beta_n(g_n, Q_n)$ and $\beta_n(g_n)$.

3.1 Data generating distribution and parameter of interest

We consider a point-treatment data structure $O = (W, A, Y)$, with $W = (W_1, W_2, W_3, W_4)$ containing four potential confounding factors, A denoting a binary treatment variable, and Y representing a continuous outcome of interest. Given a treatment mechanism $g_0(A \mid W)$ and the regression function $Q_0(A, W)$, the observed data structure was generated as follows:

1. Generate W_1, W_2, W_3, and W_4 as independent random uniform variables over the interval $[0, 1]$.
2. Generate the observed treatment variable A from the conditional distribution of A given W, $g_0(A \mid W)$.
3. Generate the observed outcome Y as $Y = Q_0(A, W) + \epsilon$ with $\epsilon \sim N(0, 1)$.

We consider the two different treatment mechanism

$$\text{logit}\Big(g_{1,0}(A \mid W)\Big) = -1 + 2W_1 - 2W_2 + W_3W_4 \tag{32}$$

and

$$\text{logit}\Big(g_{2,0}(A \mid W)\Big) = -1 + 2W_1 - 4W_2 + W_3W_4. \tag{33}$$

As described in section 2, the analytic bias estimate introduced here tends to underestimate the true bias of the estimator somewhat by focusing entirely on its asymptotic bias. This might be expected to pose a problem in situations in which the ETA assumption is practically violated, which can lead to considerable finite-sample bias that is not taken into account by the approach described here. The treatment mechanism $g_{1,0}$ is used to examine the behavior of the data-adaptive estimator in situations in which treatment probabilities are clearly bounded away from zero so that the finite-sample bias of the estimator should be negligible. The treatment mechanism $g_{2,0}$, on the other hand, is intended to represent a strong practical violation of the ETA assumption.

We consider the two regression functions

$$Q_{1,0}(A, W) = -1 + A + W_1 - W_2 + 2AW_1 + W_3W_4 \tag{34}$$

and

$$Q_{2,0}(A, W) = -1 + A + 5W_1 - W_2 + 2AW_1 + W_3W_4. \tag{35}$$

The analytic variance estimate can be biased high for an IPTW estimator that is based on an estimated treatment mechanism. This might be expected to pose a problem particularly in situations in which confounders are highly predictive of the outcome of interest so that adjusting for additional empirical confounding could lead to a considerable reduction in variability. Under the regression function $Q_{1,0}$, the confounders W are moderately predictive of Y. The regression function $Q_{2,0}$, on the other hand, is intended to examine the potential impact of overestimating the true variance of the estimator in situations in which W is highly predictive of Y.

The causal parameter of interest is defined through the simple marginal structural model

$$E[Y_a] = \beta_0 + \beta_1 a. \tag{36}$$

For the regression function $Q_{1,0}$, the true marginal structural model is given by

$$\begin{aligned} E[Y_a] &= E_W[Q_{1,0}(a,W)] \\ &= E_W[-1 + a + W_1 - W_2 + 2aW_1 + W_3W_4] \\ &= -0.75 + 2a. \end{aligned} \tag{37}$$

The regression function $Q_{2,0}$ corresponds to the true marginal structural model

$$\begin{aligned} E[Y_a] &= E_W[Q_{2,0}(a,W)] \\ &= E_W[-1 + a + 5W_1 - W_2 + 2aW_1 + W_3W_4] \\ &= 1.25 + 2a. \end{aligned} \tag{38}$$

The true value of the main parameter of interest, β_1, is thus given by $\beta_{1,0} = 2$ in both cases.

3.2 Relying on a parametric model for Q

We first consider an estimator $\beta_n(g_n, Q_n)$ that is based on a correctly specified logistic regression model for the treatment mechanism and a correctly specified linear regression model for Q. We compare the performance of this estimator to that of the non-truncated IPTW estimator as well as that of the simpler truncated IPTW estimators described in the introduction that always truncate weights at 10 or 20, or, alternatively, at a point corresponding to 10 or 20% of the sample size. Table 1 summarizes the relative efficiencies of $\beta_n(g_n, Q_n)$ as compared to the other candidate IPTW estimators for three different sample sizes and the four different data-generating distributions $(g_{1,0}, Q_{1,0})$, $(g_{2,0}, Q_{1,0})$, $(g_{1,0}, Q_{2,0})$, and $(g_{2,0}, Q_{2,0})$. Column 1 of the table shows that $\beta_n(g_n, Q_n)$ can achieve considerable gains in efficiency relative to the non-truncated estimator, with particularly strong gains, ranging from 19 to 47%, for data-generating distributions under which the ETA assumption is practically violated. Considering the four simpler truncated estimators, we note that no single estimator consistently outperforms the other three for all data-generating distributions and sample sizes considered here, underscoring the limitations of truncation schemes that do not respond to any factors beyond sample size. The estimator $\beta_n(g_n, Q_n)$, on the other hand, is seen to consistently achieve smaller mean squared errors than all four reference truncation schemes, with typical gains in efficiency in the range of 5 to 20%.

