Inference For The Smoothed Proportion Whose Average Treatment Effect Exceeds a Threshold

Jonathan Levy and Mark van der Laan
University of California, Berkeley
School of Public Health
2121 Berkeley Way, Room 5302
Berkeley, CA 94720-7360
jl@jlstiles.com

March 15, 2019

Abstract

The strata-specific treatment effect function (TE function) defines a random variable giving the average treatment effect for a randomly drawn strata and thus, has a corresponding cumulative distribution function (CDF). The CDF is of interest because it gives the analyst a percentage of the population whose average effect exceeds a given threshold and therefore provides a measure of what segment of the population will benefit (or receive a deleterious effect) beyond a certain level. However, the CDF is not pathwise differentiable, so we will estimate it with a family of pathwise differentiable kernel smoothed parameter mappings as a strategy to provide inference. We present a cross-validated targeted maximum likelihood estimator (CV-TMLE), which assumes the TE CDF is continuous. We show that the use of highly adaptive lasso (HAL) to fit the outcome model and, in the case of an observational study, the treatment mechanism, guarantees asymptotic efficiency under the condition that the models are of finite sectional variation norm and are continuous from the right with left hand limits. We also provide conditions under which we can let the bandwidth approach 0 and guarantee a normal limiting distribution if using HAL. Through simulations we verify theoretical properties of the estimator and show the importance of machine learning over conventional regression approaches to fitting the nuisance parameters. We offer a bandwidth selector that points the way for future developments to minimize MSE in estimating the TE CDF.

Background and Motivation

The stratum-specific treatment effect or TE function is defined as a random variable given by the average treatment effect for a randomly drawn stratum of baseline covariates. Estimating the cumulative distribution function or CDF of the TE function estimates the proportion of the population that has an average treatment effect at or below a given level. Because clinicians treat patients based on patient characteristics and evaluating public health policy depends on assessing whether a policy has a large positive or possibly negative effect on segments of the population, such is of interest in accounting for heterogeneous effects beyond a basic population average effect. This paper is also a continuance of previous work estimating the variance of TE function (Levy et al. 2018).

Much consideration has been given to the distribution of $Y_1 - Y_0$, where Y_a is the counterfactual outcome under the intervention to set treatment to a value of $a \in \{0, 1\}$, as per the neyman-rubin potential outcomes framework (Neyman 1923), (Rubin 1974). Neyman, 1923, realized, in estimating the mean of $Y_1 - Y_0$, that the standard errors for small samples depended estimating the correlation of Y_1 and Y_0 for which he had not the data. Assumptions needed to estimate the joint distribution of Y_1 and Y_0 are hard to verify. Fisher, 1951 suggests one can essentially create the counterfactual $Y_1 - Y_0$ by careful design. Heckman and Smith, 1998 estimate the quantiles of $Y_1 - Y_0$ via the assumption of quantiles being preserved from Y_1 to Y_0 given a strata of confounders. Without strong assumptions, using tail bounds (Frechet 1951) to estimate the quantiles of $Y_1 - Y_0$ via the marginals of Y_1 and Y_0 tends to leave too big of a measure of uncertainty to be useful (Heckman and Smith 1997). Cox, 1958, assumes $var(Y_1 - Y_0)$ = 0 for predefined homogeneous subgroups, essentially assuming the distribution of $Y_1 - Y_0$ is the same as the distribution of $E[Y_1 - Y_0 \mid W]$ for a finite set of W. The CDF of $E[Y_1 - Y_0 \mid W]$ is what we aim to estimate.

Data

Our full data, including unobserved measures, is assumed to be generated according to the structural equations (Wright 1921), (Strotz and Wold 1960), (Pearl 2000) below. We first assume a joint distribution, $U = (U_W, U_A, U_Y) \sim P_U$, an unknown distribution of unmeasured variables. $O = (W, A, Y)$ are the measured variables. In the time ordering of occurrence we have $W = f_W(U_W)$ where W is a vector of confounders, $A = f_A(U_A, W)$, where A is a binary treatment and $Y = f_Y(U_Y, W, A)$, where Y is the outcome, either binary or bounded continuous. We thusly define a distribution $P_{U,O}$, via $(U, O) \sim P_{U,O}$.

The full-data model, $\mathcal{M}^F$, consists of all possible $P_{U,O}$ which, may be non-parametric or include knowledge about the data generating system, such as the treatment mechanism, as in a randomized trial. The observed data model, $\mathcal{M}$, is linked to $\mathcal{M}^F$ in that we observe $O = (W, A, Y) \sim P$ where $O = (W, A, Y)$ is generated by P_{UO} according to the structural equations above. The observed data distribution, P, is therefore an element of the observed data model, $\mathcal{M}$. We will also assume throughout that we will observe n independent identically distributed draws from the true distribution, P_0.