We next examine the behavior of $\beta_n(g_n, Q_n)$ if the model for g is correctly specified, but the model for Q is mis-specified. The data are generated according to the distribution $(g_{2,0}, Q_{2,0})$. We consider the following four different mis-specified models for Q:

$$Q(A,W) = \gamma_0 + \gamma_1 A + \gamma_2 W_1 + \gamma_3 W_2 + \gamma_4 A W_1 \tag{39}$$

$$Q(A,W) = \gamma_0 + \gamma_1 A + \gamma_2 W_1 + \gamma_3 W_2 \tag{40}$$

$$Q(A,W) = \gamma_0 + \gamma_1 A + \gamma_2 W_1 \tag{41}$$

$$Q(A,W) = \gamma_0 + \gamma_1 A + \gamma_2 W_3 \tag{42}$$

Recall that the correct model would be given by $Q(A,W) = \gamma_0 + \gamma_1 A + \gamma_2 W_1 + \gamma_3 W_2 + \gamma_4 A W_1 + \gamma_5 W_3 W_4$. Table 2 summarizes the observed relative efficiencies of $\beta_n(g_n, Q_n)$ as compared to the other candidate IPTW

Table 1: Finite-sample performance of $\beta_n(g_n, Q_n)$ if Q estimated consistently. *This table summarizes the relative efficiencies of $\beta_n(g_n, Q_n)$ as compared to the non-truncated IPTW estimator as well as a number of simpler truncated IPTW estimators for three different sample sizes and the four different data-generating distributions $(g_{1,0}, Q_{1,0})$, $(g_{2,0}, Q_{1,0})$, $(g_{1,0}, Q_{2,0})$, and $(g_{2,0}, Q_{2,0})$.*

	$M = \infty$	$M = 10.0$	$M = 20.0$	$M = 0.1n$	$M = 0.2n$
$\mathbf{g_{1,0}, Q_{1,0}}$					
n=100	1.13	1.00	1.06	1.00	1.06
n=500	1.11	1.04	1.10	1.11	1.11
n=2500	1.05	1.06	1.05	1.05	1.05
$\mathbf{g_{2,0}, Q_{1,0}}$					
n=100	1.37	1.00	1.00	1.00	1.00
n=500	1.39	1.72	1.09	1.15	1.31
n=2500	1.19	5.71	2.06	1.19	1.19
$\mathbf{g_{1,0}, Q_{2,0}}$					
n=100	1.23	1.04	1.10	1.04	1.10
n=500	1.17	1.11	1.15	1.17	1.17
n=2500	1.08	1.22	1.08	1.08	1.08
$\mathbf{g_{2,0}, Q_{2,0}}$					
n=100	1.47	1.08	1.01	1.08	1.01
n=500	1.48	2.19	1.20	1.15	1.37
n=2500	1.25	8.04	2.78	1.25	1.25

estimators. Rows 1 through 6 of the table show that omission of the interaction terms AW_1 and W_3W_4 from the model for $Q(A, W)$ appears to have only a minor impact on the performance of the estimator, with $\beta_n(g_n, Q_n)$ based on model (39) or (40) still achieving smaller mean squared errors than any of the other estimators. If the model for $Q(A, W)$ in addition fails to include the important confounding factor W_2, $\beta_n(g_n, Q_n)$ performs favorably at large sample sizes, but tends to be outperformed at smaller sample sizes. Omission of both major confounding factors W_1 and W_2 is seen to severely compromise the performance of the estimator, with all other candidate estimators now consistently achieving smaller mean squared errors than $\beta_n(g_n, Q_n)$. These simulation results suggest that $\beta_n(g_n, Q_n)$ behaves quite well as long as the model for Q is reasonably well specified, but that gross mis-specification of that model can have a serious impact on its finite-sample performance.

3.3 Modelling Q based on the propensity score

In this section, we examine the finite-sample performance of an estimator $\beta_n(g_n)$ that avoids relying on a parametric model for Q by making use of the estimated propensity score. As before, the logistic regression model for the treatment mechanism is correctly specified. Table 3 summarizes the relative efficiencies of $\beta_n(g_n)$ as compared to the other candidate IPTW estimators for three different sample sizes and the four different data-generating distributions $(g_{1,0}, Q_{1,0})$, $(g_{2,0}, Q_{1,0})$, $(g_{1,0}, Q_{2,0})$, and $(g_{2,0}, Q_{2,0})$. The results show that, while $\beta_n(g_n)$ tends to be slightly less efficient than $\beta_n(g_n, Q_n)$ based on a correctly specified model for Q, the estimator still performs favorably as compared to the non-truncated IPTW estimators as well as the simpler truncation schemes. Under data-generating distributions that practically violate the ETA assumption, $\beta_n(g_n, Q_n)$ achieves gains in efficiency relative to the non-truncated estimator ranging from 5 to 46%. For small sample sizes, some of the reference truncation schemes occasionally achieve slightly smaller mean squared errors than does $\beta_n(g_n)$, but for sufficiently large sample sizes, the latter estimator consistently outperforms all of them, with typical gains in efficiency in the range of 1 to 15%. Since the simulation studies discussed above illustrate that mis-specification of the model for $Q(A, W)$ can have a serious impact on the

Table 2: Finite-sample performance of $\beta_n(g_n, Q_n)$ if Q estimated inconsistently. *This table summarizes the relative efficiencies of $\beta_n(g_n, Q_n)$ as compared to the non-truncated IPTW estimator as well as a number of simpler truncated IPTW estimators for three different sample sizes and four different mis-specified models for Q. The correct model for Q is given by $Y \sim A + W_1 + W_2 + AW_1 + W_3W_4$*

	$M = \infty$	$M = 10.0$	$M = 20.0$	$M = 0.1n$	$M = 0.2n$
$\mathbf{Y \sim A + W_1 + W_2 + AW_1}$					
n=100	1.48	1.08	1.01	1.08	1.01
n=500	1.48	2.17	1.19	1.15	1.36
n=2500	1.24	8.01	2.77	1.24	1.24
$\mathbf{Y \sim A + W_1 + W_2}$					
n=100	1.44	1.06	0.99	1.06	0.99
n=500	1.41	2.08	1.14	1.10	1.30
n=2500	1.21	7.81	2.70	1.21	1.21
$\mathbf{Y \sim A + W_1}$					
n=100	1.36	0.99	0.93	0.99	0.93
n=500	1.17	1.72	0.94	0.91	1.07
n=2500	1.11	7.16	2.48	1.11	1.11
$\mathbf{Y \sim A + W_3}$					
n=100	0.86	0.63	0.59	0.63	0.59
n=500	0.24	0.35	0.19	0.19	0.22
n=2500	0.05	0.31	0.11	0.05	0.05

performance of the estimator $\beta_n(g_n, Q_n)$, we recommend that it may be preferable in practice to settle for the slightly smaller gains in efficiency afforded by the estimator $\beta_n(g_n)$ whose finite-sample performance does not depend on a correctly specified model for an additional nuisance parameter.

3.4 Comparison of data-adaptively truncated estimators

The previous sections have illustrated that $\beta_n(g_n, Q_n)$ and $\beta_n(g_n)$ can achieve considerable gains in efficiency relative to the non-truncated IPTW estimator as well as to the simple truncation schemes currently in use. In this section, we compare these gains in efficiency to those achieved by two benchmark estimators that $\beta_n(g_n, Q_n)$ and $\beta_n(g_n)$ might be hoped to approximate.

The first of these two estimators, denoted by $\beta_n^{\#}(g_n, Q_n)$, selects the truncation constant by minimizing an estimate of the mean squared error based on the parametric bootstrap approach described by Wang et al. (2006). As described previously, the closed-form bias estimate developed here ignores the finite-sample bias that an IPTW estimator may be subject to if the ETA assumption is practically violated; a bias estimate based on the parametric bootstrap, on the other hand, would capture this additional source of bias. Similarly, a bootstrap-based variance estimate does not have to rely on a conservative approximation that ignores the reduction in variability that may be achieved by estimating the treatment mechanism. It is therefore of interest to evaluate to what extent a data-adaptively truncated estimator based on the proposal by Wang et al. may offer additional improvements in performance beyond those afforded by $\beta_n(g_n, Q_n)$ and $\beta_n(g_n)$. Due to the increased computational complexity of this approach, we were only able to consider an estimator that employs a fairly small number of 25 bootstrap samples for the purpose of estimating the mean squared error of the candidate truncated estimators $\beta_{M,n}(g_n)$. The second benchmark estimator, denoted by β_n^0, is based on the true mean squared error for each of the candidate estimators $\beta_{M,n}(g_n)$ and is thus infeasible in practice. For each data-generating distribution, it simply selects the truncation level that leads to the smallest mean squared error.