Parameter of interest and identification

First we define the potential outcome under the intervention to set the treatment to a value a to 0 or 1 as (Neyman 1923) $Y_a = f_Y(U_Y, a, W)$. The TE function with respect to $P_{U,O} \in \mathcal{M}^F$ is then defined as $b_{P_{U,O}}(W) = \mathbb{E}_{P_{U,O}}[Y_1|W] - \mathbb{E}_{P_{U,O}}[Y_0|W]$. Our parameter of interest is a mapping from $\mathcal{M}^F$ to R^d defined by $\Psi^F(P_{U,O}) = (\Psi^F_{t_1}(P_{U,O}), ..., \Psi^F_{t_d}(P_{U,O}))$, where $\Psi^F_{t_i}(P_{U,O}) = \mathbb{E}_{P_{U,O}}\mathbb{I}(b_{P_{UX}}(W) \leq t_i)$. We will assume $Y_a \perp A|W$, i.e. the randomization assumption, on $\mathcal{M}^F$ (Robins 1986), (Greenland and Robins 1986) as well as the positivity assumption, which states for all a and W, $0 < E_P[A = a \mid W] < 1$. We thus have that $b_P(W) = \mathbb{E}_P[Y|A = 1, W] - \mathbb{E}_P[Y|A = 0, W]$, which yields $b_{P_{U,O}}(W) = b_P(W)$. We can then identify the parameter of interest as a mapping from the observed data model, $\mathcal{M}$, to $\mathbb{R}^d$ via $\Psi(P) = (\Psi_{t_1}(P), ..., \Psi_{t_d}(P))$, where $\Psi_{t_i}(P) = \mathbb{E}_P\mathbb{I}(b_P(W) \leq t_i)$.

Ψ is not pathwise differentiable (van der Vaart 2000) so instead we consider the smoothed version of the parameter mapping, using kernel, k, with bandwidth, δ, which is pathwise differentiable and hence, provides a strategy for providing inference for the smoothed TE CDF as well as the TE CDF itself. Here we will suppress k in the notation for convenience and define the i^{th} component of the $d - dimensional$ parameter mapping as $\Psi_{\delta,t_i}(P) = \mathbb{E}_W \int_x \frac{1}{\delta} k\left(\frac{x-t_i}{\delta}\right) \mathbb{I}(b(W) \leq x)dx = \int_x \frac{1}{\delta} k\left(\frac{x-t_i}{\delta}\right) F(x)dx$ so we can write the d-dimensional parameter mapping as $\Psi_\delta(P) = (F_\delta(t_1), ..., F_\delta(t_d))$, where $F_\delta(t_i)$ is a shortened notation for the smoothed CDF, $F(t_i)$.

A Brief Note on Pathwise Differentiability Taken from van der Vaart, 2000: A parameter, Ψ, is pathwise differentiable relative to the tangent space of P, if for every submodel P_t with score function, g, in the tangent space, there exists a continuous linear map from $\dot{\Psi}_P : L^2(P) \rightarrow \mathbb{R}^k$ such that as t vanishes

$$\frac{\Psi(P_t) - \Psi(P)}{t} \longrightarrow \dot{\Psi}_P(g)$$

A classic case of a non pathwise differentiable parameter is the density for a continuous distribution (in the absence of any parametric assumptions) at a point which depends on a set of measure 0. In our case, the pathwise derivative for the TE CDF at a given value, t, does not exist because the indicator function is not differentiable where it jumps at the value of t. Many of the TE values far away from t will not be very helpful in estimating the CDF at t with much precision, so we focus on a bandwidth of TE values around t in a similar manner to a kernel density estimator. As n becomes larger we want to decrease the bandwidth so as to minimize the mean squared error.

The pathwise derivative defined above has a representation as $\int gD^*(P)dP$, where $D^*(P)$ is a unique element of the tangent space called the efficient influence curve or canonical gradient, whose variance is the cramer-rao lower bound (minimum variance possible) for any regular asymptotically linear estimator of the parameter (van der Vaart 2000). Knowing $D^*(P)$ enables the construction of estimators for non-parametric and semi-parametric models, that are asymptotically efficient in that asymptotically their variance attains the cramer-rao lower bound. Examples of such estimators are the one-step estimator and targeted maximum likelihood estimator (TMLE) or its cross-validated counterpart, CV-TMLE (van der Laan and Rubin 2006) (Zheng and van der Laan 2010) (van der Laan and Rose 2011). We prefer the CV-TMLE and TMLE, because they have the advantage of being substitution estimators and, therefore, obey natural parameter bounds which, has been shown to improve stability in finite samples (van der Laan and Rose 2011). For our case, if we plug in a model for many points on the TE CDF, we will be guaranteed that the estimates with be both monotonic and bounded within [0,1], where a non-substitution estimator holds no such guarantees. We prefer CV-TMLE over TMLE because, as we will see, it requires only one condition as opposed to two for TMLE in order to guarantee asymptotic efficiency.

The Cross-Validated Targeted Maximum Likelihood Estimator, CV-TMLE

We will employ the notation, $P_n f$ to be the empirical average of function, $f(\cdot)$, and Pf to be $\mathbb{E}_P f(O)$. Define a loss function, $L(P)(O)$, which is a function of the observed data, O, and indexed at the distribution on which it is defined, P, such that $E_{P_0} L(P)(O)$ is minimized at the true observed data distribution, $P = P_0$. We scale a continuous outcome to be in $[0, 1]$ via the transformation $Y_s = \frac{Y-m}{M-m}$ where m and M are minimum and maximum outcomes respectively, obtained from the data or known a priori. A given distribution, P, in our model defines an outcome model with conditional mean, $\bar{Q}(A, W) = E_P[Y \mid A, W]$, and loss function, $L(P)(w, a, y) = -logp_Y(y \mid a, w)$, where p_Y is the conditional likelihood of Y given A and W.