Figure 1 shows the true mean squared error for the set of candidate estimators $\beta_{M,n}(g_n)$ along with

Table 3: Finite-sample performance of $\beta_n(g_n)$. *This table summarizes the relative efficiencies of* $\beta_n(g_n)$ *as compared to the non-truncated IPTW estimator as well as a number of simpler truncated IPTW estimators for three different sample sizes and and the four different data-generating distributions* $(g_{1,0}, Q_{1,0})$, $(g_{2,0}, Q_{1,0})$, $(g_{1,0}, Q_{2,0})$, *and* $(g_{2,0}, Q_{2,0})$.

	$M = \infty$	$M = 10.0$	$M = 20.0$	$M = 0.1n$	$M = 0.2n$
$\mathbf{g_{1,0}, Q_{1,0}}$					
n=100	1.07	0.95	1.01	0.95	1.01
n=500	1.07	1.01	1.06	1.07	1.07
n=2500	1.01	1.03	1.01	1.01	1.01
$\mathbf{g_{2,0}, Q_{1,0}}$					
n=100	1.33	0.97	0.97	0.97	0.97
n=500	1.28	1.59	1.01	1.06	1.21
n=2500	1.05	5.05	1.82	1.05	1.05
$\mathbf{g_{1,0}, Q_{2,0}}$					
n=100	1.18	1.00	1.05	1.00	1.05
n=500	1.13	1.08	1.12	1.13	1.13
n=2500	1.06	1.19	1.06	1.06	1.06
$\mathbf{g_{2,0}, Q_{2,0}}$					
n=100	1.46	1.07	1.00	1.07	1.00
n=500	1.42	2.09	1.15	1.11	1.31
n=2500	1.21	7.79	2.70	1.21	1.21

the mean squared errors achieved by $\beta_n(g_n)$, $\beta_n(g_n, Q_n)$, and $\beta_n^{\#}(g_n, Q_n)$ where the latter two estimators employ a correctly specified model for the additional nuisance parameter $Q(A, W)$. Table 4 summarizes the relative efficiencies of the four data-adaptively truncated estimators as compared to the non-truncated IPTW estimator. These results show that $\beta_n(g_n, Q_n)$ and $\beta_n(g_n)$ typically perform on par with β_n^0, with $\beta_n(g_n, Q_n)$ in fact tending to achieve slightly higher gains in efficiency relative to the non-truncated estimator. In addition, we note that use of the parametric bootstrap approach developed by Wang et al. does not appear to provide significant improvements in performance over the closed-form approximation introduced here, at least at the fairly small number of 25 bootstrap samples. The performance of the former estimator can likely be improved by using a larger number of bootstrap samples, but as mentioned earlier this will make the estimator too computationally intensive for a number of applications.

4 Data analysis

In this section, we illustrate the data-adaptively truncated IPTW estimator $\beta_n(g_n)$ in an applied data analysis aimed at estimating the causal effect of vigorous leisure-time physical activity (LTPA) on all-cause mortality in the elderly. The data we analyze were collected as part of a community-based longitudinal study of physical activity and fitness (Study of Physical Performance and Age Related Changes in Sonomans - SPPARCS), in which Tager et al. (1998) followed a group of people aged 55 years and older living in and around Sonoma, CA, over a time period of about ten years.

Our measure of vigorous LTPA is defined based on a questionnaire in which participants were asked how many hours during the past seven days they had participated in twelve common vigorous physical activities such as jogging, swimming, bicycling on hills, or racquetball. Activities were assigned standard intensity values in metabolic equivalents (METs) (Ainsworth et al., 1993); one MET approximately equals the oxygen consumption required for sitting quietly. A continuous summary score was obtained by multiplying these intensity values by the number of hours engaged in the various activities and summing up over all activities