$$\begin{aligned}&\int L(P)(w, a, y)dP_0 \\ &= \int \left[- \left(ylog(\bar{Q}(a, w)) + (1 - y)log(1 - \bar{Q}(a, w))\right)\right] dP_0\end{aligned} \tag{1}$$

For continuous outcome scaled to be in [0,1], (1) is the mean of the so-called quasibinomial loss, also minimized at the truth (Wedderburn 1974) (McCullagh 1983). So, whether we scale our continuous outcomes to be between 0 and 1 or have a binary outcome we will use the same loss function and thus, the targeting portion of the CV-TMLE or TMLE procedure, performed with logistic regression (see section), will be identical for bounded continuous or binary outcomes. The scaling of continuous outcomes changes nothing of importance because when we evaluate our parameter on the original scale we are smoothing the TE CDF $\mathbb{E}(b(W) \leq t) = \mathbb{E}(b(W)/(M - m) \leq t/(M - m))$, the parameter mapping for scaled outcomes.

Let $D^\star_{\Psi_\delta}(P)(O)$ be the efficient influence curved, which is d-dimensional, defined for our smoothed TE CDF parameter (see (3) below). A targeted maximum likelihood (TML) update (van der Laan and Rubin 2006) is a map of an initial estimate, P_n^0, of the data generating distribution to P_n^* , such that $P_n L(P_n^*) \leq P_n L(P_n^0)$ and $P_n D^*(P_n^*) = o_P(n^{-1/2})$. P_n^* is called the TMLE update of P_n^0 and the TMLE estimate of the parameter of interest is the substitution or plug-in estimator, $\Psi(P_n^*)$. We refer to the initial estimate of the parameter of interest as $\Psi(P_n^0)$. To perform the TML updating procedure we define a 1-dimensional submodel, called a canonical locally least favorable submodel or clfm, (Levy 2018b):

Definition 0.1. A canonical 1-dimensional locally least favorable submodel (clfm) of an estimate, P_n^0, of the true distribution, P_0, is

$$\left\{P_{n,\epsilon}^0 \text{ s.t } \frac{d}{d\epsilon} P_n L(P_{n,\epsilon}^0)\right|_{\epsilon=0} = \|P_n D^\star(P_n^0)\|_2, \epsilon \in [-\delta, \delta]\} \tag{2}$$

where $P_{n,\epsilon}^0 = P_n^0$ and $\| \cdot \|_2$ is the euclidean norm.

The reader may be familiar with the d-dimensional locally least favorable submodel or lfm (van der Laan and Rubin 2006) or universal least favorable submodel (ulfm) which is also 1 dimensional (van der Laan and Gruber 2016), which has the advantage of not relying on an iterative procedure, but here we did not notice an appreciable difference in performance so we used the faster clfm-based procedure. To construct a clfm, one needs to know the efficient influence curve of our parameter of interest (van der Vaart 2000), which is a d-dimensional curve (for each of the d components of the parameter mapping), where the i^{th} component is given by

$$D^{\star}_{\Psi_{\delta,t_i}}(P_0)(O) = \frac{-1}{\delta} k\left(\frac{b_0(W)-t_i}{\delta}\right) * \frac{2A-1}{g_0(A|W)}(Y-\bar{Q}_0(A,W)) + \int \frac{1}{\delta} k\left(\frac{x-t_i}{\delta}\right) \mathbb{I}(b_0(W) \leq x)dx - \Psi_{\delta,t_i}(P_0) \tag{3}$$

where t_i is a given TE value (average treatment effect level), k is the kernel and bandwidth is δ. The reader may find the proof in the appendix of Levy and van der Laan, 2018. From here we will shorten the notation and refer to D^* as the $d-$dimensional efficient influence curve with components $D^{\star}_i = D_{\Psi_{\delta,t_i}}$.

Our initial estimate, P_n^0, of the true distribution, P_0, is defined by $\bar{Q}_n^0(A,W)$, an estimate of the outcome regression, $\bar{Q}_0$, g_n, an estimate of the treatment mechanism, g_0 and $Q_{W,n}$, the empirical distribution, which estimates $Q_{W,0}$, the distribution of W. We denote the empirical density as $q_{W,n}$, which estimates the true density, $q_{W,0}$, of W. We can then define the d-dimensional curve

$$\begin{aligned} H^0(A,W) &= (H_1^0(A,W), H_2^0(A,W), ..., H_d^0(A,W)) \\ &= \frac{1-2A}{\delta g_n(A|W)}\left(k\left(\frac{b_n^0(w)-t_1}{\delta}\right), ..., k\left(\frac{b_n^0(w)-t_d}{\delta}\right)\right) \end{aligned}$$

where $b_n^0(W_i) = \bar{Q}_n^0(1,W) - \bar{Q}_n^0(0,W)$. The initial empirical risk for the outcome model is given by

$$P_n L(\bar{Q}_n^0) = -\frac{1}{n}\sum_{i=1}^{n}\left[Y_i \text{log}\bar{Q}_n^0(A_i,W_i) + (1-Y_i)\text{log}(1-\bar{Q}_n^0(A_i,W_i))\right]$$

Our efficient influence curve approximation at the initial estimate is given by $D^*(P_n^0)$. Now define the elements of the clfm of initial estimate, P_n^0, by keeping g_n and $q_{W,n}$ fixed and defining $\bar{Q}^0_{n,\epsilon}(A,W)$ as

$$expit\left(logit(\bar{Q}_n^0(A,W)) + \epsilon\left\langle H^0(A,W), \frac{P_n D^*(P_n^0)}{\|P_n D^*(P_n^0)\|_2}\right\rangle_2\right)$$

We can then verify this satisfies the definition of a clfm above. This then gives rise to the iterative procedure detailed below in steps 1 through 4 below.

TML Algorithm

step 1: Obtaining Initial Estimates To perform a TMLE we use an ensemble learning package such as sl3 (Coyle et al. 2018a) or SuperLearner (Polley et al. 2017) to construct the initial fit, $\bar{Q}_n^0$, of outcome model $E_P[Y \mid A, W]$, and the initial fit, g_n, of the treatment mechanism, $E_P[A \mid W]$, thus providing the estimates $\bar{Q}_n^0(A_i,W_i)$ and $g_n(A_i \mid W_i)$, $i \in 1:n$, i.e., for all n subjects.

step 2:
Starting with $m = 0$:

If $|P_n D_j^*(P_n^m)(O)| < \frac{\hat{\sigma}(D_j^*(P_n^m)(O))}{log(n)n^{1/2}}$ for all j then $P_n^{\star} = P_n^m$ and go to step 4. Otherwise go to step 3. $\hat{\sigma}(\cdot)$ refers to the sample standard deviation of values taken over the data. To provide some clarity: If $n = 1000$ then $log(n) \approx 7$, so the above stopping criterion ensures any bias is second order at that point. More iterations after this are only time-consuming and do not help with coverage to any appreciable extent.

step 3:
Y as the outcome, offset = $logit(\bar{Q}_n^m)(A,W)$ and so-called clever covariate is computed as

$$\left\langle (H^{m-1}(A,W), \frac{P_n D^*(P_n^m)}{\|P_n D^*(P_n^m)\|_2}\right\rangle_2$$

where $\langle \cdot, \cdot \rangle_2$ is the dot-product or euclidean inner product. Assume ϵ_m is the coefficient computed from the logistic regression defined by

$$\bar{Q}_n^{m+1}(A,W) = expit\left(logit\left(\bar{Q}_n^m(A,W)\right) + \epsilon_n^m H^m(A,W)\right)$$

We then update the models by the setting $\bar{Q}_n^{m+1}(A,W) =$ $expit\left(logit(\bar{Q}_n^m(A,W)) - \epsilon_n^m\left\langle (H_1(P_n^m)(A,W), \frac{P_n D^*(P_n^m)}{\|P_n D^*(P_n^m)\|_2}\right\rangle_2\right)$ set $m = m + 1$ and return to step 2.

step 4:
The TMLE procedure yields $\bar{Q}_n^*(A,W)$ and our estimator is then a plug-in estimator, with j^{th} component:

$$\Psi_{\delta,t_j}(P_n^{\star}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\delta}\int k\left(\frac{x-t_j}{\delta}\right)\mathbb{I}(b_n^*(W_i) \leq x)dx$$

and standard errors are given by

$$\frac{\widehat{\sigma}_n(D_j^*(P_n^*))}{\sqrt{n}}$$

where $\widehat{\sigma}_n(D_j^*(P_n^*))$ is the sample standard deviation of $\{D_j^*(P_n^*)(O_i) \mid i \in 1:n\}$ and $b_n^*(W_i) = \bar{Q}_n^*(1,W) - \bar{Q}_n^*(0,W)$.

Performing a CV-TMLE To perform a CV-TMLE (Zheng and van der Laan 2010) we would define a split, B_n, which is a mapping on $1:n$, such that $B_n(i) = 1$ means the i^{th} observation is in the training set and $B_n(i) = 0$ means the i^{th} observation is in the validation set. We usually define 10 splits for which the validation sets are disjoint and comprise all n observations, as in typical 10-fold cross-validation. A CV-TMLE is defined as an average across the splits of estimates computed on the validation sets.

On the training set of each split B_n, we would use an ensemble learning package such as sl3 (Coyle et al. 2018a) or SuperLearner (Polley et al. 2017) to construct the initial fit, $\bar{Q}^0_{n,B_n}$, of outcome model $E_P[Y \mid A, W]$, and the initial fit, g_{n,B_n}, of the treatment mechanism, $E_P[A \mid W]$. We would then predict the outcome and treatment probabilities on the validation set defined by B_n. For all $i \in 1:n$, we therefore provide an estimate $\bar{Q}_n^0(A_i,W_i)$ and $g_n(A_i \mid W_i)$ for when observation i was in the validation set of one of the splits. With these predictions we can proceed with steps 2 through 4 above.

TMLE conditions for Estimating $\Psi_\delta(P_0)$

The importance of the TMLE mapping is we then have $\Psi_{\delta,t_j}(P_n^\star) - \Psi_{\delta,t_j}(P_0) = (P_n^0 - P_0)D_j^*(P_n^\star) + R_{j,2}(P_n^\star, P_0)$ where $R_2(P_n^*, P_0)$ is given by

$$\frac{-1}{\delta}\int\Big[k\Big(\frac{b_n^*(w)-t}{\delta}\Big)\Big(\Big(\frac{g_0(1|w)}{g_n(1|W)}-1\Big)\Big(\bar{Q}_0(1,w)-\bar{Q}_n^*(1,w)\Big)-$$

$$\Big(\frac{g_0(0|w)}{g_n(0|w)}-1\Big)\Big(\bar{Q}_0(0,w)-\bar{Q}_n^*(0,w)\Big)\Big)\Big]dQ_{W,0}(w)$$

$$+\frac{1}{\delta}\int\left[\int_{b(w)}^{b_0(w)}k\left(\frac{x-t}{\delta}\right)dx+k\left(\frac{b(w)-t}{\delta}\right)(b(w)-b_0(w))\right]dQ_{W,0}(w)$$

The reader may see the proof in the appendix of Levy and van der Laan, 2018.