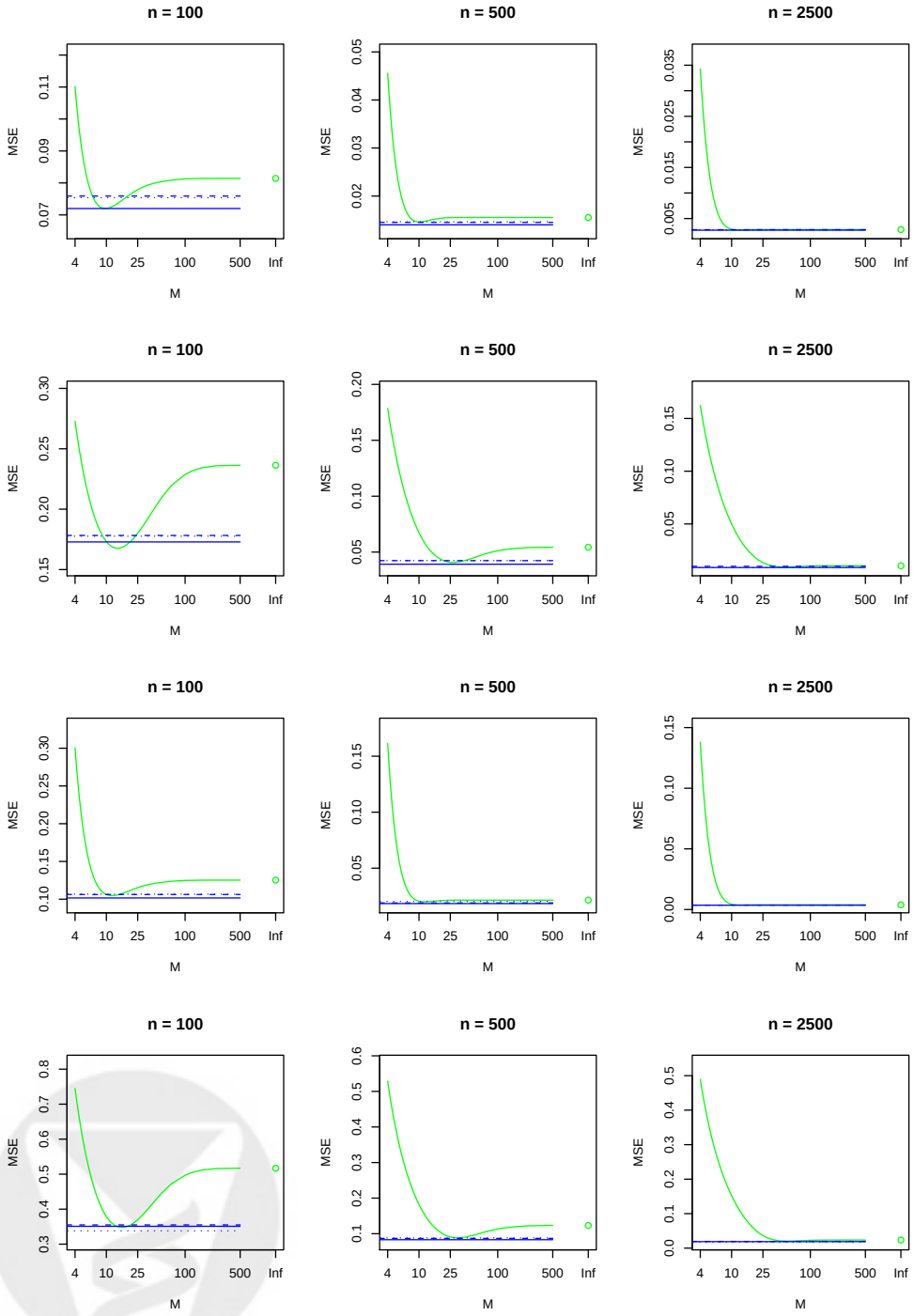

Figure 1: Mean squared error comparison. *These plots show the mean squared error for β_1 for truncated estimators $\beta_{M,n}(g_n)$ using a fixed value of M (solid green line), $\beta_n(g_n, Q_n)$ (solid blue line), $\beta_n(g_n)$ (dashed blue line), and $\beta_n^{\#}(g_n, Q_n)$ (dotted blue line). The four rows of plots correspond to the four different data-generating distributions $(g_{1,0}, Q_{1,0})$, $(g_{2,0}, Q_{1,0})$,$(g_{1,0}, Q_{2,0})$, and $(g_{2,0}, Q_{2,0})$, respectively.*

Table 4: Comparison of candidate data-adaptively truncated estimators. *This table summarizes the relative efficiencies of the two data-adaptively truncated estimators proposed here along with those of two benchmark estimators relative to the non-truncated IPTW estimators. For the estimator β_n^0, the corresponding optimal truncation level is given in parentheses.*

	$\beta_n(g_n)$	$\beta_n(g_n, Q_n)$	$\beta_n^{\#}(g_n, Q_n)$	β_n^0
$\mathbf{g_{1,0}}, \mathbf{Q_{1,0}}$				
n=100	1.07	1.13	1.08	1.13 (M=10)
n=500	1.07	1.11	1.06	1.06 (M=10)
n=2500	1.01	1.05	1.02	1.03 (M=12)
$\mathbf{g_{2,0}}, \mathbf{Q_{1,0}}$				
n=100	1.33	1.37	1.33	1.41 (M=14)
n=500	1.28	1.39	1.28	1.33 (M=29)
n=2500	1.05	1.19	1.21	1.13 (M=50)
$\mathbf{g_{1,0}}, \mathbf{Q_{2,0}}$				
n=100	1.18	1.23	1.17	1.20 (M=12)
n=500	1.13	1.17	1.06	1.07 (M=11)
n=2500	1.06	1.08	1.06	1.04 (M=12)
$\mathbf{g_{2,0}}, \mathbf{Q_{2,0}}$				
n=100	1.46	1.47	1.53	1.49 (M=17)
n=500	1.42	1.48	1.40	1.39 (M=33)
n=2500	1.21	1.25	1.33	1.17 (M=50)

considered here. The CDC currently recommends that elderly people engage in physical activity for 30 minutes at least five times a week, corresponding to an energy expenditure of 22.5 METs (CDC, 1996). The treatment variable A was therefore defined as the following dichotomous version of our summary LTPA score:

$$A = \begin{cases} 0 & \text{if } LTPA < 22.5 \text{ METs} \\ 1 & \text{if } LTPA \geq 22.5 \text{ METs} \end{cases} \tag{43}$$

Apart from sex and age, the primary confounding factor of the relationship between LTPA and all-cause mortality is likely to be given by a subject's underlying level of general health. Healthier subjects will not only tend to experience lower mortality risks, but are also more likely to engage in higher levels of vigorous physical activity. To control for this source of confounding, our analysis adjusts for a number of covariates that are intended to capture a subject's underlying level of health. Participants were asked, for instance, to rate their health as excellent, good, fair, or poor. Self-reported physical functioning was defined from a series of questions, originally developed by Nagi (1976) and Rosow and Breslau (1966), that assessed the degree of difficulty a participant experienced in various activities of daily living. On the basis of this questionnaire, we classified a participant's level of physical functioning as excellent, moderately impaired, or severely impaired. In addition, participants were asked about the previous occurrence of cardiac events such as myocardial infarctions, the presence of a number of chronic health conditions, their smoking status, as well as a possible decline in physical activity compared to five or ten years earlier. Table 5 summarizes the definition of the covariates W we adjust for as potential confounding factors.

The outcome of interest Y was defined as an indicator for death within five years of the baseline interview. We note that Y was observed for all study participants so that we do not have to adjust for right censoring. Of the 2092 participants enrolled in the SPPARCS study, 15 did not answer all the questions needed to define their level of vigorous physical activity; an additional 26 were missing information about at least one of the confounding factors described above. Our analysis is based on the remaining 2051 participants.

Table 5: Definition of indicator variables that are considered as potential confounders.

Variable	Definition
$FEMALE$	Female
$AGE.1$	≤ 60 years old
$AGE.2$	60-70 years old
$AGE.4$	80-90 years old
$AGE.5$	90-100 years old
$HTL.EX$	Excellent self-rated health
$HLT.FAIR$	Fair self-rated health
$HLT.POOR$	Poor self-rated health
$NRB.FAIR$	Moderately impaired physical functioning ($0.5 \leq$ NRB score < 1.0)
$NRB.POOR$	Severely impaired physical functioning (NRB score < 0.5)
$CARD$	Previous occurrence of any of the following cardiac events: Angina, myocardial infarction, congestive heart failure, coronary by-pass surgery, and coronary angioplasty
$CHRON$	Presence of any of the following chronic health conditions: stroke, cancer, liver disease, kidney disease, Parkinson's disease, and diabetes mellitus
$SMK.CURR$	Current smoker
$SMK.EX$	Former smoker
$DECLINE$	Activity decline compared to five or ten years earlier

The parameter of interest is defined based on the logistic marginal structural model

$$\text{logit}\Big(P(Y_a)\Big) = \beta_0 + \beta_1 a, \tag{44}$$

with β_1 identifying the causal log odds ratio for mortality comparing $a = 1$ to $a = 0$. The treatment mechanism was estimated by a logistic regression model that included main-effect terms for all indicator variables defined in table 5. According to the fit we obtained, summarized in table 6, subjects were more likely to engage in a higher level of physical activity if they were under the age of 70 or rated their own health as excellent. Likewise, subjects were estimated to be less likely to engage in higher levels of physical activity if they rated their own health as fair or poor, suffered from moderate or severe functional impairment, were female or currently smoking, or reported a decline in physical activity over the past five to ten years. We evaluated the goodness-of-fit of this model using the Hosmer-Le Cessie test introduced by Hosmer et al. (1997) as an improvement of the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980). This test yielded a p-value of 0.75, providing little evidence against the assumption that this model adequately describes the data. The non-truncated weights $wt(A, W)$ obtained on the basis of this estimated treatment mechanism range up to 33.