Theorem 0.1. *Assume the following two conditions*

1. *$D_j^*(P_n^*)$ is in a P_0-donsker class for all j.*
2. *$R_{2,j}(P_n^*, P_0)) = o_p(1/\sqrt{n})$ for all j.*
3. *$D_j^*(P_n^*) \overset{L^2(P_0)}{\longrightarrow} D_j^*(P_0)$ for all j. We note we do not need this assumption as it is implied by the previous.*

Then the TMLE outlined in steps 1 through 4 above is an asymptotically efficient estimator of Ψ_{δ,t_i} for all i in $1:d$.

The proof is given in van der Laan and Rubin, 2006. It is shown in Zheng and van der Laan, 2010 that when performing a CV-TMLE, condition 1 above is automatically satisfied for the remainder term computed on each validation fold and, therefore, we only need to satisfy condition 2.

Remark 1. If the TMLE or CV-TMLE conditions hold for the initial estimates then they will also hold for the TMLE or CV-TMLE updates (van der Laan 2016). Thus, as we show in simulations, it is crucial we use state-of-the-art machine learning methods in forming our initial estimates of the data generating distribution.

Remark 2. From here on, all theorems will apply to either TMLE or CV-TMLE so we will use the lighter TMLE notation where we need not keep track of the splits, B_n.

The Use of Highly Adaptive Lasso When using the highly adaptive lasso (HAL) (van der Laan 2016), (van der Laan and Gruber 2016) to perform the initial estimates, we are guaranteed $\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2(P_0)}$ and $\|g_n - g_0\|_{L^2(P_0)}$ are $o_P(n^{-0.25})$ under the conditions that $\bar{Q}_0$ and g_0 are of bounded sectional variation norm and continuous from the right with left-hand limits. We also have the following corollary from Levy and van der Laan, 2018.

Theorem 0.2. *Our remainder term can be bounded is as follows:*

$$R_{2,i}(P_n^0, P_0) =$$

$$\frac{1}{\delta}O\left(\|g_n-g_0\|_{L^2_{P_0}}\|\bar{Q}_n^0-\bar{Q}_0\|_{L^2_{P_0}}\right)+\frac{1}{\delta}O\left(\|b_n^0-b_0\|_\infty^2\right)$$

$$\textit{or}\ \frac{1}{\delta}O\left(\|g_n^0-g_0\|_{L^2_{P_0}}\|\bar{Q}_n^0-\bar{Q}_0\|_{L^2_{P_0}}\right)+\frac{1}{\delta^2}O\left(\|b_n^0-b_0\|_{L^2_{P_0}}^2\right)$$

The reader may see the proof in theorem B.2 in Levy and van der Laan, 2018. This shows we are not guaranteed consistent estimates based on knowledge of the treatment mechanism as in the case of doubly robust estimators. By the above stated properties of HAL we immediately have the following corollary:

Corollary 1. *When using HAL to form initial estimates of $\bar{Q}_0$ and g_0, the TML estimator of $\Psi_{\delta,t_i}(P_0)$ (fixed bandwidth, δ) will be asymptotically efficient.*

Simultaneous Estimation and Confidence Bounds We apply a procedure to give simultaneous confidence bands (a number of standard errors to cover the truth for all d smoothed parameters at a given significance level) for d points on the TE CDF that accounts for correlation between estimates via use of the correlation matrix of the efficient influence curve approximation. It will yield bands smaller than that of a bonferroni correction (Dunn 1961), only being approximately equal to such when estimates are completely uncorrelated. The reader may consult Levy and van der Laan, 2018 for the procedure.

Allowing the Bandwidth, δ, to Vanish for n Large

The reader may notice that below we bound the remainder term in two different ways, one of which has δ in the denominator and the other which has δ^2 in the denominator. If we let δ approach 0 as a function of n, then we would prefer to only have δ in the denominator so as to allow δ to approach 0 faster and hence, a lower mean squared error. However, that condition is more difficult to guarantee as we will point out.

We will refer to the following facts, where P_n^0 is an initial fit of P_0 and P_n^* is a TMLE update of P_n^0.

1. The asymptotic variance of $\sqrt{n}(\Psi_{\delta,t_i}(P_n^*) - \Psi_{\delta,t_i}(P_0))$ is of order $1/\delta$. Theorem B.3, proven in Levy and van der Laan, 2018.
2. The bias between unsmoothed TE CDF value at t_i and the smoothed parameter, $\Psi_{t_i}(P_0) - \Psi_{\delta,t_i}(P_0)$, is of order δ^J, where J is the order of the kernel (power of the kernel's first non-zero moment) and we assume the TE CDF to have J continuous derivatives. Theorem B.4, proven in Levy and van der Laan, 2018a.