Figure 2 summarizes the IPTW estimates of β_1 for the different candidate truncated estimators $\beta_{M,n}(g_n)$ as well as the data-adaptively truncated estimator $\beta_n(g_n)$. The non-truncated IPTW estimator yields an odds-ratio estimate of 0.72 (95% CI: 0.43 to 1.06). The data-adaptively truncated estimator, selecting a truncation level of $M = 20$, yields an odds-ratio estimate of 0.67 (95% CI: 0.40 to 0.94). The selected truncation level is estimated to lead to a 7% increase in efficiency relative to the non-truncated IPTW estimator. An IPTW estimator based on the fixed truncation level $M = 20$ yields an odd-ratio estimate of 0.67 (95% CI: 0.41 to 0.93), showing that the data-adaptively truncated estimator is only slightly more variable than this reference estimator.

Table 6: Logistic regression fit for the treatment mechanism.

	OR	95% CI	p-value
AGE.1	1.14	(0.84, 1.56)	0.3929
AGE.2	1.17	(0.92, 1.50)	0.1976
AGE.4	1.01	(0.68, 1.49)	0.9751
AGE.5	0.32	(0.04, 2.50)	0.2771
HLT.EX	1.37	(1.09, 1.71)	0.0059
HLT.FAIR	0.65	(0.45, 0.94)	0.0232
HLT.POOR	0.30	(0.09, 1.01)	0.0522
NRB.POOR	0.32	(0.18, 0.57)	<10e-4
NRB.FAIR	0.80	(0.64, 1.01)	0.0660
SMOKE.CURR	0.53	(0.34, 0.83)	0.0058
SMOKE.EX	1.08	(0.87, 1.34)	0.4937
CARD	1.05	(0.78, 1.41)	0.7657
CHRONIC	1.02	(0.82, 1.25)	0.8827
FEMALE	0.81	(0.65, 1.01)	0.0590
DECLINE	0.59	(0.46, 0.75)	<10e-4

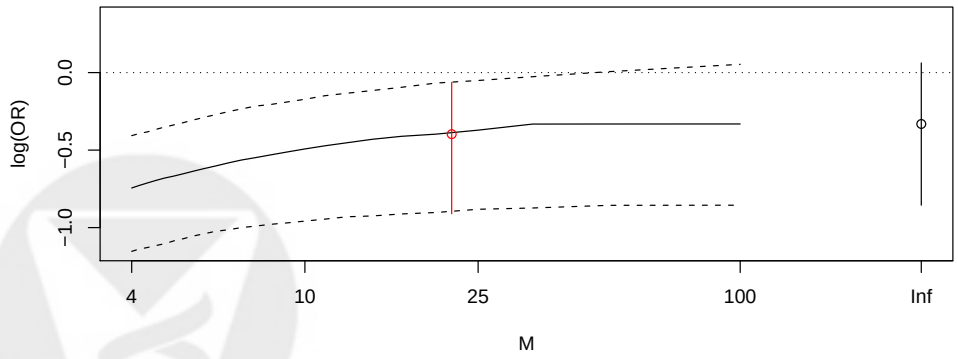

Figure 2: Log(OR) estimates. *The solid line shows estimates of β_1 as obtained by $\beta_{M,n}(g_n)$ for different truncation levels M along with pointwise 95% bootstrap confidence intervals (dashed lines). The estimate obtained by $\beta_n(g_n)$, shown in red, is based on a truncation level of 20.*

5 Discussion

In this article, we develop an approach for addressing an important open problem in the popular methodology of IPTW estimation, namely the high variability of such estimators in situations in which the ETA assumption is practically violated. We introduce an estimator that data-adaptively selects an appropriate truncation constant for the Inverse-Probability-of-Treatment weights based on the goal of minimizing the mean squared error of the estimator. While the resulting estimator requires an estimate of an additional nuisance parameter, we show that its consistency does not rely on a consistent estimate of that nuisance parameter. For the case of a binary treatment variable A, we describe an approach for obtaining an estimate of this nuisance parameter that makes use of the estimated propensity score and does not require the user to specify an appropriate parametric model. The simulation studies we present demonstrate that the methodology developed here can lead to considerable gains in efficiency over more *ad-hoc* truncation approaches currently in use. In fact, the proposed data-adaptively truncated estimator is seen to perform on par with an infeasible benchmark estimator that relies on knowledge of the true data-generating distribution. In an applied data analysis, the estimator is estimated to achieve a 7% gain in efficiency relative to the non-truncated estimator, with truncation resulting in a non-significant finding becoming statistically significant.

The marginal structural model used in our simulation studies is very simple and was selected deliberately in order to allow us to focus on a single one-dimensional parameter of interest. The methodology could, however, be extended to more complex marginal structural models that stratify on a subset V of the baseline covariates and also include interaction terms between a and V. In such cases, the mean squared error to be minimized could be defined more generally as a user-supplied weighted average of the mean squared errors of the individual coefficients of the model.

While this article focuses primarily on the estimation of marginal structural model parameters in the context of a point-treatment study, the approach can be extended in a straightforward way to a number of other parameters of interest. One particular class of parameters we would like to highlight consists of the variable importance measures described by van der Laan (2006) that are designed to capture the impact of an input variable on an outcome of interest. Such parameters have wide applications in contemporary problems in computational biology that investigate a large number of candidate input variables in the hope of identifying a subset for which there is strong evidence of an impact on an outcome of interest. This problem arises frequently, for instance, in the area of biomarker discovery. In such high-dimensional problems, the simulation-based approach developed by Wang et al. (2006) becomes computationally infeasible so that approximate analytic methods like the one introduced here become even more attractive.

Inverse-Probability-of-Censoring-Weighted (IPCW) estimators represent another area to which the approach presented here can be applied. Such estimators have become a popular tool in survival analysis and run into similar problems as IPTW estimators when censoring probabilities are estimated to be close to zero. In addition, it can be hoped that a similar approach can be developed for IPTW estimators in the longitudinal setting in which treatments are assigned at multiple time points. In this context, practical violations of the ETA assumption are even more common than in the point-treatment setting since the required weights consist of a product of time-specific weights that can therefore more easily become very large for some observations.

Another possible application of the approach described here lies in the selection of the set of confounding factors that are to be included in the model for the treatment mechanism. A data set may, for example, contain a covariate that is a very strong predictor of the treatment variable, but only a weak predictor of the outcome of interest. Such a covariate will often be only a weak confounder of the relationship between treatment and outcome, but can cause a serious practical ETA violation. Omitting it from the model for the treatment model could thus, at the price of a slight increase in bias, offer a considerable reduction in variability, thus leading to an overall reduction in mean squared error.

We propose to base inference for the data-adaptively truncated IPTW estimator on the bootstrap. Alternatively, one might attempt to use the conservative influence curve (25) for the IPTW estimator with g known for this purpose, in the hope that the downward bias caused by ignoring the additional variability due to the data-adaptive selection of the truncation constant would be offset to some extent by the upward bias caused by ignoring the reduction in variability achieved by estimating the treatment mechanism. Fu-

ture research will be needed to investigate the performance of this approach which would further reduce the computational demands of the estimator.

The data-adaptively truncated IPTW estimator developed here has been implemented in an R package called `tIPTW` that can be downloaded at `http://www.stat.berkeley.edu/~laan/Software/`. Currently, the package supports the data-adaptive selection of a truncation constant for IPTW estimators in linear as well as logistic marginal structural models. Future work will be dedicated to adding some of the possible extensions discussed in this section.

6 Acknowledgements

We would like to thank Dr. Ira Tager from the Division of Epidemiology at the UC Berkeley School of Public Health for kindly making available the dataset that was used in our data analysis. His work on the SPPARCS project was supported by a grant from the National Institute on Aging (RO1-AG09389).

References

Ainsworth, B., Haskell, W., Leon, A., Jacobs, D. J., Montoye, H., Sallis, J., and Paffenberger, Jr., R. (1993). Compendium of physical activities: classification of energy costs of human physical activities. *Medicine and Science in Sports and Exercise*, 25:71–80.

Bickel, P., Klaasen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models*. The Johns Hopkins University Press.

CDC (1996). Physical activity and health: a report of the surgeon general. Altanta, Georgia: US Department of Health and Human Services, CDC.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Hosmer, D. W., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness–of–fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Nagi, S. (1976). An epidemiology of disability among adults in the United States. *Milbank Quartely*, 54:439–468.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference*, 129:405–426.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–480 (1990).

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy survivor effect. *Mathematical Modelling*, 7:1393–1512.

Robins, J. (1987). Addendum to "a new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy survivor effect" [Math. Modelling **7** (1986) 1393-1512]. *Computers and Mathematics with Applications*, 14:923–945.

Robins, J. (1999). Marginal structural models versus structural nested models as tools for causal inference. In Halloran, M. and Berry, D., editors, *Statistical Models in Epidemiology: The Environment and Clinical Trials*, volume 116, pages 95–134. Springer Verlag.

Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association 1999*, pages 6–10.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

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

Rosow, I. and Breslau, N. (1966). A Guttman health scale for the aged. *Journal of Gerontology*, 21:556–559.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

Tager, I., Hollenberg, M., and Satariano, W. (1998). Self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population. *American Journal of Epidemiology*, 147:921–931.

van der Laan, M. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2.

van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer Verlag.

Wang, Y., Petersen, M., Bangsberg, D., and van der Laan, M. (2006). Diagnosing Bias in the Inverse-Probability-of-Treatment-Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. Technical Report 211, UC Berkeley Division of Biostatistics Working Paper Series.