Theorem 0.3. *Assume the TE CDF has J continuous derivatives. Assume we allow our bandwidth = $\delta_n = O(n^{-1/(2J+1)})$. Let $\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2(P_0)} = o_P(n^{r_{\bar{Q}_n}})$ and $\|g_n - g_0\|_{L^2(P_0)} = o_P(n^{r_{g_n}})$ Then if $r_{g_n} + r_{\bar{Q}_n} \le -\frac{J+1}{2J+1}$ and either of*

- *A1: $\|\bar{Q}_n^0 - \bar{Q}_0\|_\infty = o_P\left(-\frac{J+1}{2(2J+1)}\right)$*
- *A2: $\|\bar{Q}_n^0 - \bar{Q}_0\|_{L^2(P_0)} = o_P\left(-\frac{2J+3}{4(2J+1)}\right)$*

$\sqrt{\delta_n n}R_2(P_n^0, P_0) \xrightarrow{p} 0$

This statement follows immediately from Theorem 0.2.

Theorem 0.4. *If using bandwidth of order $\delta_n = O(n^{-1/(2J+1)})$ and HAL to form initial predictions then if we use a kernel of order $J > \frac{4r+3}{2}$ and the TE CDF has J continuous derivatives, $\sqrt{\delta_n n}R_2(P_n^0, P_0) \xrightarrow{p} 0.$*

The statement follows from the fact HAL guarantees $\|f_0 - f_n^0\|_{L^2(P_0)} = O_P(n^{-1/4-1/8(r+1)})$, when fitting a function, f_0 of finite sectional variation norm that is continuous from the right with left-hand limits (van der Laan 2016).

Remark 3. The motive for this theorem is that if we wanted to minimize the MSE based on items 2 and 3 above, as for a kernel density estimator, we would want $\delta_n = O(n^{-1/(2J+1)})$. However, we also want the remainder term to become truly second order when blown up by $\sqrt{\delta_n n}$ in order for $\sqrt{\delta_n n}(\Psi_{\delta_n,t_i}(P_n^*) - \Psi_{\delta_n,t_i}(P_0))$ to have a limiting distribution. Thus, perhaps higher order kernels can be useful in relaxing the requirements of Theorem 0.3 in fitting the treatment mechanism and, especially, the outcome model. For fitting nuisance parameters that are functions of variables of dimension 5, we would need a kernel of order 12 or greater to guarantee theorem and 12 continuous derivatives of the TE CDF. If $\|f_0 - f_n^0\|_{L^\infty} = O_P(n^{-1/4-1/8(r+1)})$ then using a kernel of order $J > \frac{2r+1}{2}$ and assuming necessary smoothness on the TE CDF, would guarantee $\sqrt{\delta_n n} R_2(P_n^0, P_0) \xrightarrow{p} 0$. Thus, if $r = 5$, we only require a kernel of order 6 and hence, only 6 continuous derivatives for the TE CDF.

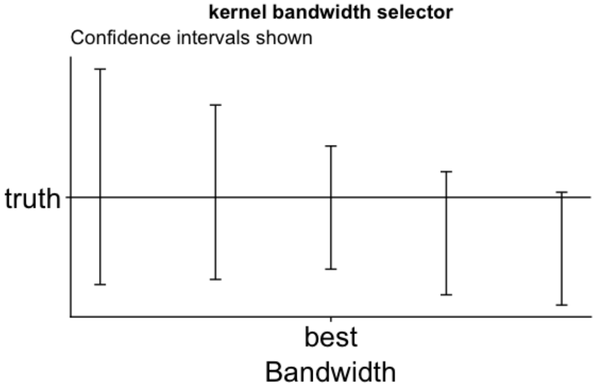
Simulations

Well-specified Models

For well-specified logistic models where the data generating system is given by the following: W is a random normal, $Pr(A = 1 \mid W) = g(A \mid W) = expit(.2 + .2 * W)$ and $E[Y \mid A, W] = expit(A + 2.5 * A * W + W)$. The TMLE's using the MLE as an initial estimate performed very well, with normal sampling distributions, nominal coverage (93% or higher) of the smoothed parameter, as expected, and did so for all kernels if we used bandwidth $n^{-1/(2J+1)}$ where J is the order of the kernel and we let n attain values of 1000, 2500, 5000, 10000, 25000 and 50000. The MSE was lowest for the well-specified MLE plug-in, also as expected, but not appreciably. In the highly unlikely scenario that we correctly specify the outcome model with a parametric form, TMLE performance appears very reliable for covering the smoothed parameter and yields vanishing standard errors as sample size grows.

A Method for Choosing Bandwidth for a Given Kernel We would like to form confidence bounds for the non-pathwise differentiable parameter or unsmoothed "true" parameter, $\Psi(P) = \mathbb{E}_P \mathbb{I}(b(W) \leq t)$ for $P \in \mathcal{M}$, and propose using some of the concepts in Chapter 25 of Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies by van der Laan and Rose, 2018. We start with a largest bandwidth of size $n^{-1/(2J+1)}$ where J is the order of the kernel. Then we divide the bandwidth into 20 equal increments from $n^{-1/(2J+1)}/20, 2n^{-1/(2J+1)}/20, ..., n^{-1/(2J+1)}$. We then find the smallest set of 5 or more consecutive bandwidths that are monotonic estimates with respect to the bandwidth. If no such 5 or more consecutive bandwidths are found then we choose the bandwidth $n^{-1/(2J+1)}$. Let us call the consecutive bandwidth sequence, $B_c = \{h_1, ..., h_c\}$, where h_1 is the smallest. We also monotonize the variance so as to force it to be increasing as the bandwidth gets smaller. We then form confidence intervals using the monotonized variance for each bandwidth in B_c. If the sequence of estimates is decreasing (increasing) as bandwidth decreases (for bandwidths in B_c), then we choose the confidence interval with the minimum (maximum) right (left) bound. The idea is that we are minimizing the MSE while maintaining nominal coverage, assuming that the smoothed parameters are monotonic as a function of the bandwidth for the bandwidths in B_c and that this monotonicity represents the monotonicity as the bandwidth approaches 0. If we apply the selector for kernel order 10, we greatly assist in covering the true parameter and maintain nominal or very near nominal coverage of the smoothed parameter for sample sizes up to 50,000. We show results for sample sizes 1000, 2500 and 5000 in Table 1. Similar results held for lesser order kernels as well. Figure 1 below displays the heuristic behind our bandwidth selector.

Figure 1:



Simulations for Misspecified Models

We call these simulations "misspecified" because we use the highly adaptive lasso or HAL (van der Laan and Gruber 2016) to recover the model without any specification on functional forms. The data generating system consisted of the following functions in the order listed. W is a random normal, $Pr(A = 1 \mid W) = g(A \mid W) = expit(-.1 - .5 * sin(W) - .4 * (|W| > 1) * W^2)$ and $E[Y \mid A, W] = expit(.3*A + 5*A*sin(W)^2 - A*cos(W))$. We simulated 1100 draws from the above data generating system and computed simultaneous TMLE's for the TE values -0.098, -0.018 0.062, 0.142, 0.222, 0.302, 0.382 and 0.462 using bandwidth $2500^{-0.2}$ and an order 1 polynomial kernel. Similar results held for the uniform kernel.

Here we show the huge advantage of data adaptive estimation in obtaining the initial estimates for CV-TMLE, using the highly adaptive lasso. TMLE_glm used used logistic regression with main terms and interactions for the initial estimates in CV-TMLE, while TMLE_HAL used HAL for the initial estimates. We can see it is catastrophic to use logistic regression here while using HAL with TMLE procedure achieved very close to nominal coverage with essentially no bias (see table 2). Targeting helped remove bias from the HAL initial estimates as well, shown in Figure 2, for one of eight points on the TE CDF simultaneously estimated by TMLE_HAL. The other seven points had very similar sampling distributions and bias.

Software

The reader may visit https://github.com/jlstiles/TECDFsim (Levy 2018c) for procedures on how to reproduce the results here-in and also visit https://github.com/jlstiles/TECDF (Levy 2018a) for software on performing the targeting step after obtaining initial estimates. This estimator is also available in the package https://github.com/tlverse (Coyle et al. 2018b), where the reader can also perform ensemble learning.

Conclusion

We have developed an estimator to efficiently estimate, under conditions, the kernel smoothed version of the TE CDF and also allow the bandwidth to approach zero and guarantee a normal limiting distribution for the TE CDF itself. Furthermore, our estimator does not rely on any parametric assumptions on the data generating distribution. We have shown our estimator hinges on data adaptive estimation, particularly the use of the highly adaptive lasso, to make our initial estimates in the targeted learning (van der Laan and Rose 2011) framework. The TML update helps eliminate bias and provides us with immediate inference for the smoothed parameter via the sample standard deviation of the efficient influence curve approximation. Our simulations have shown that for well-specified models, choosing the bandwidth of optimal order $n^{-\frac{1}{2J+1}}$ (and hence a vanishing bandwidth in n), assuming J continuous derivatives for the TE CDF, provides normal and unbiased sampling distributions for the smoothed parameter.

The next step is to develop a way to optimally (smallest MSE possible) select the bandwidth, δ_n, and kernel so that the estimator minus the truth blown up by $\sqrt{n\delta_n}$ is normally distributed and covers the TE CDF nominally. Our bandwidth selector in this paper still gives nominal or near-nominal coverage of the smoothed parameter as the bandwidth vanishes for large n, but is not yet reliable for covering the TE CDF itself, though we show it is a big improvement over setting the bandwidth to $n^{-\frac{1}{2J+1}}$. Our bandwidth selector relies on the assumption that the smoothed parameter is monotonically increasing or decreasing toward the unsmoothed parameter as the bandwidth vanishes. Our method of determining this monotonicity is somewhat arbitrary and it also remains to be seen how this monotonicity generally holds for small bandwidths. For instance, if the monotonicity changes direction for a small bandwidth, our proposed bandwidth selector might be problematic.

Table 1: coverage of smoothed parameter, kernel is order 10

Estimating smoothed TE CDF						
	n = 1000		n = 2500		n = 5000	
TE	meth	fixed	meth	fixed	meth	fixed
−0.145	0.907	0.947	0.920	0.949	0.916	0.941
−0.085	0.911	0.953	0.950	0.946	0.939	0.934
−0.025	0.925	0.944	0.950	0.960	0.958	0.948
0.035	0.916	0.940	0.929	0.949	0.942	0.966
0.095	0.934	0.951	0.934	0.949	0.946	0.942
0.155	0.933	0.952	0.942	0.946	0.936	0.948
0.215	0.927	0.958	0.927	0.951	0.932	0.941
0.275	0.893	0.955	0.913	0.955	0.905	0.951
Estimating true TE CDF						
−0.145	0.671	0.001	0.436	0	0.298	0
−0.085	0.612	0.136	0.593	0.024	0.743	0.001
−0.025	0.770	0.166	0.615	0.019	0.405	0.001
0.035	0.850	0.071	0.924	0.001	0.927	0
0.095	0.747	0.070	0.895	0	0.912	0
0.155	0.750	0.251	0.859	0.020	0.903	0
0.215	0.695	0.947	0.717	0.861	0.793	0.692
0.275	0.858	0.008	0.817	0	0.707	0

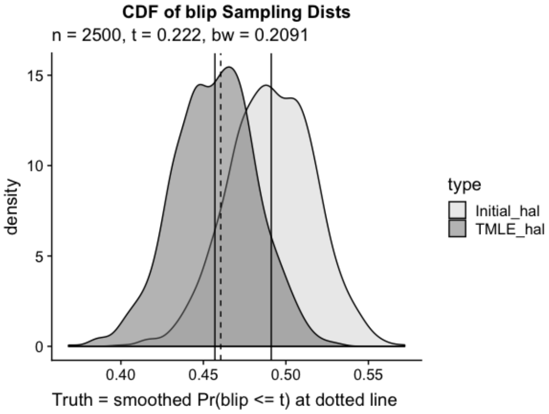
meth means we applied the bandwidth selection method, fixed means we used bandwidth $n^{-1/(2J+1)}$
where J is the kernel order.

Table 2: TMLE with HAL vs TMLE with glm

	MSE		Coverage	
TE	TMLE_hal	TMLE_glm	TMLE_hal	TMLE_glm
-0.098	0.0008	0.0200	0.9173	0
-0.018	0.0009	0.0168	0.9255	0.0182
0.062	0.0009	0.0053	0.9373	0.5846
0.142	0.0007	0.0037	0.9492	0.8100
0.222	0.0006	0.0217	0.9618	0.1046
0.302	0.0007	0.0472	0.9518	0
0.382	0.0007	0.0580	0.9446	0
0.462	0.0007	0.0453	0.9409	0

Results for smoothed parameter
TMLE_HAL used HAL for initial estimates
TMLE_glm used glm for initial estimates
simultaneous TMLE_hal coverage was 90%
TMLE_glm coverage was 3%

Figure 2: TMLE Bias Reduction



References

Cox, D. R. 1958. *Planning of Experiments*. John Wiley and Sons, Inc., 5th edition.

Coyle, J.; Malenica, I.; Hejazi, N.; and Levy, J. 2018a. sl3.

Coyle, J.; Malenica, I.; Hejazi, N.; and Levy, J. 2018b. tlverse.

Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association* 56(293):52–64.

Frechet, M. 1951. Sur les tableaux de correlation dont les marges sont donnees. *Annals University Lyon* A(14):53–77.

Greenland, S., and Robins, J. 1986. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 15(3).

Heckman, J. J., and Smith, J. 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64(4):487–535.

Heckman, J. J., and Smith, J. A. 1998. Evaluating the welfare state. *National Bureau of Economic Research* (6542).

Levy, J., and van der Laan, M. 2018. Kernel smoothing of the treatment effect cdf. *arXiv:1811.06514 [stat.ME]*.

Levy, J.; van der Laan, M.; Hubbard, A.; and Pirracchio, R. 2018. A fundamental measure of treatment effect heterogeneity. *arXiv:1811.03745 [stat.ME]*.

Levy, J. 2018a. blip cdf.

Levy, J. 2018b. Canonical least favorable submodels: A new tmle procedure for multidimensional parameters. *arXiv:1811.01261 [stat.ME]*.

Levy, J. 2018c. Tecdfsim.

McCullagh, P. 1983. Quasi-likelihood functions. annals of statistics. 11(1):59–67.

Neyman, J. 1923. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Sciences* 5(4):465–480. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original which appeared in Roczniki Nauk Rolniczych Tom X, 1923.

van der Laan, M., and Gruber, S. 2016. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The International Journal of Biostatistics* 12(1):351–378.

van der Laan, M., and Rose, S. 2011. *Targeted Learning*. New York: Springer.

van der Laan, M., and Rose, S. 2018. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies (2018)*. Springer International Publishing AG.

van der Laan, M., and Rubin, D. 2006. Targeted maximum likelihood learning. *U.C. Berkeley Division of Biostatistics Working Paper Series* (213).

van der Laan, M. 2016. A generally efficient targeted minimum loss based estimator. *U.C. Berkeley Division of Biostatistics Working Paper Series* 343.

Pearl, J. 2000. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Polley, E. C.; LeDell, E.; Kennedy, C.; and van der Laan, M. 2017. Superlearner: Super learner prediction. R package version 2.0-23-9000.

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period. *Journal of Mathematical Modeling* 7:1393–512.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

Strotz, R., and Wold, H. 1960. Recursive vs. nonrecursive systems: an attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica* 28(2):417–427.

van der Vaart, A. 2000. *Asymptotic Statistics*, volume Chapter 25. Cambridge, UK: Cambridge University Press.

Wedderburn, R. 1974. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* 61.

Wright, S. 1921. Correlation and causation. *Journal of Agricultural Research* 20(7):557–585.

Zheng, W., and van der Laan, M. 2010. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series* (273).